Procedural learning in adults with and without dyslexia: Reliability and individual differences

Cátia Margarida Ferreira de Oliveira

PhD

University of York

Psychology

June 2022

Abstract

Procedural learning is thought to play a crucial role in language and literacy acquisition through the extraction of linguistic regularities from the sensory input. Aligning with this, a procedural deficit is hypothesized to be causally implicated in neurodevelopmental disorders such as dyslexia and developmental language disorder. In this thesis, we comprehensively examined the role of procedural memory in the serial reaction time task (SRTT) in language/literacy in typical and atypical populations. We began by establishing, and potentially improving, the reliability of the SRTT in typically developing adults by examining the impact of the similarity between sequences at test and retest, the addition of an extra session, and the inclusion of an interstimulus interval. Despite numerical improvements, no experimental manipulation resulted in adequate test-retest reliability. This was further confirmed in a meta-analysis that assessed the reliability of the SRTT. Alongside the careful examination of the psychometric properties of the SRTT, the relationship between procedural learning and language/literacy was reported, with no support for associations between these measures. This was also replicated in a second meta-analysis in typical and atypical populations. Additionally, the performance of adults with and without dyslexia on the SRTT was contrasted across three sessions. This study represented the first attempt to assess the reliability of the SRTT in a disordered population. Again, the test-retest reliability of the SRTT was well-below adequate psychometric standards. There was no evidence for a procedural learning impairment in the dyslexic participants in any session, and there was only a moderate relationship between nonword repetition and procedural learning in this group. Nonetheless, procedural learning and attention were consistently correlated across experiments. To conclude, these findings do not rule out the possibility that procedural learning is involved in language/literacy development and disorders; instead, they highlight the need for more reliable measures and more testable hypotheses.

List of Contents

List of Tables

List of Figures

List of Appendices

List of Accompanying Material

The experimental chapters each have an associated page on the Open Science Framework, containing pre-registrations, data, and analyses.

Chapter 2: https://osf.io/fn9mw/
Chapter 3: https://osf.io/a65hn/
Chapter 4: https://osf.io/ev2xw/
Chapter 5: https://osf.io/dt5xs/

Acknowledgments

First and foremost, I wish to thank my supervisors Dr. Lisa Henderson and Dr. Emma Hayiou-Thomas for your continuous guidance and support throughout this journey. I could not have asked for better supervisors. I am incredibly grateful to have learnt from you not only how to be a better researcher, in all its integrity and discipline, but also how caring and generous a supervisor can be. Thank you for making my time at York such an enriching and lovely experience.

A special thanks also goes to Professor Gareth Gaskell and Dr. Tom Hartley for their thought-provoking questions and feedback as part of my thesis advisory panel. I thoroughly enjoyed these meetings as, even though they were often challenging, they always made me grow as a scientist.

Another huge thanks go to Dr. Jelena Mirkovic for your kindness and guidance. I was tremendously lucky to have had the chance to have you as a borrowed supervisor (at least that's how I always saw it).

Further thanks go to the Sleep, Language and Memory group, the developmental group, the Reproducibilitea group, and our little journal club. Thank you for always making me feel welcomed and for your insightful suggestions. This is especially true for Dr. Sam Berens. Thank you for being so patient and generous with your time. I have learnt so much from you.

To all the brilliant teachers I was lucky to meet along the way who have continuously sparked the joy for knowledge: Dr. Jenny Thomson, Dr. Antonio Benítez-Burraco, Dr. Luis M. Augusto and Dr. Wolfgang Viechtbauer, to name a few.

To my York family, without whom this journey would not have been as enjoyable: Amanda Hickey, Anna Guttesen, Juliana Olivier and Martin Scott. I could not have asked for better friends. Thank you for the walks, ridiculous jokes, long rant sessions, and for your understanding. A special thanks goes to Juliana, for always believing this was possible even at 5am when confidence wavers.

I cannot forget all the other friends and family who have kept me sane throughout the PhD, thank you.

Finally, but certainly not least, to Nuno who makes everything possible.

For my brother

For my grandfather

Author's declaration

I declare that this thesis is a presentation of original work completely solely by the author under the supervision of Dr. Lisa Henderson and Dr. Emma Hayiou-Thomas. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as references.

The research was supported by a grant from the Fundação para a Ciência e a Tecnologia, Portugal (DFA/BD/6560/2020).

PUBLICATIONS

Chapter 2 of this thesis has been published as follows:

- Oliveira, C. M., Hayiou-Thomas, M. E., & Henderson, L. (2022, May 10). Reliability Of the Serial Reaction Time Task: If At First You Don't Succeed, Try Try Try Again. https://doi.org/10.31234/osf.io/hqmy7

CONFERENCE PRESENTATIONS

Several findings have also been presented at conferences as follows:

Chapter 2:

- Oliveira, C. M., Hayiou-Thomas, M. E., & Henderson, L. M. (2020, June). Reliability of the SRT task: if at first you don't succeed, try try try again. Poster presented at IMPRS Interdisciplinary Approaches in the Language Sciences, online.
- Oliveira, C. M., Hayiou-Thomas, M. E., & Henderson, L. M. (2020, November). Reliability of the SRT task: if at first you don't succeed, try try try again. Poster presented at White Rose Doctoral Training Partnership, online.
- Oliveira, C. M., Hayiou-Thomas, M. E., & Henderson, L. M. (2021, April). Procedural learning in the serial reaction time task: a long way to stability. Poster presented at EPS, online.
- Oliveira, C. M., Hayiou-Thomas, M. E., & Henderson, L. M. (2021, April). Procedural learning in the serial reaction time task: a long way to stability. Poster presented at BNS, online.

Chapters 4 and 5:

- Oliveira, C. M., Hayiou-Thomas, M. E., & Henderson, L. M. (2022, June). Procedural learning on the SRTT: Sensitivity to group level and individual differences in language and literacy. Poster and flash talk presented at Statistical Learning conference, San Sebastian, Spain
- Oliveira, C. M., Hayiou-Thomas, M. E., & Henderson, L. M. (2022, June). Procedural learning on the SRTT: Sensitivity to group level and individual differences in language and literacy. Poster and flash talk presented at Celebrating Dorothy: A Festschrift for Dorothy V. M. Bishop, London, UK.

Chapter 1. Introduction

Language acquisition is a fundamental human ability, crucial for educational success, social development, and even mental health (McGregor, 2020). It is vital that we understand the mechanisms of language learning so that we can optimise this process over the lifespan. Models of language acquisition have been domain-specific (Chomsky, 1965; Karmiloff-Smith, 2015) or domain-general (i.e., drawing on general cognitive memory systems). One such theory in the latter camp suggests that, in addition to a wide range of cognitive functions such as object and scene perception (e.g., Coppola et al., 1998; Lauer et al., 2018) and social cognition (e.g., Monroy et al., 2017; Norman & Price, 2012; Ruffman et al., 2012), a domain-general cognitive ability in pattern sensitivity is involved in language acquisition (Batterink et al., 2019; Batterink & Paller, 2019). Indeed, seminal work by Saffran and colleagues (1996) deemed the detection and extraction of patterns in the speech stream as fundamental to language development. A key neurocognitive system that is thought to underpin the learning of patterns and regularities in the environment is the procedural memory system, which according to Ullman and colleagues' proposal is pivotal to language acquisition (Ullman, 2001a, 2004, 2016c; Ullman et al., 2020) and impairments to this system may represent the cause, or at least a risk factor, for developmental language disorder (DLD) and dyslexia. However, despite this breadth of interest, accumulating evidence for this theory has been somewhat inconsistent, with recent concerns over the psychometric properties of tasks used to measure procedural learning as well as the construct itself.

This literature review begins by providing a brief contextual overview of human memory organisation, before narrowing in on procedural memory. We will review the definition of this construct and the tasks used to measure it, paying particular attention to the commonly used Serial Reaction Time task (SRTT). We then outline key theories that incorporate a role for procedural memory in typical and atypical language and literacy acquisition (i.e., the procedural/declarative model and the procedural deficit hypothesis, respectively), and critically review behavioural and neurological evidence for these theories. Importantly, evidence from typical and atypical (namely, dyslexia and developmental language disorder) populations will be considered to provide a richer understanding of the processes underlying language acquisition across language and literacy spectra. We then reflect on current challenges for the field, focussing on issues with capturing individual differences, primarily the poor psychometric properties of the tasks used to assess procedural memory. Experimental tasks designed to elicit group-level effects, such as the SRTT, may not be suitable for individual differences research. Finally, we conclude by proposing future directions for addressing the role of procedural memory in language acquisition and impairments, including considering both extrinsic and intrinsic factors that might influence the psychometric properties of procedural learning tasks. Ultimately, a better understanding of how to measure procedural learning will be fundamental to evaluating the role of this construct in typical and atypical language development. If procedural learning is indeed impaired in atypical populations, this research not only has the potential to advance theories of language development but could ultimately inform both the assessment and remedial practices for language disorders such as dyslexia and developmental language disorder.

Human memory systems: a brief overview

The conceptualisation of memory not as a single faculty, but instead as comprising multiple memory systems has been widely agreed upon (Eichenbaum & Cohen, 2001; L. R. Squire et al., 2004; L. R. Squire, 2004), yet the division of memory systems has continued to be debated; dichotomous classifications of memory are the most common, but three or more dissociable memory systems have also been proposed (Schacter, 1987). The most central classification separates declarative and nondeclarative memory systems (L. R. Squire, 2004; L. R. Squire & Zola, 1996). The distinction between these systems was originally based on cognitive neuropsychological investigations of individuals with amnesia (e.g., Knowlton et al., 1992; L. R. Squire & Alvarez, 1995) and Parkinson's disease (e. g., Knowlton et al., 2006; J. Smith et al., 2001), and on animal studies (e.g., Gaffan, 1974; Teng et al., 2000). In individuals with amnesia and Parkinson's disease, there is often a functional dissociation between performance on declarative and nondeclarative tasks. While patients with amnesia tend to show impaired formation of new memories related to facts and events (declarative memory), they often exhibit preserved gradual acquisition of habits and skills (procedural memory). The opposite pattern is often observed in individuals with Parkinson's disease (Hay et al., 2002; Knowlton et al., 1992; J. Smith et al., 2001).

The dichotomisation of memory systems does not preclude their parallel operation to support behaviour. Instead, it suggests that declarative and nondeclarative memory systems can be distinguished based on the type of information they process and the principles by which they operate. Declarative memory has been implicated in the encoding, storage, and retrieval of general knowledge about the world (semantic memory) or specific events (episodic memory) across sensory modalities and cognitive domains (L. R. Squire, 1994; Tulving & Markowitsch, 1998). This memory system is proposed to be dynamic (L. R. Squire & Wixted, 2011; Ullman & Pullman, 2015), such that rather than being limited to its involvement in the learning and retention of the new material, once learned, knowledge acquired by the declarative memory system tends to be generalised and used flexibly in different contexts. Traditionally, declarative learning is considered to be a relatively rapid process, which only requires a small number of exposures to the stimuli to be encoded (Ullman, 2013; Gais & Born, 2004; Holz et al., 2012), yet despite the relatively quick encoding stage, more recent models emphasise a role for longer-term declarative memory consolidation processes (Gais & Born, 2004; Holz et al., 2012). However, this memory system requires attentional resources and benefits from the intention to learn (Knowlton & Moody, 2008).

Declarative memory is distinguished from nondeclarative memory, specifically, a heterogeneous group of learning abilities that comprise procedural memory, priming, and perceptual learning, simple classical conditioning, and non-associative learning (P. J. Reber, 2008; L. R. Squire, 2004). Knowledge and habits in this broad nondeclarative system are acquired gradually based on repeated experience, often without conscious awareness, and, in some circumstances, have been claimed to be less flexible and more context-bound than declarative memory (Squire & Dede, 2015).

There is support for a distinction between declarative and nondeclarative memory at the neurobiological level. Specifically, functional magnetic resonance imaging studies have shown that the declarative memory system relies primarily on the medial temporal lobe and connected cortical regions (L. R. Squire et al., 2015; Zola-Morgan & Squire, 1991). Over time and through offline replay (especially during sleep), these memory traces, which are initially temporary and labile memories coded in the hippocampus, are thought to be reorganised for long-term retention in the neocortex, especially in the temporal and temporo-parietal regions (McClelland & O'Reilly, 1995; L. R. Squire & Zola, 1996; Zola-Morgan & Squire, 1991), and gradually become more stable and less susceptible to interference (L. R. Squire et al., 2015). Conversely, nondeclarative memory (e.g., the procedural memory system) relies on a fronto-striatal circuit (including the frontal cortex, the basal ganglia, and the thalamus) and other neural substrates such as the cerebellum (Eichenbaum, 2012; Ullman et al., 2020) (considered in further detail in the following section).

The declarative and nondeclarative memory systems have been found to work cooperatively, independently and in competition, yet it is still unclear how their engagement is modulated to optimise learning and behaviour (Foerde et al., 2006). Whilst learning in most tasks can proceed by engaging more than one memory system, other tasks may be more efficiently mastered by one system over another (L. Squire & Dede, 2015), with this relative weighting at least partly modulated by task demands (Foerde et al., 2006). For example, studies using the SRTT (described in more detail in the section *Serial Reaction Time task(s)*), which indexes procedural memory, have shown enhanced procedural learning/consolidation in this task when declarative memory has been suppressed through the use of repetitive transcranial magnetic stimulation (Ambrus et al., 2020; Galea et al., 2010), due to long-term effects of alcohol (Virag et al., 2015) or in participants' experiencing cognitive fatigue

(Borragán et al., 2016) or under hypnosis (Nemeth, Janacsek, Polner, et al., 2013). This suggests that, due to the competitive relationship between declarative and nondeclarative memory systems, the disruption of the declarative memory system in the SRTT can improve performance of the nondeclarative memory systems, possibly by reducing resource competition (Galea et al., 2010).

An overview of the procedural memory system

Procedural memory, one of the systems under the umbrella of nondeclarative memory, underlies the acquisition, consolidation, and automatization of motor, perceptual and cognitive skills that involve the integration of sequenced, statistical or probabilistic knowledge across time and space (Eichenbaum, 2002; Eichenbaum & Cohen, 2001; Knowlton & Moody, 2008; Ullman, 2004). Procedural learning is a gradual process that requires multiple exposures to stimuli but fewer attentional resources than declarative memory (Reber, 1993; Ullman, 2013), is mostly independent from general intelligence (Danner et al., 2017; Feldman et al., 1995; A. S. Reber et al., 1991) and not necessarily accessible to consciousness (Knowlton & Moody, 2008). Thus, procedural memory is often referred to as an implicit memory system, yet these terms are not interchangeable since implicit memory encompasses other nondeclarative memory systems (L. R. Squire, 1994).

Evidence from both healthy individuals and those with neurological dysfunctions has repeatedly demonstrated that, at a neurological level, the procedural memory system comprises a network of brain structures that includes the basal ganglia (specifically, the striatum), the cerebellum, and portions of the parietal and frontal cortices (Packard & Knowlton, 2002; Kandel, Schwartz, & Jessell, 2012; Packard & Knowlton, 2002; Parent & Hazrati, 1995; PascualLeone, Wassermann, Grafman, & Hallett, 1996; Ullman, 2004, 2006). Notably, these structures may be differentially activated depending on the nature and stage of learning. Whilst the associative striatum has been shown to play a crucial role in earlier stages of procedural learning, at later stages, after extended training, a reduction in the activation in the associative striatum has been observed (Ashby et al., 2010; Coynel et al., 2010), alongside an increase in activation of the putamen and in the sensorimotor striatum (Ashby et al., 2010; Patterson & Knowlton, 2018).

The challenges of defining procedural memory

Despite decades of research referring to the construct of procedural memory, there remain several challenges regarding its definition. Not least, various experimental paradigms have been used to measure procedural learning, and subsequently there is little agreement on what constitutes the gold-standard proxy for procedural learning (Bogaerts et al., 2021). This potentially reflects the vagueness of the definition. According to Bogaerts et al. (2021), procedural learning as a construct can be defined either by its neuroanatomical demarcations, the nature of the knowledge that is acquired or the computations underlying procedural learning. A neuroanatomical view would regard learning that relies on the basal ganglia and its circuitry as procedural learning. Extrapolating this criterion to experimental tasks, paradigms that tap into the neuroanatomical structures associated with procedural memory would be considered adequate proxies of procedural learning. However, behavioural tasks rarely map cleanly onto a predefined neuroanatomical substrate, and tasks designed to index procedural memory are no exception. Some of the most widely used paradigms in this field are the SRTT (described in detail in the following section), artificial grammar learning, speech segmentation, and probabilistic classification, and neuroimaging evidence from all of these experimental tasks has shown activation of the basal ganglia (e.g., artificial speech segmentation paradigm: e.g., McNealy et al., 2010; SRT: e.g., Janacsek et al., 2020; artificial grammar learning: e.g., Forkstam et al., 2006; Petersson et al., 2012; probabilistic classification task: e.g., Poldrack et al., 1999), most consistently for the SRTT (Williams, 2020). On the other hand, there is also evidence of concurrent activation of the medial temporal lobes for some of these tasks (artificial grammar learning task: e.g., Forkstam et al., 2006; Lieberman et al., 2004; artificial speech segmentation paradigm: e.g., Schapiro et al., 2014; SRTT: e.g., Naismith et al., 2010), as well as for others considered to index similar processes (Contextual cueing: e.g., Chun & Phelps, 1999; Greene et al., 2007; Hebb repetition: e.g., Attout et al., 2020; Kalm et al., 2013).

The picture is further complicated by apparently contradictory evidence from neuropsychological studies. Despite the evidence in neurotypical adults showing activation in the medial temporal lobe in some procedural memory tasks, patients with medial temporal lobe damage (amnesia) show preserved learning effects on these tasks (artificial grammar learning task: e.g., Knowlton et al., 1992; Knowlton & Squire, 1996; SRTT: e.g., P. J. Reber & Squire, 1994, 1998; Hebb: e.g., Gagnon et al., 2004; probabilistic classification task: e.g., Knowlton et al., 2006). This pattern suggests that, while it is possible for 'procedural' learning to occur without the engagement of the medial temporal lobe, in typically developing children and adults both memory systems are often active during learning (Batterink et al., 2019). This is further supported by the impaired performance of patients with Parkinson's disease on procedural memory tasks when compared to healthy controls (SRTT: Clark & Lum, 2017a). Surprisingly, given the damage to the basal ganglia that characterises Parkinson's disease, patients nevertheless show preserved performance on the artificial grammar learning task (P. J. Reber & Squire, 1999; Witt et al., 2002) suggesting that they may rely on the intact medial temporal lobe to support learning (Batterink et al., 2019; P. J. Reber & Squire, 1999). Despite incongruencies between imaging and patient studies for most experimental paradigms, unlike these tasks, there is convergent evidence for the involvement of the basal ganglia in the SRTT and a dissociation from the hippocampus based on amnesic patients. Although, there is still some imaging evidence for hippocampal involvement that needs to be better understood (Williams, 2020), overall, this suggests that the SRTT is possibly a purer measure of procedural memory than other paradigms.

A further distinction within the definition of procedural memory focuses on the nature of the learning and knowledge acquired. Definitions of the procedural memory construct often focus on its implicit and incidental nature, where individuals are not alerted to the regularities in the stimuli and learning on procedural memory paradigms proceeds without conscious awareness of the process and the knowledge acquired (Batterink et al., 2019). In the Weather prediction task (Knowlton et al., 1994), participants are presented with one or more cards on every trial; these cards are, unknowingly to the participant, independently associated with a fixed probability of the weather outcome. Participants are instructed to predict the weather outcome (sunshine or rain) based on the card combinations; both outcomes occur equally often. Similarly, in the artificial grammar learning paradigm (A. S. Reber, 1967), participants are first exposed to a training set that comprises combinations of strings generated by a finite-state grammar without being alerted to the nature of the task. After exposure, participants are asked to judge whether the new stimuli conform to the grammatical rules of the artificial language. In both tasks, procedural learning is reflected in the participants' ability to extract probabilistic information, which in these tasks would be reflected in the ability to predict above chance the weather outcome on the weather prediction task, and correctly judge the grammaticality of new stimuli on the artificial grammar paradigm. Crucially, despite evidence of procedural learning, participants' explicit knowledge of the underlying regularities is usually limited (e.g., artificial grammar learning: Dienes et al., 1991; artificial speech segmentation paradigm: Virtala et al., 2018; weather prediction task: Kemény, 2014), though it may vary depending on task characteristics (e.g., interstimulus interval: Destrebecqz & Cleeremans, 2001, 2003). We return to the issue of task purity later in this chapter in the section on *Task-pure measures*

Finally, procedural memory may be defined by its underlying computations, whereby learning in this memory system requires the extraction of statistical, probabilistic, or sequential information. Two processes can be distinguished as underlying procedural memory depending on the regularities that are assimilated: sequential and statistical learning. Sequential learning refers to the processes involved in acquiring serial information that occurs in the same order. Statistical learning, on the other hand, takes place when learning occurs through the assimilation of probability-based information (Simor et al., 2019; Tóth-Fáber et al., 2021). Behavioural and neural differences have been found between these procedural learning systems (Nemeth, Janacsek, & Fiser, 2013; Simor et al., 2019) (described in detail in section *Tracking the stages of procedural memory via the SRTT*). Thus, a task that involves the acquisition of regularities would be considered an adequate proxy of procedural memory. As previously described, most experimental paradigms described involve the tracking of probabilities and continuous binding of temporal or spatial information. Whilst the specificity of the computations required may vary depending on the task, the paradigms typically used in the field of statistical learning (including the SRTT) still share a significant degree of uniformity and thus may fail to represent the heterogeneity of real-world regularities (Frost et al., 2019). For example, in the 20 years of research using the artificial speech segmentation paradigm that followed Saffran et al's seminal publication in 1996, 84% of studies used syllables as linguistic units, 82% embedded either triplets or pairs of stimuli in the input, and over 90% used transitional probabilities of 1.0. The similarity of tasks used in this literature is not limited to the individual elements or the high probability of the transitional probabilities, but also pertains to the duration of the exposure to the patterned stimuli, and to the subsequent testing phase which usually takes a 2 alternative-forced-choice format. More recent research in the field has begun to use more diverse paradigms: the original artificial speech segmentation paradigm, as originally designed by Saffran et al. (1996) is no longer used by most studies (a drop from 60% to 35% between 2016 and 2018), yet, for those studies that examined auditory stimuli, syllables continue to be used as linguistic units in 60% of the studies. Similarly, the duration of the familiarisation phase, the number and length of the stimuli remain broadly uniform across experiments.

Due to its underspecification and lack of precision as a construct, procedural memory is often reduced to the properties of the tasks used to measure it (Frost et al., 2019). However, as indicated by Bogaerts et al. (2021), a single framework of procedural memory may not be the solution. The various definitions of procedural memory are not mutually exclusive and instead offer a means for discussion about the definition and operationalisation of procedural memory. Furthermore, while the various tasks are often used interchangeably, assuming a unitary procedural memory capacity is problematic as there is only limited evidence for correlations between different procedural memory paradigms (Arnon, 2020; Kalra et al., 2019; Siegelman & Frost, 2015; West, Shanks, et al., 2021). This raises the question of whether these weak correlations emerge because these tasks capture non-overlapping constructs or whether true correlations are being attenuated due to measurement error (Buffington et al., 2021; Rouder et al., 2019).

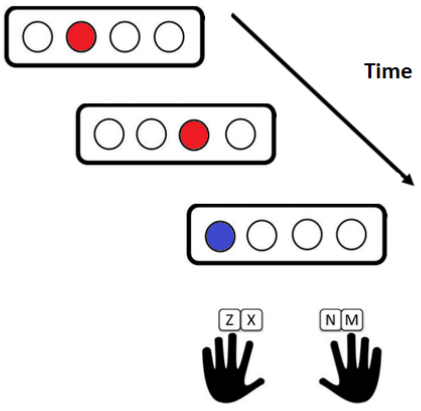
In this thesis, our definition of procedural learning is based on what is captured by the SRTT, primarily due to its widespread use for assessing the role of procedural memory in language and literacy acquisition and impairments. Compared to other procedural memory tasks, the SRTT provides the most consistent evidence for basal ganglia involvement, with both neuroimaging and patient studies demonstrating the involvement of the basal ganglia in the SRTT (Williams, 2020). Additionally, the seminal groundwork examining the psychometric properties of this task (e.g., Kalra et al., 2019; Siegelman & Frost, 2015; West et al., 2018) informs the validity of the available findings for advancing theories of the role of procedural memory in language and literacy development. For the remainder of this chapter, we first describe this task in detail and consider how it captures procedural learning. We then review behavioural and neuroimaging studies that utilise the SRTT in populations with and without dyslexia and developmental language disorder with the aim of assessing evidence related to the role of procedural memory in language and literacy development. We conclude by reviewing, in more detail, some of the barriers to individual differences research.

Serial Reaction Time Task(s)

The SRTT (Figure 1.1) is one of the most commonly used tasks for measuring procedural learning. The SRTT has undergone several adaptations since the original version first developed by Nissen and Bullemer (1987), but the core design involves the appearance of a visual target in one of four possible spatial locations, with participants being required to react as quickly as possible by pressing a button on a keyboard that corresponds to the location of the stimulus on screen. Unknown to the participants, the presentation of the stimuli follows a pattern. A decrease in response times for the patterned trials when compared to random trials is taken as evidence of learning, on the basis that participants' knowledge of the sequence allows them to anticipate the position of the following stimulus. Evidence of procedural learning in the SRTT in the absence of explicit awareness in probe tests is generally taken as evidence that the learning remains implicit (Wilkinson & Shanks, 2004).

Figure 1.1

Example of a probabilistic serial reaction time task sequence (red dots represent probable trials and blue improbable trials)



Three main hypotheses have been put forward to explain learning in the SRTT: the stimulus-based hypothesis, the response-based hypothesis, and the stimulus-response (S-R) rule hypothesis. Whilst the stimulus-based hypothesis suggests that learning is perceptual, depending on associations between stimuli but not on the response (Remillard, 2003), the response-based hypothesis, instead proposes that learning occurs through the association of sequential responses (Bischoff-Grethe et al., 2004; Willingham, 1999). Alternatively, the S-R rule hypothesis postulates that procedural learning occurs due to an associative process whereby mappings between stimuli and responses are made (Schumacher & Schwarb, 2009; Schwarb & Schumacher, 2012). Thus, an appropriate response must be selected from a set of stimuli-response pairs active in working memory (Curtis & D'Esposito, 2003; Schwarb & Schumacher, 2012). Oculomotor versions of the SRTT (for which participants show comparable procedural learning to button press response tasks) have shown evidence that procedural learning may occur without the need for a manual response (Kinder et al., 2008; Medimorec et al., 2021a; Vakil et al., 2017). Yet, there is still debate as to whether these experiments capture purely perceptual learning as the involvement of eye movements when following the stimuli on screen may support procedural learning (Willingham, 1999). More recently, Medimorec et al. (2021b) examined whether the inhibition of eye movements by using a fixation target during the learning phase would disrupt spatial sequence learning and observed a reduced proportion of correct anticipations in the restricted version of the SRTT when compared to the standard oculomotor version. Thus, suggesting that learning in this task may not be entirely perceptual since it relies on a motoric response (i.e., eye movements). Crucially, however, the consistent evidence for the independence of procedural learning on the SRTT from a manual response highlights its wider applications in populations for whom a manual response may not be possible.

Types of Serial Reaction Time tasks

The type of sequences used in SRTTs can be deterministic or probabilistic in nature. In deterministic sequences, which tap sequential learning, there are usually two types of blocks – random and sequenced. Stimuli appear in either repeating or random sequences. In the sequenced trials, a pattern is continually repeated, with no clear demarcation of the beginning and end of any sequence; this is done to avoid explicit awareness of the underlying sequence. So called 'random' sequences are in fact quasi-random in that they are at least constrained by the rule that a stimulus cannot immediately reappear in the same stimulus location on the next trial. In typical deterministic sequences, the first blocks contain the repeating sequence, with a sudden switch to a random block (interference effect), followed by a final sequenced block. Reaction times tend to decrease progressively during practice in sequenced blocks but then increase in random blocks, and this is taken as evidence of learning as the participants learn to predict where the next stimulus will appear. Procedural learning in deterministic SRTTs is usually measured as the difference in response time between the performance in the adjacent sequenced and random blocks.

More recently, two variants of the SRTT have been created for measuring sequential and statistical learning (Kóbor et al., 2020; Nemeth et al., 2010; Song et al., 2008; Takács et al., 2021), as unlike in the deterministic SRTT, in probabilistic and alternating SRTTs, the task contains both sequential and statistical information. Thus, whilst in the deterministic SRTTs the underlying sequence is repeated within a block, in probabilistic and alternating SRTTs elements of the sequence are interleaved with random trials. Potentially a consequence of the decreased salience of the underlying sequence, no, or less, evidence of explicit awareness is observed for these variants, thus, these are thought to be purer measures of procedural memory when compared to deterministic SRTTs (Kóbor et al., 2020; Nemeth et al., 2010; Nemeth, Janacsek, & Fiser, 2013; Song et al., 2007; Stark-Inbar et al., 2017; Takács et al., 2021; Vékony et al., 2022). Furthermore, alternating and probabilistic SRTTs allow for procedural learning to be tracked across time as patterned and random trials are interleaved throughout the task. Thus, procedural learning in these tasks is likely less confounded with fatigue as compared to deterministic SRTTs, given that learning is not computed as a difference in later blocks (Pan & Rickard, 2015). Finally, alternating and probabilistic versions of the SRTT show better psychometric properties than deterministic ones, especially for the probabilistic versions, which thus far have been found to present the best psychometric properties in adults (Kalra et al., 2019; Stark-Inbar et al., 2017; West, Shanks, et al., 2021).

In probabilistic SRTTs, participants are exposed to two second-order conditional sequences, one that occurs with a higher probability than the other (e.g., sequence A (85%): 121432413423; sequence B (15%): 323412431421; Siegelman & Frost, 2015). In these tasks, trials containing elements of the probable sequence are intermixed with trials from the improbable sequence within each block of trials. The signal-to-noise ratio often varies between studies. Each block starts with a random bigram (e.g., 43) and the next location cued will be either the location that followed that bigram in sequence A (i.e., 2) or the location that followed that bigram in sequence B (i.e., 1). Procedural learning in probabilistic SRTTs is most commonly measured as the simple difference in response times between improbable and probable trials. Although for a discussion of the various ways in which the procedural learning effect can be calculated, see section *Reliable measures*.

A third variant of the SRTT is the alternating SRTT, in which sequenced stimuli alternate with random elements. Thus, the location of the second stimulus on screen is determined randomly. If, for instance, the sequence is 1243, with each number representing a distinct location on screen, in the alternating SRTT the sequence of stimuli would be 1r2r4r3, with r representing a random stimulus. With practice, individuals respond more quickly to high-frequency triplets (e.g., 1r2) than to low-frequency triplets (e.g., r2r). Procedural learning in alternating SRTTs is measured as the difference in response times to high-frequency triplets compared to low-frequency triplets. This type of sequence is considered a probabilistic second-order conditional since the determination of which stimulus (n) will follow is based on element n-2.

Alongside manipulations of the nature of the sequence, the sequences used in the SRTT can also vary in length (usually 8 to 12 items long) and structure, that is, in the nature of the dependencies. In a sequence with first-order dependencies, learning to predict the following event can be based on the information contained in the previous element (n-1), whilst second-order conditional sequence learning requires knowledge of the two previous positions (n-2). Longer sequences (e.g., D. V. Howard & Howard, 1989; D. Y. Howard & Howard, 1992) and higher-order transitions are typically more difficult to acquire than shorter and lower order transitions (e.g., Deroost, Kerckhofs, Coene, Wijnants, & Soetens, 2006; Deroost & Soetens, 2006b; Reed & Johnson, 1994; Remillard & Clark, 2001; Soetens, Melis, & Notebaert, 2005).

Evidence for the SRTT as a measure of procedural memory

Support for the use of SRTTs as indices of procedural memory has come from a broad array of behavioural, neuroimaging, and clinical studies. First, the SRTT elicits robust procedural learning effects and, due to the nonverbal nature of the SRTT, this task has the potential to be used across development and with clinical populations. Even though most research using the SRTT has focussed on populations aged 4 years and older (Adi-Japha, Badir, Dorfberger, & Karni, 2014; Howard, Howard, Dennis, & Kelly, 2008; Janacsek, Fiser, & Nemeth, 2012; Zwart et al., 2017), new evidence using the oculomotor version of the SRTT has allowed for the examination of the developmental trajectory of procedural learning from infancy to adulthood, thus overcoming limitations imposed by the reliance on motor responses (Koch et al., 2020). This suggests that the SRTT may be useful to document and compare the developmental trajectory of procedural memory across the lifespan, irrespective of the motor and verbal abilities of the population of interest.

Second, there is consistent evidence demonstrating that adequate learning on the SRTT involves the recruitment of the basal ganglia. A recent functional neuroanatomical meta-analysis of SRTT studies (Janacsek et al., 2020) has found evidence of basal ganglia activation for the procedural learning effect, in the more anterior than posterior portions of the striatum, namely in the anterior/mid caudate and putamen and in the globus pallidus. On the other hand, the visual and motor elements of the SRTT were linked to cerebellar and premotor activation.

Finally, beyond converging evidence to support the view that learning on the SRTT is dependent on the intact functioning of the basal ganglia (Kandel et al., 2012; Packard & Knowlton, 2002; Parent & Hazrati, 1995; Ullman, 2004), given its flexibility and nonverbal nature, the SRTT task has been widely used in typical and atypical populations. In atypical populations, the SRTT has allowed for a better understanding of their behavioural and neuroanatomical basis. In populations with amnesia, Alzheimer and Korsakoff's syndrome preserved performance on the SRTT lent support to hypotheses of spared procedural memory in the face of damage to the declarative memory system (Nissen and Bullemer, 1987; Nissen et al., 1989; Reber and Squire, 1994, 1998; Vandenberghe et al., 2006, Knopman, 1991). Conversely, in individuals with Parkinson's and Huntington's disease, populations characterised by damage to the basal ganglia, deficits in procedural learning on the SRTT are observed even when using a verbal version of the task (Westwater et al., 1998; Sommer et al., 1999; Smith and McDowall, 2006; Clark, Lum & Ullman, 2014; Doyon et al., 1997, 1998; Molinari et al., 1997; Shin and Ivry, 2003).

Tracking the stages of procedural memory via the SRTT

Brain imaging studies have suggested that procedural learning in the SRTT depends on different neural correlates, which are preferentially activated during different stages of learning (Doyon & Benali, 2005; Simor et al., 2019; Tzvi et al., 2015). An initial fast acquisition stage begins with the first exposure to the stimuli/task. This stage is characterised by a rapid improvement in performance which may be observed as faster response times and better accuracy. At this stage, cortico-striatal and cortical-cerebellar anatomical systems are crucial, with recruitment of the striatum, the cerebellum, and motor cortical regions as well as the prefrontal, parietal, and limbic areas (Doyon et al., 2009). During earlier stages, anterior portions of the striatum tend to show more activation, whilst the opposite pattern is observed for posterior portions (Doyon et al., 2009). Ullman et al. (2020) hypothesise that this anterior/posterior striatal distinction may reflect the activation of parallel circuits, as anterior striatal circuits are thought to support motivation, working memory, and executive functions. On the other hand, posterior striatal structures underlie motor and visual learning (Doyon et al., 2009).

This initial stage is followed by a decrease in the learning rate and a trend towards an asymptote (Doyon et al., 2009; Verstynen et al., 2012), possibly indicating that learning saturation is a necessary stage for "consolidation" to occur (Hauptmann & Karni, 2002; Hauptmann et al., 2005; Karni et al., 1998). Consolidation refers to the process involved in increasing the robustness and resistance to interference of the initial memory (see Doyon et al., 2009; Robertson, Pascual-Leone, & Press, 2004). This intermediate stage is measured as the contrast between performance at the end of the session and performance at the beginning of a subsequent session, without further practice between the two sessions. This offline period can be expressed by an improvement in performance (behaviourally reflected in a larger procedural learning effect in the subsequent session) or by memory stabilisation (reflected in similar performance at both time points) (Robertson, Pascual-Leone, & Miall, 2004).

It is still not clear whether acquired memories are stabilised or enhanced during offline periods (Pan & Rickard, 2015; Peigneux et al., 2006; Simor et al., 2019), yet some evidence suggests that this process might differ for sequence and statistical learning. For deterministic sequences, enhanced procedural learning has been observed after an offline sleep period; conversely, learning in probabilistic versions of the SRTT does not appear to benefit from post-learning offline sleep periods (King et al., 2017; Nemeth et al., 2010; Robertson, Pascual-Leone, & Press, 2004). Whilst this pattern of findings may be explained by the nature of the sequence, two caveats need to be considered. Firstly, probabilistic SRTTs are not pure measures of statistical learning as they also contain sequential information. Secondly, evidence suggests that memory consolidation may be influenced by explicit awareness of the sequence. Robertson, Pascual-Leone and Press (2004) manipulated the mode of learning (explicit or implicit) in the SRTT and whether the offline period included sleep or awake. Sequence learning was measured before and after the 12-hours interval. The explicit groups only showed offline improvements when sleep was included, with overnight consolidation correlating with the amount of NREM sleep. For the implicit groups, on the other hand, offline learning occurred irrespective of whether the interval included a period of sleep or awake. Beyond the role of sleep in the consolidation of explicit knowledge of the sequence, sleep has also been found to favour the emergence of explicit awareness when learning occurs in implicit conditions (Fischer et al., 2006; Wilhelm et al., 2013). More recently, Simor et al. (2019) aimed to explore the training-dependent and offline changes involved in sequence and statistical learning using a modified version of the alternating SRTT. In this task, the authors were able to isolate both statistical, computed as the difference between random low and high-frequency triplets, and sequence learning, indexed as the difference between random and patterned high-frequency triplets. After training, 78 adults were assigned to one of three conditions: active wakefulness, quiet rest, and daytime sleep. During training, sequence learning increased throughout, whilst statistical learning plateaued rapidly. After the offline period (1 hour), irrespective of vigilance state, sequence and statistical knowledge was preserved, with no significant offline changes. Post-offline training led to improvements in sequence learning but not statistical learning. It is unclear whether the lack of differences in offline gains between sequence and statistical learning reflects the shorter interval between test and retest; thus, further evidence is required to determine whether distinct consolidation processes are involved in the acquisition of statistical and sequential regularities.

The final stage of procedural learning involves the automaticity of the learned skill, which allows the behaviour to be performed effortlessly, demanding fewer attentional resources. This can be observed in dual-task conditions, where the interference of a second task is attenuated when the learned skill has been automatised (Seger & Spiering, 2011). Whilst in the early, fast stage of skill acquisition the cerebellum plays a crucial role, in the later stages there is a decrease in its involvement, with the striatum being more actively involved (Doyon & Benali, 2005).

Neuroimaging studies have examined the time course of procedural learning on the SRTT task and have observed that brain region activation varies substantially depending on the learning stage (Doyon et al., 2009; Doyon & Benali, 2005; Tzvi et al., 2015). Doyon and colleagues (2002) conducted an fMRI study analysing the performance of healthy adults in an SRTT across three separate sessions. Changes during the course of the sessions were observed in the activation in the cerebellar cortex and deep cerebellar nuclei (dentate nucleus). Whilst activity in the deep cerebellar nuclei increased from session 1 to session 2, activation in the cerebellar cortex diminished. Thus, the recruitment of the cerebellar cortex appears to contribute primarily at the early stages of procedural learning in the SRTT, having a less predominant role in motor sequence consolidation. From session 2 to session 3, there was evidence of sequence learning becoming gradually more dependent on the neocortex (right striatum, SMA, inferior parietal, precuneus, and ventrolateral prefrontal cortex) and less dependent on the cerebellar-cortical circuit.

To conclude, even though later stages of procedural learning tend to be less frequently examined, there is evidence suggesting that procedural learning abilities transition through distinct stages, from an initial stage of rapid improvements in performance to a later stage of automatisation. These stages can be detected both at the behavioural and neural levels and are primarily modulated by the amount of practice (Hauptmann et al., 2005).

Procedural/declarative model

Having provided an overview of procedural and declarative memory, we start by outlining the procedural/declarative model and the procedural deficit hypothesis, before reviewing the evidence in subsequent sections. Based on the multiple systems view of memory, Ullman and colleagues (Ullman, 2001a, 2001b, 2004, 2016b, 2016c; Ullman et al., 2020) proposed that the declarative and procedural memory systems differentially support language processing and learning. Unlike modular theories of language acquisition (Chomsky, 1980; Fodor, 1983; Pinker, 1994), the declarative/procedural model takes a domain-general approach and proposes that language development relies on pre-existing neural networks or regions involved in other cognitive functions (Earle & Ullman, 2021; Hamrick et al., 2018; Ullman, 2016b). According to the procedural/declarative model, the declarative memory system encodes, stores, and retrieves arbitrarily related information, such as the content of the mental lexicon (i.e., word forms and their meanings). In line with multiple memory systems theory, declarative memory is responsible for idiosyncratic and rapidly acquired knowledge which is thought to be initially reliant on the medial temporal lobe and its associated circuitry (Zola-Morgan & Squire, 1991). In contrast, the procedural memory system is proposed to underlie the learning and processing of rule-based knowledge and is thought to rely on the frontal and basal ganglia structures. The procedural memory system is hypothesised to be involved in the acquisition of syntax (i.e., the set of rules for how linguistic units are combined to convey meaning; van der Lely & Pinker, 2014), morphology (pertains to how words and morphemes are combined to form new words; van der Lely & Pinker, 2014), and phonology (refers to the abstract representations and rules that allow for the combination of sounds into more complex linguistic structures such as morphemes; van der Lely & Pinker, 2014), which are acquired through gradual exposure to statistical, probabilistic and sequential structures (Ullman, 2001a, 2001b, 2004, 2016b, 2016c; Ullman et al., 2020). Other aspects of language may also be learnt through the procedural memory system, such as word boundaries in a speech stream, articulatory knowledge, and speech perception (Ullman et al., 2020).

In Ullman (2001b) multiple lines of evidence focussing on the formation of past tense in the English language are presented to support the dissociable role of the declarative and procedural memory systems in language acquisition. Psycholinguistic evidence has demonstrated frequency effects for irregular but not regular past forms, which is taken to indicate the real-time demands of the composition of regulars (Ullman, 2001b). Evidence from developmental disorders, such as DLD and William’s disease, and neurodegenerative diseases such as Aphasia, Alzheimer’s disease, Parkinson’s disease, and Amnesia, previously reviewed, have been instrumental given the pattern of dissociations and associations that provide evidence for the dissociation between declarative and procedural memory in language (Ullman, 2001b). Specifically, evidence has been widely explored in the formation of past tense, whereby individuals with Parkinson’s disease often show impaired regular past tense, despite preserved irregular past tense production. The opposite pattern is observed in individuals with Alzheimer’s disease (Pinker & Ullman, 2002). Thus, procedural memory is argued to be involved in the acquisition of regular past forms through the assimilation of probabilistic information, whilst the acquisition of irregular past tense forms depends on the declarative system given its reliance on more arbitrary associative learning (Pinker & Ullman, 2002; cf. Plunkett & Marchman, 1991; Rumelhart & McClelland, 1987).

Despite the emphasis on the role of declarative and procedural memory systems for language acquisition, this model does not preclude other neural structures and cognitive or computational components from being involved in language acquisition, nor does it suggest that the declarative and procedural memory systems subserve only language (Ullman, 2001a). Similarly, this division of labour does not prevent the two systems from acquiring the same linguistic knowledge, since, similarly to non-linguistic domains, some redundancy between systems is expected to occur. Thus, given the learning flexibility of the declarative memory system, this long-term memory system is expected to be involved in the acquisition of grammar through the learning (both explicitly and implicitly) and storage of the rules through associative learning mechanisms such as chunking (Ullman & Pullman, 2015).

Given the later maturational trajectory of the neuroanatomical networks involved in declarative memory, first language acquisition is expected to rely more on the procedural memory system (Finn et al., 2016; Ullman, 2015). As the declarative memory system develops from childhood up to adulthood, language learning becomes increasingly reliant on this memory system. Whilst at least some degree of proceduralisation of the second language knowledge is thought to occur with further practice, evidence suggests that this process, even after years of exposure, may not result in the same levels of attainment (Hamrick et al., 2018; Ullman, 2015).

Aligning with the dynamic nature of this model, the dependence on either system has been found to be modulated by several factors, including the age of the participants, the nature of the task, and the instructions. Regarding the effect of age, whilst declarative memory is expected to improve from childhood to adulthood, the procedural memory system reaches adult-like levels of functioning earlier in development as its underlying structures tend to mature earlier than those involved in declarative memory (Bauer, 2008). In line with this, Finn et al. (2016) compared the performance of 10-year-old children and adults in a battery of declarative and procedural memory tasks and observed that children showed less evidence of declarative learning than adults, yet both age groups exhibited equivalent procedural learning. Secondly, the nature of the task and instructions may encourage greater engagement of one system other than the other (Destrebecqz & Cleeremans, 2001; Stefaniak et al., 2008). For example, tasks with explicit instructions, which raise participants' awareness to the patterns in the input, are expected to show increased learning in the declarative memory systems (Schendan et al., 2003). On the other hand, incidental and dual-task paradigms would be expected to exert the opposite effect (Ashby & Maddox, 2011; Hazeltine, 1997; Ullman et al., 2020).

Procedural deficit hypothesis

Given the postulated role of the procedural and declarative systems in language acquisition, Ullman and colleagues (e.g., Ullman, 2014; Ullman et al., 2020; Ullman & Pierpont, 2005) extended this model to account for the pattern of linguistic - as well as non-linguistic - impairments observed in developmental language disorder and dyslexia. Before outlining this model, we first turn to consider the behavioural and cognitive features of these neurodevelopmental disorders.

DLD and dyslexia are neurodevelopmental disorders with unknown aetiology (Catts et al., 2017; Bishop, 2017) that affect the typical development of language and literacy skills, respectively (Bishop & Hayiou-Thomas, 2008; Peterson & Pennington, 2012). These profiles are present in approximately 3% - 10% of the school-aged population (DLD: Tomblin et al., 1997, dyslexia: Astle et al., 2019; Butterworth & Kovas, 2013). Beyond these difficulties, these populations often show broader non-linguistic impairments, including in executive function (DLD: Snowling et al., 2019; dyslexia: Romani et al., 2011) and working memory (DLD: e.g., Baird et al., 2010; dyslexia: e.g., Fostick & Revah, 2018).

Several theories have proposed causal explanations for the difficulties experienced by children with DLD and dyslexia (see Ramus, 2003). The phonological theory (M. Snowling, 2000), one of the most prominent theories of dyslexia, suggests that individuals have a specific impairment in the representation, storage, and retrieval of speech sounds. Support for the phonological theory comes from phonological awareness tasks, where participants are required to segment or manipulate speech sounds. A meta-analysis conducted by Melby-Lervåg et al. (2012) observed a large deficit in phonemic awareness in children with dyslexia when compared to aged-matched TD children (pooled effect size estimate: -1.37) and reading-matched children (pooled effect size estimate: -0.57). Furthermore, performance on phonemic awareness tasks strongly correlated with word reading abilities.

Language specific accounts of DLD, on the other hand, have suggested that this disorder may be due to a domain-specific grammatical impairment (Van Der Lely, 2005; van der Lely et al. 1998; van der Lely & Pinker, 2014). This theory has primarily focussed on the difficulties experienced by children with DLD with morphology and syntax, with meta-analytical evidence (Krok & Leonard, 2015) suggesting that children with DLD show less accurate regular (g = -1.801) and irregular (g = -0.805) past tense production abilities than typically developing children across Germanic languages. Despite considerable support for these, and other non-linguistic (e.g., deficits in general speed of processing: e.g., Breznitz & Misra, 2003; auditory processing: e.g., Goswami et al., 2011; Tallal & Benasich, 2002; phonological working memory: e.g., Gathercole & Adams, 1993; Gathercole & Baddeley, 1993), explanatory theories of dyslexia and DLD, most fail to account for the secondary non-linguistic impairments and the highly heterogeneous profiles presented by these populations (Ullman & Pierpont, 2005).

The procedural deficit hypothesis, proposed by Ullman and Pierpont (2005), and based on the multiple memory systems model, posits a dysfunction in the brain systems responsible for procedural memory as a causal mechanism of both the language learning impairments in developmental language and literacy disorders, and the non-linguistic impairments that are reliant on procedural memory circuitry. Based on the profile commonly observed in children with DLD and dyslexia of impaired grammatical and phonological skills while vocabulary is relatively spared (Hsu & Bishop, 2014), this hypothesis proposes that the declarative memory abilities of children with dyslexia and DLD remain relatively intact; while an impairment in the procedural memory system can account for the cognitive and behavioural profile of children with DLD and dyslexia, as well as weaknesses in motor skills and working memory often exhibited by these populations (Ullman & Pierpont, 2005). Abnormalities of different structures and dysfunction of different portions of the structures of the procedural memory circuitry are proposed to be associated with distinct behavioural manifestations, thus accounting for the heterogeneity within dyslexia and developmental language disorders, but also potentially explaining the comorbidity between neurodevelopmental disorders due to overlapping abnormalities of procedural memory circuitry (Ullman et al., 2020; Ullman & Pierpont, 2005).

Additionally, and in accordance with the declarative-procedural model, the procedural deficit hypothesis proposes that both systems can interact competitively and cooperatively in the learning process, which may lead the preserved memory system to compensate for the dysfunctional system. Therefore, in neurodevelopmental disorders, the declarative memory system may perform a compensatory role when procedural memory is affected. The procedural deficit hypothesis (Ullman & Pierpont, 2005) predicts that, firstly, individuals with neurodevelopmental disorders should at least partly compensate for their procedural memory impairments by adopting cognitive and behavioural strategies reliant on the relatively stronger declarative system. Despite the flexibility of the declarative memory system, the ease with which this memory system compensates for the procedural memory dysfunction may vary. For example, declarative memory would more easily compensate for skills that rely on local (e.g., successive dependencies between determiner and noun) rather than long-distance (e.g., subordinating conjunctions and their modal verbs) temporal dependencies (Purdy et al., 2014). Secondly, the increased reliance on declarative memory should translate into increased activation of its neural structures, when probing tasks expected to have been compensated by this memory system or at the earlier stages of adoption of new strategies. Finally, individuals with better declarative memory are expected to more successfully compensate for their procedural memory impairments (Ullman & Pierpont, 2005). In the context of language acquisition, individuals with procedural memory impairments may rely on declarative memory to compensate for their linguistic and literacy skills by relying on explicit rules as well as other strategies such as chunking (Ullman, 2004; Ullman & Pierpont, 2005).

The predictions of the procedural deficit hypothesis have been assessed using various experimental paradigms purporting to measure procedural learning, yet it is unclear whether the populations with dyslexia and/or DLD would be expected to show impairments across all these tasks

(Hsu & Bishop, 2014). For the purpose of this review, we will focus on the evidence from the SRTT for the reasons previously highlighted including the consistent evidence from neuroimaging and clinical studies showing the involvement of the basal ganglia when performing the SRTT; more information about its psychometric properties (reviewed in the section *Reliable measures*); and its non-verbal nature. However, it should be acknowledged that whilst the procedural learning effect measured by the SRTT requires the learning of probabilistic and/or sequential information it likely does not tap into all computations required for language and literacy acquisition (Bogaerts et al., 2021), thus it is possible that procedural memory in the SRTT does not correlate with all measures of language and literacy.

Evidence for the procedural/declarative model and the procedural deficit hypothesis

This section considers the evidence for the procedural deficit hypothesis as applied to DLD and dyslexia. First, we examine evidence at the biological level (i.e., from neuroimaging studies) given the clearer predictions by the procedural deficit hypothesis (Ullman et al., 2020; Ullman & Pierpont, 2005) for abnormalities in the brain structures underlying procedural memory, namely the cortico-basal ganglia-thalamo-cortical circuitry in dyslexia and DLD, although it is worth noting that the direction of the difference between typically developing and disordered groups is not clear. Furthermore, a mapping between structural and functional abnormalities and these disorders is still missing, instead (Ullman et al., 2020) clarify that the location(s), severity, and extent of the neuroanatomical abnormalities may vary within and across disorders. We then move on to consider the procedural deficit hypothesis model as applied to the broader population by considering group-level and individual differences studies. At the group-level, the procedural deficit hypothesis (Ullman, 2014; Ullman et al., 2020; Ullman & Pierpont, 2005) predicts that individuals with dyslexia and DLD will underperform when compared to TD controls in experimental paradigms that tap into procedural memory. Yet, it is unclear whether these populations are expected to show impaired performance in all tasks. Instead, given Ullman's suggestion (Ullman et al., 2020) for potential differences within and across disorders in the neuroanatomical profile, it is likely that these populations may show impaired performance in some but not all tasks. Given the consistent evidence for the involvement of the cortico-basal ganglia-thalamocortical circuitry in the SRTT as previously described (section *Serial Reaction Time Task*), this task is potentially a good candidate for evaluating the validity of the procedural deficit hypothesis (Conway et al., 2019; Williams, 2020). Finally, the predicted role of procedural memory in the acquisition of language and literacy abilities thought to rely on the ability to extract and manipulate probabilistic knowledge will be examined in typical and atypical populations. Whilst the procedural/declarative model (Lum et al., 2012; Ullman & Pierpont, 2005) postulates a positive association between language and literacy and procedural memory in TD populations, the predictions for atypical populations are more convoluted, given the possible compensatory role of declarative memory. Thus, the same positive association between language and literacy and procedural memory is expected to occur in populations with dyslexia and DLD, unless their language and literacy deficits have been compensated for by the declarative memory system, resulting in a null association between these abilities (Lum et al., 2012). Specifically, if procedural memory is severely impaired in this population, as predicted by the procedural deficit hypothesis, then associations between procedural learning and language/literacy measures are not expected since these populations have relied on the declarative memory system for language and literacy acquisition. Rather, language and literacy abilities which would be expected to rely on the procedural memory system (e.g., grammar, phonology) would instead be expected to correlate with declarative memory.

Neuroimaging evidence

Development language disorder

The structural and functional neurological underpinnings of DLD remain poorly understood relative to other neurodevelopmental disorders. It has long been agreed, however, that the brain of individuals with DLD is characterised by more subtle structural differences (e.g., in volume) as opposed to gross lesions (Jernigan et al., 1991). A systematic review conducted by Liégeois et al. (2014) which included four structural neuroimaging studies using voxel-based morphometry reported evidence for cortical abnormalities within the temporal region in participants with language disorders. However, the direction of the abnormalities was inconsistent across studies, whilst reduced grey matter was observed by Badcock et al. (2012) in the right posterior superior and middle temporal gyri and left posterior superior temporal sulcus, Soriano-Mas et al. (2009) found an increase in volume within a posterior “perisylvian” area extending from the posterior superior temporal gyrus to the angular and the supramarginal gyrus. Consistent with the procedural deficit hypothesis, atypical development was also observed in the basal ganglia, specifically in the caudate nucleus (reductions: Badcock et al., 2012; Lee et al., 2013; increase: Soriano-Mas et al., 2009) and putamen (Lee et al., 2013). Abnormalities in the basal ganglia have also been reported in other studies not included in the systematic review (e.g., Herbert et al., 2003; Jernigan et al., 1991; Watkins et al., 2002).

Further support for the role of structural abnormalities and language proficiency comes from the study by Lee et al. (2013) where a negative correlation was observed between a composite language measure and the relative volume of basal ganglia substructures, namely putamen, the globus pallidus and the nucleus accumbens with larger substructures being associated with poorer language scores. Similarly, volumetric reductions in most of the cortical and subcortical regions examined by Lee et al. (2013) were associated with poorer language abilities, yet there was no evidence for a reduction in fractional anisotropy - a measure of connectivity in the brain that assesses the degree of anisotropic diffusion of water molecules - in the caudate or putamen. These findings are in line with those obtained by Verhoeven et al. (1999). Thus, whilst an atypical brain structure appears to be frequently observed in the temporal regions and cortico-basal ganglia-thalamo-cortical circuitry, lending support for the procedural deficit hypothesis, the direction of the difference varies between studies, likely due to sampling and methodological differences.

Beyond structural abnormalities, the procedural deficit hypothesis predicts functional differences between TD and DLD populations in the same regions, yet the direction of the activation abnormalities is not yet clear. The functional imaging studies included in Liégeois et al. (2014) reported hypoactivation of the posterior superior temporal gyrus (Badcock et al., 2012; de Guibert et al., 2011), as well as right-sided hyper-activation within the right insula extending to the inferior frontal gyrus-pars opercularis/pars triangularis, and caudate head during a phonological task (de Guibert et al., 2011), thus providing some evidence for functional abnormalities in the basal ganglia as predicted by the procedural deficit hypothesis.

Noting the small sample sizes that characterise many functional neuroimaging studies of DLD, as well as variability in the inclusion/exclusion criteria, Krishnan and colleagues (2021) conducted an fMRI experiment using a simple verb generation task in large samples of children aged 10–15 years (DLD N = 50, typically developing N = 67). In this well-powered study, the grammatical task evoked activity in the same regions and to a similar degree in both TD and DLD children. There was no evidence of group differences in the left inferior frontal gyrus and putamen bilaterally. In a subgroup of children (N = 14), with the poorest performance on the task, sub-threshold group differences were observed in the left inferior frontal gyrus and caudate nuclei. Whilst these findings lend limited support to the procedural deficit hypothesis, the authors acknowledge that the results may have been influenced by the nature of the task (i.e., striatal regions may be involved in more complex linguistic tasks), age of the participants (i.e., there could be developmental differences in the involvement of the striatal circuit) and the nature of the analyses (i.e., more detailed analyses could have been more sensitive to connectivity or microstructural differences in the striatum). Thus, further research is needed to systematically examine these possibilities.

In sum, previous evidence shows some support for structural and functional abnormalities in the basal ganglia as predicted by the procedural deficit hypothesis. Yet, there is variability in the direction of the abnormalities, possibly related to methodological differences. Further research is required to determine whether the neuroanatomical abnormalities depend on the severity, or comorbid disorders (e.g., speech impairments: Silveri, 2021), and are thus not characteristic of the population with DLD more broadly.

Dyslexia

In contrast to the above literature on DLD, the neuroimaging evidence-base for dyslexia has a much longer history, and the pattern of findings has been more consistent (Paulesu et al., 2014; Yan et al., 2021). We start by presenting the neurological underpinnings of reading in typical development. The dual-pathway architecture suggests that two mechanisms are involved in reading, whereby unfamiliar or nonwords are read primarily via alphabetic decoding, whereby the graphemes are mapped to phonemes, in contrast for familiar words meaning is immediately retrieved from spelling without reliance on phonology (Castles et al., 2018 for a detailed review). Neuroimaging and clinical studies provide supporting evidence for this dual-pathway. The meta-analysis conducted by Taylor et al. (2013) proposes two pathways involved in skilled reading: the dorsal pathway involved in phonologically mediated reading and the ventral pathway involved in the direct access to meaning from print. This dissociation is further supported by clinical studies. Patients with damage in the dorsal pathway, which encompasses the parietal lobe, superior temporal gyrus, and inferior frontal gyrus, show poor nonword reading (Woollams & Patterson, 2012). In contrast, those with damage in the ventral pathway, i.e., the ventral occipitotemporal and anterior inferior frontal gyrus regions, show difficulties reading irregular words (e.g., Woollams et al., 2007). Across development, reading gradually switches its reliance from the dorsal to the ventral pathway (B. A. Shaywitz et al., 2002) and this transition is associated with speeded word reading (Castles et al., 2018).

Structural abnormalities in brain areas associated with reading have also been observed in developmental dyslexia. Eckert (2004) reviewed the evidence from structural imaging studies and reported consistent differences between controls and dyslexic participants on the inferior frontal gyrus, temporal-parietal region, the medial occipital lobe, and the cerebellar anterior and posterior lobes. These findings were further supported in a more recent meta-analysis with nine voxel-based morphology studies (Richlan et al., 2013), which observed reduced grey matter volume in individuals with dyslexia compared to controls in the right superior temporal gyrus and left superior temporal sulcus, with some research suggesting that these structural abnormalities may be present before reading onset (e.g., Raschle et al., 2011). In addition, a small body of evidence points to structural abnormalities in subcortical structures; specifically, and in line with the procedural deficit hypothesis, structural abnormalities in the basal ganglia have been reported (putamen: Eckert et al., 2005; Pernet et al. 2009; caudate head: Brown et al. 2001).

One of the most consistent findings in the functional neuroimaging research on dyslexia is the underactivation of the left temporo-parietal and left ventral occipitotemporal brain areas in dyslexia (Devoto et al., 2021; Martin et al., 2016; Raschle et al., 2012). In an activation likelihood estimate (ALE) meta-analysis (Maisog et al., 2008), comprising 96 foci from nine publications, greater activation was found in controls than in individuals with dyslexia on two left extrastriate areas within BA 37, precuneus, inferior parietal cortex, superior temporal gyrus, thalamus, and left inferior frontal gyrus, as well as on the right hemisphere in the fusiform, postcentral, and superior temporal gyri. Conversely, a second ALE meta-analysis (Maisog et al., 2008), which included 75 foci from six papers, reported hyperactivity for the dyslexic participants compared to controls on the right thalamus and anterior insula. These results converge to a degree with the findings from Richlan et al. (2009) in a meta-analysis with 17 studies on 128 foci that examined the patterns of brain activation during reading or reading-related tasks. In this study, dyslexic brains showed underactivation in the inferior parietal, superior temporal, middle and inferior temporal, and fusiform regions of the left hemisphere. Overactivation, on the other hand, was observed in the inferior frontal gyrus which was accompanied by overactivation in the primary motor cortex and the anterior insula. Overall, these meta-analyses point to a pattern of underactivation in the left hemisphere during reading or reading-related tasks. However, there was no clear evidence for overactivation of the right hemisphere temporoparietal regions which have been hypothesised to compensate for the left temporoparietal under-activation (e.g., Shaywitz & Shaywitz, 2005)**.** This disparity is likely due to the distinctions in sampling and methodology, but potentially intensified by the role of orthographic depth on reading proficiency (see (Martin et al., 2016) for a meta-analysis comparing the patterns of activation in dyslexia in deep and shallow orthographies.

Overall, there seems to be only limited evidence for structural abnormalities in the cortico-basal ganglia-thalamocortical circuitry in dyslexia. Instead, functional imaging provides consistent evidence for structural and functional abnormalities in the dorsal pathway, which is associated with phonologically dependent reading.

Behavioural evidence

We now turn to the available evidence at the behavioural level for both procedural learning and consolidation on the SRTT. Given the similar predictions from the procedural deficit hypothesis, the similar pattern of findings for both groups, as well as the inclusion of both disorders in many samples, the evidence for DLD and dyslexia was collapsed.

Learning

Evidence for a procedural deficit in the SRTT has been often observed in children and adults with dyslexia (e.g., Earle & Ullman, 2021; Gabay et al., 2012; J. H. Howard et al., 2006; Jiménez-Fernández et al., 2011; Menghini et al., 2006, 2008; Vicari, 2005) and DLD (e.g., Conti-Ramsden et al., 2015; Hedenius et al., 2011; Hsu & Bishop, 2014; Kuppuraj et al., 2016; Lukács & Kemény, 2014; Lum et al., 2010, 2012; Sengottuvel & Rao, 2013b, 2014; Tomblin et al., 2007) when compared to typically developing individuals. In other studies, however, children and adults with dyslexia (e.g., Du & Kelly, 2013; Henderson & Warmington, 2017; Kelly et al., 2002; Menghini et al., 2010; Perlant & Largy, 2011; Rüsseler et al., 2006; Vakil et al., 2015; Yang et al., 2013; Yang & Hong-Yan, 2011) and DLD (e.g., Gabriel et al., 2011, 2012, 2013, 2015; Lee et al., 2013, 2016; Lum & Bleses, 2012; Mayor-Dubois et al., 2014) showed similar performance on the SRTT to controls. These mixed results may be associated with sampling and methodological differences, as these studies not only differ in the ages of participants, the definitions of dyslexia/DLD, but also in the type of SRTT adopted and the number of trials, amongst other differences. The impact of these factors on the magnitude of the difference between groups has been systematically examined in the following meta-analyses.

Meta-analyses conducted by Lum and colleagues (Lum et al., 2013, 2014) comparing the procedural learning effect of individuals with and without dyslexia and DLD on the SRTT were able to shed some light on the mixed findings. Even though there was considerable variability in the characteristics of the SRTTs used, individuals with dyslexia and DLD demonstrated a significantly smaller difference between random and sequence response times in the SRTT than TD children, with an average mean effect size of 0.449 in the case of dyslexia (95% CI [.204 - .693] (Lum et al., 2013) and .328 for DLD (95% CI [.071 - .58] (Lum et al., 2014), respectively. These group differences provide clear support for the procedural deficit hypothesis. Across meta-analyses (Lum et al., 2013, 2014), the difference in procedural learning between clinical and control groups was moderated by participants' age and the number of exposures to the sequence. Specifically, the difference in the procedural learning effect between groups decreased with participants' age or a larger exposure to the sequence. For dyslexic populations, the interaction between type of sequence and age predicted the size of the effect between individuals with dyslexia and controls, whereby smaller effects were observed for older participants when a second-order conditional sequence was used; that is, perhaps surprisingly, individuals with dyslexia showed a smaller deficit for the more complex second-order conditional sequences (Lum et al., 2014). The age effect is thought to reflect compensatory mechanisms of the declarative memory systems or a delay in the maturation of the procedural memory system. The decrease in the differences between groups due to increased exposure to the sequence suggests that individuals with DLD and dyslexia may require further practice to show comparable learning to TD populations.

These findings were replicated for the DLD group, as participants with DLD were also reported to underperform when compared to controls (medium effect size: Hedge's g of 0.46) in the meta-analysis conducted by Obeid et al. (2016) using a larger dataset that was not restricted to the SRTT. More recently, West, Melby-Lervåg et al. (2021), in a meta-analysis with 610 participants with developmental language disorder and dyslexia and 698 TD participants, revealed a procedural learning impairment on the SRTT in the clinical groups when compared to TD controls (*g* = -.30), yet the size of the difference between groups did not differ between dyslexia (*g* = .28) and DLD participants (*g* = .33). In line with the findings from Lum et al. (2013, 2014), West, Melby-Lervåg et al. (2021) observed a larger, but not significant, group difference between disordered and TD groups in children than adults. Unlike the meta-analyses conducted by Lum and colleagues which only included deterministic SRTTs, West, Melby-Lervåg et al. (2021) reported a significant moderator effect of type of SRTT, whereby the overall effect size in deterministic SRTTs (*g* = .28, 95% CI [.42, .14], k = 22) was higher than for alternating SRTTs (*g* = .04, 95% CI [.33, .24], k = 4). Whilst this finding may point to computational differences between these versions of the SRTT, where the deterministic SRTT taps primarily into sequential learning and the latter taps into both sequential and statistical learning, the disparity in the sample size suggests that further research is needed to understand these results. Finally, unlike previous studies, there was no evidence for a moderating effect of sequence length or the number of sequence repetitions for the deterministic SRTTs. Even though these findings contrast those from Lum and colleagues (Lum et al., 2013, 2014), the moderating effect of the number of exposures in the dyslexic population (Lum et al., 2013) only emerged in the interaction term with age, thus, suggesting the effect of these variables may differ according to participants' age.

In summary, despite some degree of variability at the study level, meta-analyses show evidence for impaired procedural learning on the SRTT as predicted by the procedural deficit hypothesis.

Consolidation

Whilst procedural memory pertains to the learning, consolidation, and automatisation of skills and habits after multiple exposures to stimuli (L. R. Squire & Zola, 1996), most studies have examined learning within a single session, as opposed to across multiple sessions. Furthermore, the few studies that have examined consolidation in dyslexia and DLD have produced a mixed pattern of results.

To our knowledge, four studies have examined the consolidation of procedural memory on the SRTT in children with dyslexia (Gabay et al., 2012a; Hedenius et al., 2013, 2021; Henderson & Warmington, 2017). In Hedenius et al. (2013), twelve children with dyslexia and 17 typically TD children completed the alternating SRTT in two separate sessions, 24 hours apart. This study aimed to determine whether TD and dyslexic groups differ in learning and offline consolidation. Analysis of RTs revealed no evidence of group differences in overall RTs or procedural learning at each point in the learning phase examined: an early learning stage (blocks 1–10), an intermediate learning stage (blocks 11–20) and a late learning stage (blocks 21–25), which occurred after an overnight interval and consolidation, where the procedural learning effect on the final block was compared to the first block on the second day. For accuracy, both groups showed similar accuracy levels across the task. However, despite no significant group differences in the first practice session, children with TD showed a larger procedural learning effect in the second session after extended practice than the dyslexic group. This difference was not accounted for by group differences in offline consolidation of the learned sequence (i.e., comparison of the final block (85 trials) of session 1 and the first block (85 trials) of session 2), instead appearing to suggest that TD children benefited to a great extent from further practice in session 2. Procedural learning accuracy scores for the second session for both groups were significantly associated with reading ability (*r* = .470).

Hedenius et al. (2021) extended these findings by examining whether prolonging the exposure to the sequence by 50% in session 1 would promote consolidation of the underlying pattern in the alternating SRTT in children with dyslexia. Using a similar design, 31 children with dyslexia and 34 TD children were again asked to perform the alternating SRTT in two separate sessions, with a 24-hour interval. As before, there were no group differences between groups on overall RTs or accuracy in any of the sessions and groups showed comparable procedural learning in session 1 (measured by the difference in response times between sequenced/non-sequenced trials) as evidenced by a non-significant group by block interaction. However, at the 24-h follow-up session, there was a significant group difference in the procedural learning effect suggesting that extended exposure to the sequence was insufficient to reduce group differences in consolidation at the follow-up test. Unlike the previous study, when comparing the last epoch of session 1 and the first epoch of session 2, children with dyslexia showed a decrease in procedural learning, while TD children showed the opposite pattern. In a separate model, inattention symptoms were entered as a covariate to determine whether differences in attention between groups could explain the group differences in procedural learning in session 2. There was no effect of attention on the procedural learning effect. One explanation for these data is that the children with dyslexia showed weaker consolidation than their typical peers (consistent with reports of weaker consolidation in other domains, such as word learning: F. R. H.

Smith et al., 2018). Other studies, however, have reported no evidence of group differences in consolidation following procedural learning in samples of individuals with dyslexia (Gabay et al., 2012 (day 1 and 24h later); Henderson & Warmington, 2017 (days 1, 2 and 8)) when compared to TD adults in a deterministic SRTT. This suggests that the differences in follow-up sessions observed by Hedenius et al. (2021) could also reflect relearning/practice during the first epoch of the follow-up session rather than consolidation processes. Though this is unlikely given that there were no group differences in the procedural learning in the first session. Instead, this discrepancy may reflect developmental differences, given evidence from previous meta-analyses (Lum et al., 2013; West, Melby-Lervåg, et al., 2021) showing that differences between individuals with dyslexia and TD controls in procedural learning decrease with age. Hence, if the same pattern occurs for later stages of procedural learning, children would be expected to show more marked difficulties with consolidation on the SRTT. However, there is currently very little empirical evidence examining this possibility directly.

Few studies have investigated the procedural learning abilities of individuals with DLD across sessions. Similarly to the findings by Hedenius et al. (2013, 2021), other studies (Desmottes et al., 2016; Desmottes, Maillart, et al., 2017; Hedenius et al., 2011) observed that children with DLD showed comparable learning in the earlier stages of procedural learning (session 1), yet showed poorer gains with further practice. Whilst in Desmottes et al. (2016), children with DLD (N = 42) only demonstrated less gains than TD children 1-week post-training, in Desmottes, Maillart, et al. (2017) gains were only observed in the TD group at the 24h and 1-week follow-ups. As before, it is unclear whether these group differences reflect differences in offline learning or reduced gains from further practice. Evidence pointing to differences between groups in offline changes in procedural learning come from Hedenius et al. (2011) where children with DLD showed a trend towards loss of procedural knowledge between the first and second sessions, approximately 3 days apart, whilst children with TD improved between sessions. This pattern was further accentuated when comparing children based on their performance on tests of grammatical ability. Children with grammatical impairments not only lost procedural knowledge between sessions, but the procedural learning effect on session 2 was comparable to that shown in the first epoch (out of four) of session 1, thus suggesting that grammar impaired children might have lost most of the procedural knowledge previously acquired.

Whilst these studies seem to point to somewhat preserved procedural learning at earlier stages, another study conducted by Earle and Ullman (2021) reported poorer procedural learning in adults with DLD across sessions. Differences in study design may explain this difference, as children in previous studies had considerably more practice in the first session than those included by Earle and Ullman (2021). This hypothesis is in line with the findings from Lum et al. (2014), which suggest that the size of the difference between groups was negatively moderated by the amount of practice in the first session. However, it fails to explain why children with DLD underperform in follow-up sessions despite further practice, leaving open the possibility that differences in consolidation may at least contribute to these effects.

Interim Summary

Despite mixed evidence from individual studies for an impairment in procedural learning in populations with dyslexia and DLD, group-level meta-analyses to date have consistently observed differences between disordered and TD groups in procedural learning. This discrepancy likely reflects the poor psychometric properties of the SRTT, whereby correlations between procedural learning and language/literacy are attenuated due to measurement error. We analyse this issue in more detail in the section *Reliable measures*. Whilst the evidence for procedural learning impairments in DLD and dyslexia is insufficient to determine whether procedural memory is involved in the development of language and literacy skills, at present, there is only partial support for the procedural deficit hypothesis. Yet, it is premature to interpret these group differences as suggesting that procedural memory has a role in language/literacy development and disorders, especially given that the SRTT is not a process-pure measure of procedural memory, and thus these group differences may be better explained by a third variable (e.g., attention: West, Shanks, et al., 2021). Therefore, in the next section, we examine the relationship between procedural memory and language and literacy across populations.

Individual differences evidence

Following from the declarative/procedural model, if the ability to extract and assimilate regularities from the input is involved in language and literacy acquisition, individuals with better procedural memory skills are expected to also show better performance in the language and literacy domains that draw on this ability, specifically syntax, phonology, morphology and reading (Lum et al., 2012; Ullman et al., 2020). Indeed, individual differences research in TD individuals has observed a positive relationship between language and literacy skills and procedural memory on the SRTT (e.g., Jiménez-Fernández et al., 2011; Vakil et al., 2015; van der Kleij et al., 2019). Specifically, procedural learning and consolidation on the SRTT have been found to be positively associated with grammar (Conti-Ramsden et al., 2015; Desmottes, Maillart, et al., 2017; Lum et al., 2012), sentence repetition (Desmottes, Maillart, et al., 2017), decoding and orthographic reading (Hedenius et al., 2013; Jiménez-

Fernández et al., 2011)**,** phonological awareness (Mayor-Dubois et al., 2014; Vakil et al., 2015) and rapid automatized naming (Mayor-Dubois et al., 2014).

In four meta-analyses comprising 16 studies, Hamrick et al. (2018) examined the relationship between grammar and lexical abilities and tasks designed to measure the procedural (i.e., mostly deterministic SRTT) and declarative (i.e., verbal paired associates recall) memory systems in first (children) and second language learners (adults). In children, grammatical abilities (e.g., syntactic comprehension, morphosyntactic comprehension) were observed to positively correlate with measures of both declarative memory (*r* = .160) and procedural memory (*r* = .269), whilst lexical abilities (e.g., receptive vocabulary, picture naming) were only correlated with measures of declarative memory (*r* = .409) and not procedural memory (*r* = .086). Given heterogeneity in the degree of second language proficiency, adults were divided into lower and higher proficiency groups. In the higher proficiency group, grammar positively correlated with procedural (*r* = .548) but not declarative memory (*r* = -.069). The opposite pattern was observed for the low proficiency group (procedural memory and grammar: *r* = -.099; declarative memory and grammar: *r* = .455). Together, these findings suggest a temporally dynamic relationship between declarative and procedural memory systems and grammatical abilities, whilst a clearer pattern was observed for lexical abilities. According to the procedural/declarative model, due to its flexibility, grammar acquisition is expected to rely primarily on the declarative memory systems at earlier stages of grammatical development, as observed here in children and adults with low proficiency, with increased dependence on the procedural memory system with further exposure and greater proficiency.

Despite the growing support for the procedural/declarative model, a considerable number of studies have failed to find any evidence of a relationship between language abilities and procedural memory in TD children and adults (e.g., grammar: Gabriel et al., 2011, 2013, 2015; Hsu & Bishop, 2014; Siegelman & Frost, 2015; phonology: Desmottes et al., 2016; Desmottes, Maillart, et al., 2017; Henderson & Warmington, 2017; reading: Henderson & Warmington, 2017; Schmalz et al., 2019; Vakil et al., 2015; spelling: Henderson & Warmington, 2017). Even though some null findings are to be expected even when a true effect exists under null hypothesis testing, more recently, two meta-analyses examining the relationship between language/literacy and procedural memory found no evidence for such an association (Lammertink, Boersma, Wijnen, et al., 2020; West, Melby-Lervåg, et al., 2021). Specifically, West, Melby-Lervåg, et al. (2021) conducted a series of meta-analyses examining the relationship between procedural memory and language abilities across different procedural learning tasks (e.g., SRTT, probabilistic category learning tasks, contextual cueing task, Hebb serial order learning task, artificial grammar learning and statistical learning tasks). Data from 441 participants and 5 independent studies was available for the SRTT and yielded a negligible association between procedural learning and measures of language and decoding in children and adults. Similar findings were also obtained by Lammertink, Boersma, Wijnen, et al. (2020) when examining the relationship between grammar and procedural memory (measured via the SRTT) in a sample of 139 children with DLD and 573 TD children, with a negligible association between procedural learning and expressive (*r* = .072, N effect sizes = 29) and receptive (*r* = .05, N effect sizes = 27) grammar. Importantly, the overall association did not differ significantly between groups.

The results from these meta-analyses for the TD children do not support the procedural/declarative model, given its clear predictions for an association between procedural learning measures and language/literacy abilities. However, the predictions by the procedural deficit hypothesis are more nuanced for individuals with neurodevelopmental disorders, given the predicted compensatory role of declarative memory. Firstly, a positive relationship between procedural learning and grammar would still be expected given the causal role of a procedural deficit in grammatical skills, whereby individuals with poorer procedural learning abilities are expected to show more impaired grammatical abilities. However, if the declarative memory system is able to compensate for the procedural deficit, such that individuals with better declarative memory show better grammatical skills, then a weaker association between grammar and procedural learning would be expected to occur (Lum et al., 2012). Whilst some degree of compensation by the declarative memory system is likely to occur in the TD group, given that individual differences in procedural learning are still expected, this is likely less problematic for this group given the higher variability in their procedural learning abilities compared to the disordered groups. Thus, the negligible association between grammar and procedural memory in children with DLD is not necessarily in conflict with the procedural deficit hypothesis. However, these results should be interpreted in light of the findings for the TD group, where there are testable predictions that are not supported by the available evidence.

Interim summary

Growing neuroimaging and behavioural evidence for the role of procedural memory in language and literacy acquisition and for group differences in procedural learning and consolidation in populations with dyslexia and DLD lend some support to the procedural/declarative model and the procedural deficit hypothesis. However, whilst group-level meta-analyses show a consistent pattern for impaired procedural memory on the SRTT (Lum et al., 2013, 2014; West, Melby-Lervåg, et al., 2021), individual differences meta-analyses (with the exception of Hamrick et al. (2018)) demonstrate only negligible associations between language and literacy and procedural memory in children and adults with and without dyslexia and DLD (Lammertink, Boersma, Wijnen, et al., 2020; West, Melby-

Lervåg, et al., 2021). Thus, it is still unclear whether a procedural memory deficit may be causally implicated in the core behavioural features of these disorders, whether it provides an explanation of the distribution at the extremes that does not apply to individual differences in the normal range, or whether the group differences observed in the SRTT may reflect extraneous variables involved in successful task performance, such as attention (Sengottuvel & Rao, 2013a; West, Shanks, et al., 2021). Support for the latter hypothesis comes from a recent meta-analysis which has highlighted the strong association between visual attention and reading development (Gavril et al., 2021) and consistent evidence highlighting the role of attention in language development (Gomes et al., 2000; Marini et al., 2020). Thus, in light of the absence of clear evidence for an association between language and literacy abilities and procedural memory, this evidence raises questions about the validity of the procedural/declarative model. Crucially, the disparity between group-level and individual differences research may be explained by the poor psychometric properties of the tasks used to measure procedural memory. Furthermore, even though the procedural/declarative model conceptualises procedural learning as a domain-general capacity, it is unclear whether procedural learning performance in all procedural memory tasks is expected to correlate with language abilities, as well as being impaired in children with DLD and dyslexia. We return to these challenges for assessing the validity of the procedural/declarative model and the procedural deficit hypothesis in the following sections.

Effective measurement of individual differences in procedural memory

Despite some evidence demonstrating an impairment in procedural learning in these populations when compared to typically developing children and adults, the role of procedural learning in language acquisition remains unclear. Individual difference studies (Henderson & Warmington, 2017; West et al., 2018; West, Shanks, et al., 2021) and meta-analyses (Lammertink, Boersma, Wijnen, et al., 2020; West, Melby-Lervåg, et al., 2021) have repeatedly failed to observe a relationship between language and literacy and procedural learning (cf. Hamrick et al., 2018). Furthermore, correlations among different tasks thought to index procedural memory have been small or absent (Kalra et al., 2019; Siegelman & Frost, 2015; West et al., 2018). This has led to questions over whether procedural memory represents a unitary construct, whether it is relevant to language and literacy acquisition, and/or whether the tasks that purport to measure procedural learning are psychometrically limited (e.g., Arnon, 2020; Kalra et al., 2019; Siegelman & Frost, 2015; West et al., 2018). Thus, in order to adequately address theoretical questions regarding the role of procedural learning as a mechanism of language acquisition there are a number of ‘standards’ that must be met for individual differences research to be effective in its aims, including but not limited to a well-defined mechanism/construct under investigation and a measure of this mechanism/construct that is both “task-pure” and meets acceptable psychometric standards. In the following sections we will review the procedural memory literature in light of the standards required for effective individual differences research.

Well-defined constructs

A specific definition of the construct of interest is required to distinguish procedural memory from other related constructs, to formulate testable hypotheses and to establish the validity of the procedural memory measures (von Bastian et al., 2020). Establishing construct validity, the extent to which a measure adequately captures the construct of interest, is thus a necessary step for defining procedural memory. Often, establishing convergent and divergent validity requires the examination of the relationship between measures thought to tap into the same underlying construct (e.g., SRTT and artificial grammar learning task), with high correlations between measures interpreted to reflect shared variance, as well as with outcome variables thought to be influenced by procedural memory (e.g., performance on the SRTT and on a language measure) (Cronbach & Meehl, 1955; von Bastian et al., 2020).

In the context of procedural memory, whilst this ability is thought to reflect the capacity for detecting and processing regularities from the environment across modalities, there is considerable disagreement regarding the nature of the regularities and the knowledge acquired, as well as its neural and computational underpinnings. As a consequence, there is considerable variability in the tasks used to measure procedural memory, the manipulations of which often serve practical purposes, instead of being informed by theory, and their validity has been seldom examined. Kalra et al. (2019) found that a number of implicit learning tasks (i.e., SRTT, artificial grammar learning, probabilistic classification task) showed inter-correlations among these measures ranging from -.18 to .32. Similarly, in children, West et al. (2018) observed correlations between the verbal and nonverbal versions of the SRTT and Hebb tasks ranging from -.18 to .24. Thus, there is only limited evidence for a relationship between procedural memory across experimental paradigms.

Irrespective of these limitations, and as previously mentioned, there is considerable evidence of face validity for the SRTT as a measure of procedural memory since this task has shown consistent involvement of the neurological structures underlying procedural memory (e.g., basal ganglia) and involves the incidental detection and integration of regularities either of sequential and/or statistical nature, depending on the version. Crucially, the SRTT is well-known for producing robust procedural learning effects at the group-level. However, large between-group effect sizes are not sufficient to determine whether a task is a reliable measure for individual differences research as evidenced by "the reliability paradox", whereby experimental tasks known for eliciting robust group-level effects show poor reliability (Hedge et al., 2018). The poor reliability in this context has been attributed to the discrepancy in the concepts of reliability for experimental and individual research. In experimental research, paradigms are designed with the goal of isolating the effect of interest, often by reducing variability between individuals. Conversely, in individual differences research, reliability refers to the ability of a test to consistently rank individuals, thus requiring high between subject variability (Hedge et al., 2018). We return to this issue in the following section.

Reliable measures

Reliability can refer to the ability of an instrument to consistently rank an individual's performance across time points (i.e., test-retest reliability or the *stability* of the test scores over different sessions) or between subsets of the instrument's items within a session (i.e., split-half reliability) (Nunnally & Bernstein, 1994). Whilst the ability of a test to consistently rank individuals' performance is an essential component of a measurement tool, it may be insufficient to ensure that a test shows agreement, that is that it will produce identical scores at test and retest. The latter concept, though less frequently assessed, is particularly important for clinical or educational applications where participants' scores may be used to track changes in individuals' symptoms or abilities of interest, for example.

Poor reliability of the rank order may occur due to measurement error, in accordance with classical test theory (Fleiss, 1986). In the face of high measurement error, participant's observed scores are expected to vary between measurements (Berchtold, 2016; Nunnally & Bernstein, 1994). This poses a serious problem for individual differences research as it can lead to noisier estimates and attenuation of the association between measures (Loken & Gelman, 2017; Rouder et al., 2019; Rouder & Haaf, 2019; von Bastian et al., 2020) and undermine the replicability of correlational findings (Hedge et al., 2018). Beyond the issues already highlighted, poor psychometric properties may also limit theory building as the interpretation and reliance on available evidence is dependent on the reliability of the measures. If these tasks are capturing trial noise, instead of stable effects, inconsistencies between studies' conclusions may reflect situational variation. Furthermore, poor reliability may also contribute to the underspecification of cognitive constructs assessed with experimental tasks that are known to have poor psychometric properties (e.g., inhibition: Tiego et al., 2018; attention: von Bastian et al., 2020). In relation to procedural memory, as noted by Bogaerts et al. (2021), there is considerable vagueness in the demarcation between constructs that are thought to tap into rule-based learning (e.g., statistical learning, procedural learning, implicit learning); in the mapping between experimental tasks and these constructs; and, in how these constructs are related to language acquisition and difficulties. Therefore, a necessary step forward will be to establish the reliability of the SRTT and its impact on the interpretation of existing evidence.

There is emerging evidence that speaks to the reliability of the procedural learning effect measured by the SRTT. Despite adequate or near-adequate split-half reliability (children: *r*s = .49-.75; adults *r*s = .84-.92; Lammertink, Boersma, Wijnen, et al., 2020; van Witteloostuijn et al., 2021; West et al., 2018; West, Shanks, et al., 2021), measured as the consistency in the rank order between odd and even trials, test-retest reliability (measured across sessions separated by a few days to 3 months) has been found to be considerably poorer (i.e. r <.70; children: *r*s = .21-.26; adults *r*s = .07-.70; Kalra et al., 2019; Siegelman & Frost, 2015; Stark-Inbar et al., 2017; West, 2018; West, Melby-Lervåg, et al., 2021).

Given the considerable degree of methodological variability across studies, the factors that contribute to the poor stability of the procedural learning effect across sessions remain speculative. Two influential factors may be chronological age and the number of trials. Indeed, the test-retest reliability of the SRTT has been found to be poorer in children (West et al., 2018; West, Shanks, et al., 2021) than adults (Kalra et al., 2019; Siegelman & Frost, 2015; Stark-Inbar et al., 2017; West, Shanks, et al., 2021). Furthermore, West and colleagues observed poor stability for the procedural learning effect in a probabilistic SRTT in children when using both 500 (*r* = .21; West et al., 2018) and 1000 (*r* = .26; West, Shanks, et al., 2021) trials in children. Conversely, increasing the number of trials from 500 to 1000 in adults led to an improvement in test-retest reliability from .53 to .66, but the improvement in test-retest reliability was minor for 1500 trials (*r* = .68). Another factor may be baseline response time (RT): A slight increase was observed once participant's average speed was accounted for (1000 trials, *r* = .71; 1500 trials, *r* = .70; West, Shanks, et al., 2021). This suggests that in adults increasing the number of trials may have contributed to a decrease in measurement error and thus an improvement in reliability (Rouder et al., 2019; Rouder & Haaf, 2021). For children on the other hand, it is likely that the longer task duration and the additional demands on attentional resources (West, Shanks, et al., 2021) may counteract the benefits of increasing the number of trials. We return to this issue later in this chapter. The findings from this study (West, Shanks, et al., 2021), have also highlighted how the method used for computing the procedural learning effect may impact the reliability of the SRTT. To our knowledge, thus far West, Shanks, et al. (2021) has been the first to compare the reliability of the SRTT when using different methods of measuring procedural learning and observed that ratio scores (which account for differences in average speed) are numerically more reliable than difference scores. Other measures of procedural learning have also been adopted, such as random slopes extracted from hierarchical models (Lammertink, Boersma, Wijnen, et al., 2020; van Witteloostuijn et al., 2021), but these have not been directly compared with other methods. Future research should aim to determine whether these analytical choices impact the reliability of the SRTT.

As of yet there are no current studies that directly and systematically examine which factors influence the psychometric properties of the SRTT. However, other factors may influence the reliability of the SRTT, such as the type of SRTT and the duration of the interval between sessions. For instance, deterministic tasks appear to show numerically lower stability than the alternating SRTT. Stark-Inbar et al. (2017) analysed the test-retest reliability of deterministic and alternating sequences in TD adults and found that the test-retest reliability for deterministic sequences (*r* = .07) was much lower than for alternating sequences (*r* = .46). The difference in test-retest reliability between deterministic and alternating SRTTs was hypothesised to be due to differences in levels of explicit awareness as participants who performed the deterministic SRTT reported more confidence in the presence of a repeating pattern than those performing the alternating SRTT. As explained by the authors (Stark-Inbar et al., 2017), it is unlikely that the contamination with explicit awareness is inherently unreliable, instead it is the individual differences in explicit awareness which contributes to poor reliability. Not only may participants become aware of the sequence at distinct stages of the task, but also explicit awareness may affect performance across participants in distinct manners, whilst it may promote procedural learning in some, it may hinder performance in others, thus resulting in changes in the ranking order between sessions.

In light of the evidence demonstrating offline consolidation and practice effects on the SRTT, it is crucial to determine whether individual differences in consolidation processes may contribute to changes in rank order between test and retest. The study by Siegelman & Frost (2015), which asked TD adult participants to perform a probabilistic SRTT twice with an interval between test and retest of approximately 3 months, observed a moderate test-retest reliability (*r* = .47), with participants performing significantly better at retest. The mean RT difference more than doubled from test to retest, thus raising the possibility of consolidation and/or practice effects. Even though the shorter interval between test and retest may account for the difference in reliability, the findings by West, Shanks, et al. (2021) of superior reliability to that of Siegelman and Frost (2015), may also suggest that the use of alternate forms of the SRTT may lead to more stability of the procedural learning effect as these have been often used to avoid practice effects (Beglinger et al., 2005). Future research is thus required to systematically investigate the psychometric properties of the SRTT, and which factors may impact its reliability, starting by examining whether the adoption of the same or alternate versions of the sequences affects the stability of procedural learning.

Considering the poor reliability of the SRTT, the inconsistent findings from individual differences research may at least be partially explained by attenuation (i.e., underestimating the true correlation between two different measures), with studies with better reliability are more likely to capture the true association between measures and more attenuation being expected to occur in children. Thus, the lack of evidence for an association between procedural memory tasks (Kalra et al., 2019; Siegelman & Frost, 2015; West et al., 2018), which is often interpreted as suggesting that these tasks tap into different sources of variation and thus arguing against a unitary view of procedural memory, may instead reflect the lack of reliability of the scores (Baugh, 2002). Similar reasoning is applied to the evidence examining the procedural/declarative model which posits that rule-based linguistic abilities such as grammar and phonology depend on the procedural memory system, according to which individuals with better procedural memory are expected to demonstrate better linguistic abilities. However, as previously reviewed, individual differences research has produced mixed findings. Whilst it is plausible for this pattern of findings to indicate that procedural learning, as indexed by the SRTT, is not involved in language and literacy and thus call into question the validity of the procedural/declarative model; it is also possible that the poor reliability of the SRTT, which is not specific to the SRTT task, but instead observed across procedural memory tasks (Arnon, 2020; Kalra et al., 2019; West et al., 2018), may have contributed to attenuation of the effect sizes. To be able to discriminate between these possibilities, experimental paradigms with better psychometric properties or more efficient techniques for disattenuating effect sizes are required (Rouder et al., 2019).

In the absence of such options, the study conducted by West, Shanks, et al. (2021) provides a unique opportunity to examine the predictions of the procedural/declarative model as it represents the only study that has reported adequate test-retest reliability. In this experiment with adults between 18 and 61 years of age, only negligible correlations between procedural learning and literacy measures were observed (rs from -.06 to -.11; West et al., 2021). Whilst this study does not provide support for the procedural/declarative model, one study is not sufficient to fully test this theory, especially if there are developmental changes in the relationship between literacy and procedural memory that need to be taken into consideration as suggested by the meta-analysis conducted by Hamrick et al. (2018). Thus, future research should aim to routinely report the psychometric properties of the SRTT to allow the determination of which factors contribute to its poor reliability to better inform task design. Such an endeavour requires researchers to not only examine sample and task characteristics, but also to resolve some of the debates related to the computational and cognitive demands of these procedural memory tasks. Whilst some researchers have hypothesised procedural learning in the SRTT to be independent of attention (e.g., Frensch et al., 1998; Heuer & Schmidtke, 1996; Schmidtke & Heuer, 1997) and explicit awareness (Destrebecqz & Cleeremans, 2001; D. R. Shanks et al., 2005), evidence regarding the role of attention and explicit awareness in the SRTT has produced mixed findings and suggests that adequate performance on procedural memory tasks relies on the involvement of other cognitive abilities. Not only are these considerations relevant for a better understanding of procedural memory as a construct, but task impurity is also likely to exert a negative effect on the stability of measures of procedural memory.

Task-pure measures

The role of attention on the SRTT

Attention is one potential factor that may influence performance on the SRTT and the reliability of the procedural learning effect. Sustained attention consists of the ability to maintain alertness and focus directed at a task for a long period (Sarter et al., 2001) and it has been shown to support other cognitive processes such as learning, memory, and executive functions (Barker, 2012; Sarter et al., 2001). The development of attention seems to follow a U-shaped trajectory with improvements from early adolescence into adulthood and a plateau during young to middle adulthood, with a decline in older adults in a sample with ages between 12 and 75 years (McAvinue et al., 2012).

The role of attention in learning is still under debate, yet there is evidence suggesting that, given that more information than can be processed is available at any one time, this is an important mechanism for reducing perceptual uncertainty (Gottlieb, 2012). Dichotomous models of attention control describe it in terms of top-down and bottom-up control (or endogenous and exogenous control, respectively). Whilst top-down control is thought to be determined by the goals of the learner and can be voluntarily allocated towards a task in a goal-driven manner (Leber & Egeth, 2006; Posner et al., 1980); bottom-up control, on the other hand, is determined by the physical characteristics of the stimuli and thus may be involuntarily diverted to stimuli with salient properties (Theeuwes, 2010). More recent evidence, however, suggests that these mechanisms are insufficient to account for strong selection biases, whereby previous attentional deployments persist despite no longer matching current goals or the salience of the stimuli. Thus, a new framework postulates that these lingering selection biases may also influence attention selection (Awh et al., 2012; Failing & Theeuwes, 2018). In relation to procedural learning, this conceptualisation of attention proposes a mechanism by which a subset of all stimuli is selected and maintained in working memory, thus allowing for the detection and extraction of probabilistic and sequential regularities (van Moorselaar & Slagter, 2019; J. Zhao & Luo, 2017). Conversely, there is also evidence suggesting that the learning of these regularities may itself affect attention, whereby procedural learning allows for the prediction of upcoming stimuli and thus may facilitate perceptual processing, reducing the demands on attention. Furthermore, learning of the regularities of distractors has also been found to reduce the amount of attention captured by these distractors and increase the efficiency of selecting the target stimuli (Wang & Theeuwes, 2018). With this broad perspective on how attention might interact with the learning of sequential information, we now turn to review the literature on the role of attention in the SRTT specifically. We start by examining the literature on procedural learning using the SRTT in individuals with attention deficit hyperactivity disorder (ADHD), then consider the performance on the SRTT in dual-task paradigms and finish by reporting the findings of individual differences research.

One approach to examining the role of attention in procedural learning has been to focus on individuals with ADHD. ADHD is a neurodevelopmental disorder characterised by a pattern of sustained attentional deficit and impulsive and hyperactive behaviours (Mirzakhany - Araghi et al., 2013; Polanczyk et al., 2014). Given the often-repetitive nature and long duration of the SRTT, it would be expected for sustained attention decrements to occur when performing this task, which could contribute to a decline in attentional resources across time and lead to disengagement from the task (Fortenbaugh et al., 2017). Considering the attentional difficulties experienced by individuals with ADHD, this pattern would likely be more accentuated for individuals with ADHD. Thus, if the successful acquisition of procedural learning in the SRTT is attention-dependent, this population would be expected to underperform in the SRTT when compared to TD individuals. Whilst some studies find preserved procedural learning in the SRTT (adolescents: Karatekin et al., 2009; Laasonen et al., 2014; Pedersen & Ohrmann, 2018; children: Takács et al., 2017), others observed impaired performance of individuals with ADHD compared to TD groups (children: Barnes et al., 2010). Barnes et al. (2010) found similar procedural learning in an alternating SRTT in children with ADHD and controls at initial (epochs 1 and 2) and later stages (epochs 4 and 5), yet reduced procedural learning compared to controls at the midpoint (epoch 3) of the task. Conversely, (Karatekin et al., 2009), observed that, even though adolescents with ADHD demonstrated comparable sequence learning to controls in manual and oculomotor versions of the SRTT, they showed fewer anticipatory oculomotor movements than controls. Authors suggest that the group differences in anticipatory oculomotor movements may be due to an overall slower rate of processing information and anticipating stimuli, or impairment in processing temporal information or in anticipatory processing.

Several experimental studies have utilised dual task paradigms to examine the role of attention in the SRTT task, in which the focus of learners' top-down attention is manipulated to focus on a secondary task performed alongside the SRTT. The expectation is that, if procedural learning is dependent on attentional resources, participants in the dual-task condition would show less (or no) evidence of procedural learning when compared to those in a single task condition. Whilst some studies have observed intact procedural learning in both dual and single-task conditions (deterministic sequences: Barker, 2012; Coomans et al., 2014 (children); Frensch et al., 1994; probabilistic sequences: Jimenez & Mendez, 1999), others have reported that the presence of a secondary task significantly disrupts learning in the SRTT (deterministic sequences: Coomans et al., 2014; Schumacher & Schwarb, 2009; Shanks & Channon, 2002; Wierzchoń et al., 2012; probabilistic sequences: D. R. Shanks et al., 2005).

Multiple theories have been put forward to explain impaired performance under dual-task conditions. The attentional resource theory suggests that impaired procedural learning is expected to occur in dual-task conditions as the cognitive resources have to be shared between concurrent activities (Kahneman, 1973). Yet, the attentional resource theory is not sufficient for explaining the inconsistent patterns in the dual-task SRTT experiments, as only a few studies report impaired performance in dual-task conditions. Thus, Frensch et al. (1998) proposed the suppression hypothesis which suggests that procedural learning is an automatic process independent from attention, with the secondary task only negatively affecting the expression of sequence knowledge without disrupting procedural learning. According to this hypothesis, regardless of learning condition, if the transfer block is conducted in a dual-task condition, evidence of procedural learning will be lower than in a single-task testing condition. Heuer and Schmidtke (1996), on the other hand, suggest that procedural learning is often impaired in dual-task conditions when both tasks are not processed separately with participants attempting to integrate both stimuli into one sequence (Heuer & Schmidtke, 1996; Schmidtke & Heuer, 1997). More recently, Schumacher and Schwarb (2009) provided an alternative explanation that attempts to account for the inconsistency in the literature regarding the dual performance cost. They propose that performance in dual-task conditions should only disrupt procedural learning when responses to the primary and secondary tasks are required in parallel. For example, when performing an SRTT whilst using the tone counting task as a secondary task, both stimuli would be presented simultaneously, and participants would be required to produce a response in every trial for both stimuli.

These theories were contrasted directly in a recent study conducted by Röttger and colleagues (2019). In their first experiment, they replicated the study by Schumacher and Schwarb, (2009) and observed impaired procedural learning in the dual-task conditions (one requiring participants to respond to high- and low-pitched tones by saying “high” or “low” and the other by saying “blue” or “yellow”). Furthermore, they found that even when reducing the dimensional similarities between primary and secondary tasks, by requiring auditory responses with no spatial information, procedural learning was still similarly disrupted. In experiment 2, it was observed that, if the tone-counting task only required a response in a few trials (in this experiment 30%), some evidence of procedural learning was observed which could account for the results obtained by Frensch et al. (1998, 1999) where participants were only required to provide a response in 50% of the trials. Furthermore, when participants were required to perform an SRTT in a dual-task condition in which the visual and auditory information was correlated, participants demonstrated evidence of procedural learning. Preserved procedural learning was also observed in experiment 3 when participants were required to perform the SRTT in a dual-task condition that required them to listen, but not repeat, the words. Finally, experiment 4 observed that participants only learned elements of the sequence which had been consistently paired with tones, thus hinting that within-trial predictability affects procedural learning. Together, these findings lend some support to the task integration hypothesis as the requirement of a parallel response is not a sufficient condition to disrupt learning on the SRTT. Instead, it appears that the randomness of the secondary task is a determining factor. As suggested by Rah et al. (2000) the disruption of procedural learning in dual-task paradigms occurs due to the conceptualisation of the SRT and the secondary tasks as a single task, with participants analysing the input for potential patterns of covariation. Finally, these findings also do not lend support to the suppression hypothesis (Frensch et al., 1999) as all testing was conducted in single-task conditions yet impaired procedural learning was still observed under certain circumstances.

Irrespective of which hypotheses better explains the dual-task cost, this paradigm may not only divert attention from the SRTT, but also interfere with the emergence of procedural learning (Franklin et al., 2016, Kiss et al., 2019, Schmidtke & Heuer, 1997). For example, if the secondary task extends the period between trials, it may prevent the perceptual binding between the cues (Kiss et al., 2019; Schmidtke & Heuer, 1997). Thus, mind-wandering was explored as an indicator of lapses in attention during a deterministic SRTT (Franklin et al., 2016). In this study, participants were exposed to 12 thought probes and asked to rate whether their attention was “completely focussed on the task” to “completely unrelated concerns” on a 5-item scale. Using this approach, the authors found a negative correlation between procedural learning and mind-wandering (r = -.31), thus suggesting that learning on the SRTT is enhanced when the participants are focusing their attention on the task (Franklin et al., 2016). Yet, this result was not replicated in Brosowsky et al. (2021), as a negative correlation between mind wandering and procedural learning was only observed for participants in the explicit, but not implicit, condition. Crucially, mind wandering was observed to increase with practice. Sustained attention has been found to be positively correlated with procedural learning on the SRTT in children

(Sengottuvel & Rao, 2013a; West, Shanks, et al., 2021). Furthermore, (West, Shanks, et al., 2021), observed significant and positive correlations (ranging from .22 to .32) between procedural memory and measures of attainment (on measures of reading, grammar, and arithmetic). However, once attention, declarative and procedural measures were entered as predictors of children's attainment (on measures of reading, grammar, and arithmetic) in a latent variable path model, only attention and declarative measures were significant unique predictors, whereas performance on the SRTT was not. This suggests that once attentional abilities are accounted for, procedural learning may not add explanatory power, thus raising questions about whether the correlations between language and literacy and procedural learning as well as group differences in procedural learning between individuals with dyslexia and DLD and TD controls may be partially explained by weaknesses in attention often observed in these disordered populations (Ebert & Kohnert, 2011; Paulesu et al., 2014).

In summary, these findings suggest that the role of attention in the SRTT is not yet fully understood, yet there is some evidence suggesting that focussing attentional resources on the SRTT may contribute to better performance. No experiment has examined the impact of attentional abilities on the stability of the procedural learning effect. Yet, not only has previous research examining the stability of attention abilities shown poor reliability (see von Bastian et al., 2020 for a discussion), but it is plausible that fluctuations in attention during task performance may contribute to the between-session variability in the procedural memory scores, thus resulting in poor reliability. If this were the case, and in line with previous evidence (West et al., 2018; West, Shanks, et al., 2021), the stability of the procedural memory effect would be expected to be lower in children than adults given previous evidence suggesting more intra-individual variability in children (Boen et al., 2021). Similarly, given the pattern of weaknesses in attention frequently observed in populations with dyslexia and DLD (Ebert & Kohnert, 2011; Paulesu et al., 2014), the stability of the procedural learning effect would also be expected to be lower for those with neurodevelopmental disorders than TD populations. Yet, thus far, to our knowledge, no study has examined the test-retest reliability of the SRTT in atypical populations.

The role of explicit awareness on the SRTT

Another factor which may impact procedural learning and its reliability is explicit awareness. Firstly, the emergence of explicit awareness on the SRTT may contaminate the measurement of procedural learning, thus providing an impure measure of this construct and limiting the conclusions that can be drawn regarding the role of procedural learning in language and literacy. Secondly, explicit awareness may impact the stability of the procedural learning effect, as whilst explicit awareness is not inherently problematic, individual differences in explicit awareness may impact procedural learning in the SRTT differently across individuals, therefore contributing to changes in participants' ranking order between test and retest (Stark-Inbar et al., 2017). Thus, we provide an overview of the role of explicit awareness on the SRTT and reflect on these issues in more detail.

Even though the SRTT has a long history of being used as an index of implicit learning, it is still unclear whether learning in this paradigm is independent of explicit awareness (D. R. Shanks et al., 2005). At least some individuals demonstrate procedural learning without awareness of the underlying sequence (Esser et al., 2021; Esser & Haider, 2017), but there is also evidence that some individuals become aware of at least parts of the sequence they are learning (D. R. Shanks et al., 2005; Stark-Inbar et al., 2017). However, the operationalisation of implicit and explicit knowledge is not always clear in the SRTT literature, and there is little agreement on the most suitable way to measure explicit knowledge of the underlying sequence (see Schwarb & Schumacher, 2012 for a discussion). There are currently two dominant theories on the generation of explicit knowledge in this literature. The simple-system view (Cleeremans & Jiménez, 2002, further elaborated in Cleeremans, 2008, 2011) posits that explicit awareness of the sequence depends on the strength of the procedural knowledge. Thus, whilst procedural learning is implicit at earlier stages, representations of the underlying sequence will gradually become stronger and result in explicit awareness (Cleeremans & Jiménez, 2002). Multiple-systems views, such as the Unexpected Even Hypothesis, hypothesise that implicit and explicit learning rely on different memory systems. According to the Unexpected Event Hypothesis (Frensch et al., 2003; Rünger & Frensch, 2008), explicit awareness occurs as a result of the discrepancy between expected and actual behaviour when performing the task, whereby the search for the cause of the unexpected event will prompt an inferential process which will likely lead to the discovery of regularities. In the SRTT this may occur when participants switch between sequenced and random trials as noticeable changes in response times and accuracy of their responses may trigger reflection resulting in explicit awareness (Frensch et al., 2003; Rünger & Frensch, 2008). From the Unexpected Event Hypothesis, one can hypothesise that deterministic SRTTs would potentially more easily elicit explicit awareness of the sequence given that the transitions between sequenced and random blocks would be expected to lead to more abrupt changes in RTs and accuracy.

Irrespective of the mechanisms involved in the generation of explicit awareness, the SRTT is not a process-pure paradigm that only taps into implicit learning. Instead, both the SRTT and measures of explicit awareness are likely to capture both implicit and explicit awareness (Destrebecqz et al., 2005), though the degree of involvement of each component may differ depending on task characteristics. Whilst prior knowledge of the underlying sequences has been found to improve performance on deterministic SRTTs, on probabilistic sequences no such benefits were observed (Song et al., 2007;

Stefaniak et al., 2008). Regarding sequence complexity, Pascual-Leone et al. (1993) observed that the magnitude of the difference between patients with Parkinson's disease and controls was larger when presented with a 12-item first-order conditional sequence than an 8-item first-order conditional sequence, whilst Kelly et al. (2004), in a dual task condition, observed that patients with Parkinson's disease were predominantly impaired for second-order conditional but not first-order conditional procedural learning. This suggests that first-order transitions may be explicitly learnt, but that the learning of the more complex second-order dependencies is implicit. Thus, the degree to which the procedural learning effect will be contaminated by explicit awareness may differ depending on task design (Reed & Johnson, 1994; Schwarb & Schumacher, 2012).

Given previous findings, it is possible that the poor reliability of the deterministic SRTT can be attributed to explicit awareness. In the study conducted by Stark-Inbar et al. (2017), two groups were asked to perform the alternating and deterministic versions of the SRTT, with the former showing a test-retest reliability of .46 whilst the latter showed a negligible .07. Despite longer practice sessions in the alternating SRTT (Willingham et al., 1989), participants showed more evidence of explicit awareness on the deterministic task. However, other variables may also have contaminated task performance, namely fatigue, decreases in motivation and attention. Whilst these variables would also affect performance on the alternating SRTT, since learning is measured throughout the task, instead of only at the end as is commonly done for deterministic tasks, the procedural learning effect for the deterministic SRTTs would likely be more affected by these extraneous factors. Crucially, it is likely not the emergence of explicit awareness itself that contributes to poor reliability, since if it were the case that explicit awareness exerts the same effect across participants, the ranking order between sessions would be maintained. However, if explicit awareness (or other extraneous factors) impact participants differently within and across sessions, then poor reliability is expected to occur. The latter reflects the most likely explanation given previous evidence demonstrating that explicit awareness varies depending on sampling characteristics such as age (Verneau et al., 2014) and working memory (Martini et al., 2013). Thus, studies aiming to test the predictions of the procedural/declarative model and procedural deficit hypothesis need to consider task characteristics in order to obtain more process-pure and reliable measures of procedural learning.

Interim summary

In summary, whilst the SRTT is well-known for producing robust procedural learning effects at the group level (both across populations and task manipulations), there are several issues which limit the validity of the findings obtained using this task in individual differences research. Firstly, given the poor specification of procedural memory and the limited understanding of the underlying computations of the SRTT it is unclear with which language and literacy measures the SRTT is expected to correlate. Secondly, there is considerable evidence suggesting that the SRTT is not a process-pure measure of procedural learning, which not only may have implications for the inferences that can be made from individual differences, but also for the reliability of the SRTT. Thirdly, the reliability of the SRTT has been shown to be poor, especially across sessions (i.e., test-retest reliability). Due to the limited number of studies which have examined the test-retest reliability of the SRTT recommendations for how to optimise reliability are not possible at this stage. However, there is preliminary evidence suggesting that individual differences in attention and explicit awareness may be partially responsible for the poorer psychometric properties of this task. Additionally, it is also likely that design and analytical choices may affect the stability of the scores within and across sessions. Whilst the poor reliability of the SRTT does not preclude its use in group-level comparisons, it may lead to attenuation of the correlations of interest. This may help to explain the disparity between individual differences and group-level findings presented throughout this review. However, in light of task impurity, it is reasonable to be cautious about whether these group-level differences reflect differences in procedural learning or whether they may be explained by other factors.

Conclusion

Evidence for the role of procedural memory in language and literacy development is still inconclusive, especially in light of issues with reliability of the procedural memory measures and the vague delineations between procedural memory and other cognitive abilities required for adequate performance on these measures. Although some group-level findings point to a procedural learning impairment on the SRTT for individuals with dyslexia and DLD when compared to TD controls, thus providing some support for the procedural deficit hypothesis, it is still unclear whether the poorer performance of individuals with dyslexia and DLD may be explained by a third variable that has not been accounted for such as attention. Further investigation is thus crucial to determine whether impairments in procedural memory represent a risk factor for neurodevelopmental disorders such as dyslexia and DLD. A better understanding of the role of procedural memory in language and literacy development has the potential to inform current instructional and remediation practices.

Considering present challenges, future research should focus on experimentally improving the psychometric properties of the SRTT by determining which factors affect its reliability. As a starting point, given the proneness of memory tasks to practice effects, it is crucial to determine the impact of using the same or alternate versions of the sequence at test and retest before exploring whether other design features, such as the presence of an interstimulus interval, affect reliability. Whilst procedural memory, as measured through the SRTT task, has consistently been shown to have poor reliability, this does not indicate that the SRTT is inherently unreliable. Instead, since the psychometric properties of a task reflect the characteristics of the scores produced in a particular setting and context (Feldt & Brennan, 1989), knowledge of how task manipulations affect reliability may allow for researchers to optimise the psychometric properties of this task. Until then, individual differences research may be limited in the inferences that can be drawn to inform how performance on the SRTT relates to language and literacy. Finally, there is a clear need for more research focussing on the role of attention in the SRTT, as this will not only contribute to a better understanding of the mechanisms involved in procedural learning, but also to be informative of whether group differences between individuals with DLD and dyslexia may be partially, if not fully, explained by comorbid attentional deficits in these groups.

Outline and research questions

In this thesis we take a multi-method approach to provide an in-depth examination of procedural memory and its relation to language and literacy proficiency, as well as attentional abilities, of adults with and without DLD and dyslexia. This includes group level and individual differences behavioural studies, both online and laboratory-based, and meta-analyses using state-of-the-art statistical techniques.

We address five key aims, namely:

(1) whether a procedural learning effect on the SRTT can be detected at the individual- and group-level in adults with and without dyslexia

(2) whether the procedural learning effect on the SRTT is stable within- and across-sessions in adults with and without dyslexia

(3) whether individual differences in procedural learning ability correlate with language and literacy proficiency in adults with and without dyslexia

(4) whether the magnitude and stability of the procedural learning effect relates to attentional abilities, and

(5) whether there are group-differences in the procedural learning effect between adults with and without dyslexia.

As outlined above, a crucial first step is to comprehensively examine the psychometric properties of the SRTT. **Chapter 2** systematically examined the psychometric properties (reliability and agreement) of the SRTT by manipulating the similarity of the sequences at test and retest (experiment

1), increasing the number of sessions (experiment 2) and including an interstimulus interval (supplementary experiment). Alongside these experimental manipulations, the role of sustained attention on the magnitude and stability of the procedural learning effect was also analysed in light of recent evidence suggesting an association between attention and procedural learning (experiment 2 and supplementary experiment). This initial effort in establishing the reliability of the SRTT was crucial not only because the lack of stability of the procedural learning effect may lead to attenuation of the associations between the variables of interest and thus limit our inferential capabilities, but also because if poor reliability of the procedural learning effect were consistently observed, it would suggest that information obtained through these tasks could not be used to drive educational or clinical decision-making.

In **Chapter 3**, we establish the overall stability of the SRTT in a meta-analysis which examined the stability of the SRTT within (split-half reliability) and across sessions (test-retest reliability). This allowed us to assess whether the findings from **Chapter 2** are replicated across studies, as well as explore the effects of sampling (participant age), methodology (number of trials, sequence type, inclusion of an interstimulus interval, version of the SRTT) and analytical decisions (whether all trials were included when computing the procedural learning scores; using difference scores, ratio scores, or random slopes as an index of learning) on the stability of the SRTT.

Finally, and in light of the findings pertaining to the psychometric properties of the SRTT, we examine the predictions of the procedural/declarative model and the procedural deficit hypothesis at the individual (**Chapters 4** and **5**) and group-level **(Chapter 5**). **Chapter 4** presents a meta-analysis of the literature examining the relationship between language and literacy and procedural memory on the SRTT in populations with and without DLD and dyslexia. This meta-analysis also includes the individual differences research conducted in experiment 2. In this chapter, we assess the predictions of the procedural/declarative model which proposes that individuals with better procedural memory abilities are expected to show better proficiency in language and literacy domains that rely on the ability to detect and assimilate regularities in the input. Whilst, according to this model, TD adults would be expected to show this pattern of associations, the findings for groups with dyslexia and DLD are less clear given that a deficit in procedural memory might be, at least partly, compensated by declarative memory. Hence, for the dyslexic and DLD groups procedural learning is expected to correlate with language and literacy abilities, unless declarative memory has compensated for the procedural deficits.

Chapter 5 culminates with an overall examination of the learning and consolidation of procedural memory on the SRTT of adults with dyslexia and typically developing adults at the group and individual level across three sessions, whilst analysing the stability of the procedural memory effect within and across sessions. The procedural deficit hypothesis postulates that a procedural memory impairment may be causal in dyslexia; therefore, the dyslexic group is predicted to show less evidence of procedural learning and/or consolidation than the comparison group of TD adults.

All pre-registrations, data and scripts for analyses described in this dissertation are available at Open Science Framework (OSF) project pages. All publications resulting from the work presented here will be openly published.

Chapter 2. Reliability Of the Serial Reaction Time Task: If at First You Don't Succeed, Try Try Try Again

Abstract

Procedural memory is involved in the acquisition and control of skills and habits that underlie rule and procedural learning, including the acquisition of grammar and phonology. The Serial Reaction Time task (SRTT), commonly used to assess procedural learning, has been shown to have poor stability (test-retest reliability). We investigated factors that may affect the stability of the SRTT in adults. Experiment 1 examined whether the similarity of sequences learned in two sessions would impact stability: test-retest correlations were low regardless of sequence similarity (r < .25). Experiment 2 added a third session to examine whether individual differences in learning would stabilise with further training. There was a small (but nonsignificant) improvement in stability for later sessions (session 1-2: r = .43; session 2-3: r = .60). Stability of procedural learning on the SRTT remained suboptimal in all conditions, posing a serious obstacle to the use of this task as a sensitive predictor of individual differences and ultimately theoretical advance.

Introduction

Procedural memory underlies the encoding, storage and retrieval of motor, perceptual and cognitive skills that involve the integration of sequenced, statistical, and probabilistic knowledge across the lifespan (Eichenbaum, 2002; Eichenbaum & Cohen, 2001; Koch et al., 2020; Ullman, 2004). Learning in this system relies on the basal ganglia (specifically, the striatum), the cerebellum and portions of the parietal and frontal cortices (Packard & Knowlton, 2002; Parent & Hazrati, 1995; Poldrack & Packard, 2003) and tends to be gradual, yet once the skills have been learnt they are used rapidly and automatically. The procedural memory system is proposed to be involved in language acquisition. Specifically, Ullman and colleagues (Ullman, 2004; Ullman et al., 2020) propose that the procedural memory system supports the acquisition of rule-based linguistic knowledge, such as phonology and grammar; whilst the declarative system is mostly associated with acquisition of more arbitrary and explicit knowledge, such as vocabulary. Supporting this, language and procedural memory share brain systems, including basal ganglia and frontal cortex, especially Broca's area (Ullman, 2001b; Ullman & Pierpont, 2005), and clinical populations with impairments of the basal ganglia tend to show both motor and linguistic impairments (Ullman & Pierpont, 2005). Aligning with the declarative/procedural model, some previous studies have shown small to moderate correlations between procedural learning and language and literacy abilities (Clark & Lum, 2017a; Desmottes et al., 2016; Desmottes, Meulemans, et al., 2017; Lum et al., 2012). However, other studies have failed to replicate these associations (Gabriel et al., 2015; Henderson & Warmington, 2017; Siegelman & Frost, 2015; Vakil et al., 2015; West et al., 2018). This inconsistency, coupled with recent concerns about the psychometric properties of tasks used to measure procedural learning (SRTT: Kalra et al., 2019; Siegelman & Frost, 2015; Stark-Inbar et al., 2017; West et al., 2018; West, Shanks, et al., 2021; contextual cueing and Hebb tasks: West et al., 2018; statistical learning tasks: Arnon, 2020), calls for further research to systematically examine the reliability of markers of procedural learning.

The SRTT (Nissen & Bullemer, 1987) is the most widely used measure of procedural (or sequence, depending on the type of task) learning that requires participants to connect a series of events and form high-order associations to predict future positions (Keele et al., 2003). It has been shown to rely on the same neural networks as other measures of procedural learning (Clark et al., 2014; Hardwick et al., 2013). For example, patients with basal ganglia disorders (e.g., Huntington's disease) show impaired procedural learning on the SRTT (Willingham & Koroshetz, 1993), and fMRI studies demonstrate that procedural learning captured by the SRTT elicits activation in the basal ganglia (putamen: Janacsek et al., 2020; ventral striatum: Doyon et al., 1996; and the cerebellum: Hardwick et al., 2013). In the SRTT, a stimulus is presented in an array (e.g., 4 squares presented horizontally across a screen) and participants are required to press a corresponding button on a keypad or button box to the position of the stimulus on screen as quickly as possible. Unbeknown to the participant, some of the stimulus transitions follow a sequence, with procedural learning being measured as the response time difference between the sequenced and random trials. Faster responses to sequenced than random trials are taken as a “procedural learning effect”, indicating that the participant has learnt the sequence and is therefore able to anticipate the next position.

SRTTs can be deterministic or probabilistic. Deterministic sequences usually comprise random and sequenced blocks. The first blocks typically contain the repeating sequence, with a sudden switch to a random block, followed by a final sequenced block. Reaction times (RTs) tend to decrease progressively during practice in sequenced blocks but then increase in random blocks; this difference in RT is taken as evidence of procedural learning. In contrast, probabilistic SRTTs usually comprise two second order conditional sequences, one that occurs with a higher probability than the other (e.g., sequence A (85%): 121432413423; sequence B (15%): 323412431421; Siegelman & Frost, 2015). Each block starts with a random bigram (e.g., 43) and the next location selected will be either the location that followed that bigram in sequence A (i.e., 2, termed a ‘probable’ trial) or the location that following that bigram in sequence B (i.e., 1, termed an ‘improbable’ trial). Procedural learning in probabilistic SRTTs is measured as the difference in response times between probable and improbable trials. Importantly, despite participants showing evidence of procedural learning, they often have little to no awareness of the presence of a probabilistic sequence (Destrebecqz & Cleeremans, 2001). Deterministic sequences, on the other hand, have been found to yield more explicit awareness of the sequence (Jiménez & Vázquez, 2005; Stark-Inbar et al., 2017; Stefaniak et al., 2008). Thus, the probabilistic sequences may represent purer measures of implicit procedural learning (Stefaniak et al., 2008).

The SRTT is well-known for producing robust effects at the group level, thus recently there has been increased interest in using the SRTT as a marker of individual differences (Siegelman & Frost, 2015). However only a few studies have explored the psychometric properties of the task. Reliability refers to the ability of a task to rank individuals’ performance consistently across time, with higher reliability indicating stable scores obtained at test and retest (Hedge et al., 2018). Split-half reliability, a measure of internal consistency within a single session that reflects the correlation between scores within a test (Nunnally & Bernstein, 1994), has been shown to be moderate to adequate on the SRTT in children and adults, respectively (children: *r*s = .49-.75; adults *r*s = .84-.92; West et al., 2018; West, Shanks, et al., 2021). However, test-retest reliability (i.e., the *stability* of the test scores over different sessions) is notably poorer and below acceptable psychometric standards (i.e., *r* <.70, Burlingame et al., 1995; Nunnally & Bernstein, 1994) in children (*r* = .21, West et al., 2018 (500 trials); *r* = .26, West, Shanks, et al., 2021 (1000 trials)) and adults (deterministic SRTT: Kalra et al., 2019; Stark-Inbar et al., 2017; probabilistic SRTT: Siegelman & Frost, 2015; West et al., 2018; alternating SRTT: Stark-Inbar et al., 2017). In one exception, West, Shanks, et al. (2021) obtained a test-retest reliability of .71 using a probabilistic SRTT with 46 adults aged between 18 and 61 years. The unusually high stability reported here could be due to one or more of a number of methodological differences (e.g., a large number of trials (i.e., 1500), the same sequence was administered twice, the gap between tests was two-three days and use of a 250ms interstimulus interval (ISI)).

According to classical test theory (Fleiss, 1986), observed scores reflect true scores and measurement error and higher degrees of measurement error lead to greater fluctuations in scores across time. This translates into poor test-retest reliability as participants' relative ranking will change between test and retest (Berchtold, 2016; Nunnally & Bernstein, 1994). Poor reliability may contribute to noisier predictions; increased uncertainty in parameter estimation (Loken & Gelman, 2017); and attenuation of the association between measures (Rouder et al., 2019; Rouder & Haaf, 2019, 2020). In small samples, as demonstrated by Loken and Gelman (2017), measurement error can lead, by chance, to overestimation of the effect size. Thus, the poor reliability of the SRTT may contribute to the inconsistently reported correlations between language/literacy measures and procedural learning (LeBel & Paunonen, 2011). It is however important to note that in the one study to date which reports adequate test-retest reliability for the SRTT (*r* = .71; West, Shanks, et al., 2021), only negligible correlations were observed between procedural learning and word and nonword reading measures (*r*s from -.06 to -.11; West, Shanks, et al., 2021). Thus, even in the face of adequate stability, this lack of association remains contrary to the predictions of the declarative/procedural model. Nevertheless, it is a single study, and identifying optimal conditions for achieving better reliability remains imperative. Indeed, only a robust and reliable task can test the boundaries of the procedural/declarative model of language acquisition, including the procedural deficit hypothesis, and permit a better understanding of the role of procedural learning and language development and disorder (Matheson, 2019). Systematically examining the stability of the SRTT also has clear methodological value, in revealing design modifications to enhance its psychometric properties, and clinical value, in working towards developing a tool that can identify procedural learning weaknesses (Berchtold, 2016). Generally, it has been claimed that a larger number of trials in any task tends to increase reliability, due to a reduction in measurement error (D. H. Baker et al., 2021; Rouder et al., 2019; Rouder & Haaf, 2019, 2021). However, studies by West and colleagues (West, 2018; West, Shanks, et al., 2021) showed only modest (and non-significant) numerical improvements in test-retest reliability when they increased the number of trials in their SRTT.

In addition to examining reliability, agreement, also called repeatability, was examined using the Bland-Altman method (Bland & Altman, 1986, 1999, 2003, 2010). As argued by Berchtold (2016), the concept of test-retest refers to both the reliability and agreement of a measurement tool, with agreement referring to the ability of a test to produce the same scores when participants are tested under the same conditions. Thus, whilst reliability reflects the test's ability to rank participants consistently within or across sessions, agreement instead focuses on the consistency of the scores, independently of the range and distribution of the variables. Thus, proving particularly important for clinical applications whereby participants' scores, instead of ranking order, may be used to track response to intervention.

Therefore, here, we examine further factors that may influence stability. Of particular focus here are the similarity of the sequences to be learned (Experiment 1) and the number of sessions across which learning is assessed (Experiment 2). To allow for a comprehensive understanding of reliability, a multi-measurement analytic approach will be taken: we will assess the psychometric properties of the SRTT across different measures of procedural learning (difference scores or random slopes) and different psychometric measures (split-half reliability, test-retest reliability, and agreement).

Experiment 1

There are several reasons why the similarity of sequences to be learned over two or more sessions may influence both the size of the procedural learning effect and potentially also its stability, and each predicts that greater similarity between sequences should result in better learning at later sessions. First, learning the same or similar sequences reduces the likelihood of proactive interference, in which the memory of the first-learned sequence disrupts the learning of the second-learned sequence (Borragán et al., 2015; Darby & Sloutsky, 2015). Second, greater similarity increases the likelihood that consolidation of the first sequence will benefit learning of the second, such that individuals benefit from prior knowledge when exposed to the new material (e.g., Robertson et al., 2004; Siegelman & Frost, 2015; Nemeth et al., 2010). Third, the well-established phenomenon of practice effects is likely to lead to an improvement in performance for later sessions (Hausknecht et al., 2007; Scharfen et al., 2018), which is why the use of alternate forms is generally recommended (e.g., Beglinger et al., 2005; although see Scharfen et al. (2018) for evidence that alternate forms do not reduce practice effects in working memory capacity tasks). Finally, greater similarity may also lead to increased explicit awareness of the sequence at subsequent sessions and improve performance (e.g., Rüsseler et al., 2003) as explicit knowledge has been shown to increase with extended training in the SRTT and is more likely to lead to offline consolidation (Robertson, Pascual-Leone, & Press, 2004).

While greater similarity in sequences used in different sessions may result in larger procedural learning effects in later sessions, they may also reduce the stability of procedural learning (Stark-Inbar et al., 2017). Individual differences in any one of the above factors would introduce variability in procedural learning at retest, thus leading to changes in the rank order of scores (Hedge et al., 2018; Stark-Inbar et al., 2017). Practice effects have been shown to vary according to participants' characteristics (e.g., age: R. M. Brown et al., 2009; Hodel et al., 2014), and cognitive skills (Schaefer & Duff, 2017), thus introducing additional variability at retest. To our knowledge there has been no direct examination of the effect of sequence similarity on either the magnitude of the procedural learning effect, or the test-retest reliability of the SRTT. However, two recent studies in the literature are consistent with our prediction: Siegelman and Frost (2015) used the same sequences at both testing sessions and reported lower test-retest reliability than West, Shanks, et al. (2021) who used different sequences. Whilst West, Shanks, et al. (2021) showed no significant differences in the learning effect between sessions, Siegelman and Frost (2015), on the other hand, reported that after 3 months the majority of participants (64 out of 75) showed a better performance at retest.

Experiment 1 examined the effect of similarity of the two sequences to be learned, in order to ascertain a) the impact on the magnitude of the procedural learning effect, and b) the effect on test-retest reliability (referred to here as stability). Similarity was operationalised in terms of the Levenshtein Distance, which has been widely used to determine the distance between strings across fields such as biology, computer science and linguistics (e.g., Berger et al., 2021; N. Eriksen & Tougaard, 2006; Faes et al., 2016; Konstantinidis, 2005). Three types of operations are considered - substitutions, deletions, and insertions - with a small distance between sequences indicating higher similarity and a large distance revealing that the sequences are dissimilar (Levenshtein, 1966). We used sequences of varying similarity in a probabilistic SRTT to test four main hypotheses:

- H1: Participants will demonstrate procedural learning in both sessions, as indexed by faster responses to probable vs improbable elements of the sequence
- H2: Similarity between sequences will impact the magnitude of the procedural learning. Higher levels of similarity between Sessions 1 and 2 will result in a larger procedural learning effect in Session 2, whereas lower levels of similarity between Sessions 1 and 2 will result in a relatively smaller of procedural learning effect
- H3: Within session reliability (indexed by the split-half correlation coefficient) will be higher than stability across sessions, indexed by test-retest reliability.

- H4: Sequence similarity will be negatively associated with stability: more similar sequences at Sessions 1 and 2 will be associated with lower test-retest reliability.

Methods

Participants

One hundred and three undergraduate students from the University of York (91 females), aged between 18 and 25 years (M = 19.18, SD = 1.09), participated in exchange for course credit. The sample included monolingual, bilingual and multilingual individuals from various nationalities; all identified as fluent English speakers. The sample size was determined based on (West, Shanks, et al., 2021), doubling the number of participants to allow for a median split of participants based on similarity of the sequence. The experiment was approved by the Ethics Committee of the Psychology Department at the University of York and each participant gave written informed consent.

Measures

Serial reaction time task (SRTT)

A nonverbal probabilistic SRTT was used, following West and colleagues (2018; 2021) given the task used in this previous study has produced the highest reported stability in the existing literature. On each trial, four black outlined rectangles were presented horizontally, and a stimulus appeared in one of the four rectangles, with participants asked to respond as quickly and accurately as possible by pressing one of four corresponding keys (Z, X, N, M) on the keyboard. The stimulus remained visible until the key press. Participants rested their index and middle fingers of each hand on the four keys so they were ready to respond.

Two versions of this task were generated, each containing two different underlying second-order conditional 12 item sequences. The first two sequences were taken from Shanks et al. (2003) probable sequence A – 314324213412; improbable sequence B – 431241321423, while the second sequences were taken from Schvaneveldt and Gomez (1998): probable sequence C – 121342314324; improbable sequence D – 123413214243. In second order conditional sequences, each trial can be predicted based on the previous two trials (Schwarb & Schumacher, 2012). For each SRTT, each block started with the consecutive generation of two random digits (e.g., 21), with that bigram then followed by the digit in sequence A (e.g., 3) with 90% of probability or followed by the digit in sequence B (e.g., 4) with 10% probability (after West et al., 2018; West, Shanks, et al., 2021). After each response a new bigram was created which continuously followed the same principles. See Appendix A for a series of simulations manipulating i) the overall number of trials, and ii) the ratio between trials per condition).

The task comprised 1000 trials, as in West, Shanks, et al. (2021), divided into 20 blocks of 50 trials each. Within each block, trials immediately followed the participants' response, with no interstimulus interval. Breaks between blocks comprised a fixation cross presented centrally on screen for a random duration between 8 and 12 seconds. The stimuli were programmed in *Psychopy* 2 (Peirce et al., 2019); response accuracy and RTs (from stimulus onset) were recorded.

Sequence Similarity

Varying the degree of similarity between inputs (Appendix B) was achieved by generating a new stimulus set for each participant, by randomly matching the digits of the sequence [1, 2, 3, 4] to a different position on screen [left, centre left, centre right, right]. In order to achieve variability in the stimulus sets, half of the participants were exposed to stimuli that conformed to the same sequence structure at session 1 and 2 whilst others were exposed to stimuli that were generated by different sequence structures at both time points. Crucially, none of the participants was exposed to the exact same stimuli set at both time points. A measure of similarity of the resulting sequences actually presented to each participant was computed using the Levenshtein distance. Levenshtein distance computes the minimum operations required (insertion, deletion and substitution) for both strings to be identical, thus providing an indication of similarity between stimulus sets (Levenshtein, 1966). The Levenshtein distance was calculated for each participant by comparing the stimulus sets, i.e., 2 sets of 1000 trials. Across participants, the Levenshtein distance between pairs of stimuli varied between 248 - 437. The ratio index of the total number of triplets in common between sequences was also computed. Given the use of second-order conditional sequences, whose minimum unit of sequential information is three sequential locations or triplets, this additional computation ensured that these triplets were captured by the Levenshtein distance scores. Pearson's correlations between the Levenshtein distance scores and the ratio index revealed a high correlation between measures (r = .86).

Procedure

All participants were tested individually or in a quiet testing room in groups of up to six. All participants performed the SRTT at both sessions (SRTT1 refers to SRTT at Session 1; SRTT2 for Session 2). Each session lasted approximately 30 minutes, with Session 2 occurring one week after Session 1 for all but two participants, who were tested 9 and 10 days apart. Once the SRT2 task was completed, task enjoyment and explicit knowledge were assessed via a question and a generation task, to ensure that the levels of explicit awareness were equivalent to previous studies using probabilistic tasks (see Appendix C).

Statistical analyses

R software - version 4.1.1 (Rstudio Team, 2020) and *lme4* package (Bates et al., 2015) were used to perform two separate linear mixed effects analyses of the performance of the participants on the SRTT and all figures produced using the package *ggplot2* (Wickham, 2016). P-values were obtained for the linear mixed effects model using the *lmerTest* package (Kuznetsova et al., 2017) and corrected for multiple comparisons using the Holm-Bonferroni method (Holm, 1979). Thus, all statements regarding significance reflect the analyses after correction for family-wise error rates.

For the following data analyses, RTs were grouped into epochs of 4 blocks, comprising 200 trials. The first two trials of each block were removed as these were not predictable since the sequence follows a higher order structure with the third trial being predicted based on the previous bigram (2 trials). All incorrect trials were removed from the analyses. Due to the unequal number of probable and improbable trials, a moving criterion based on sample size was used to identify outlier RTs (Cousineau & Chartier, 2010; Van Selst & Jolicoeur, 1994). Participants with overall RTs > 2.5 standard deviations (SD) from overall mean were excluded from the analyses (based on z scores averaged across probable and improbable conditions for each group/session separately). Two participants were removed from the analyses for both sessions whilst the remaining two participants were removed for one of the sessions.

Since RTs were right skewed based on visual inspection and tests of normality, a log transformation was used to normalise the distribution of RTs (Brysbaert & Stevens, 2018). Visual inspection of the residual plots after log transformation did not reveal any obvious deviations from homoscedasticity or normality.

The fixed-effects structure represented the maximal-fixed-effects structure. The random intercept structure included solely participants, as item order was not consistent across participants due to randomisation procedures. The random structure followed the forwards best path approach (Barr et al., 2013) starting from the minimal intercepts-only structure and building the random structure according to Likelihood-ratio tests ($p < 0.2$) (Barr et al., 2013) and the Akaike Information Criterion (AIC; Akaike, 1974) to avoid overfitting (Brewer et al., 2016).

H1 And H2: The Procedural Learning Effect and Similarity

The first model - RT model, designed to explore the procedural learning skills of the sample, included the within-group variables – *probability* (probable or improbable), *epoch* (contrasts between successive epochs 2-1, 3-2, 4-3, 5-4) and *session* (1 or 2) into a linear mixed effects model, with *participants* as a random effect, in order to account for participant variability in performing the SRTT, and *Session*, *Epoch* and *Probability* as random slopes. The second model - similarity model - was formulated to explore the relationship between similarity and procedural learning in more detail. Due to the continuous nature of the similarity variable, it was centred and standardised before running the analysis. Across all models, dichotomous variables were contrasted using effect coding. The model with similarity included only RTs from the last 3 epochs to avoid the inclusion of epochs where procedural learning is not yet robust as suggested by Conway et al. (2019). *Probability* (probable or improbable), *Session* (1 or 2) and *Similarity* were entered as fixed effects and *Participants* as a random effect. Thus, unlike the first model, *Epoch* was not included as the goal was to explore the role of similarity when procedural learning was more robust independently of its progression across epochs. After building the random structure following the method previously described, *Session* and *Probability* were included as a random slope.

After model selection, the *influence.ME* package was used to detect influential data as these values may lead to changes in regression estimates (Nieuwenhuis et al., 2012). Dfbetas were standardised and participants whose z-scores were greater than +/-3.29 were identified as influential cases for as opposed to the 2.5 SD threshold to avoid loss of a high number of participants (Walker et al., 2020). Three participants were identified as influential cases for the response times model and four for the similarity model.

H3 and H4: Reliability and agreement

Test-retest and split-half reliability of the RTs were analysed using Pearson correlations, with reliability of .70 or greater being considered adequate (Nunnally & Bernstein, 1994). Two[1] different indices of procedural learning, commonly used in previous studies, were computed to better capture stability. Simple difference scores, the most commonly used measure for the SRTT, were computed for each participant as the simple difference between improbable and probable RTs, with a positive

[1] Ratio scores were also computed taking individual differences in baseline RT into account by dividing participants' difference scores by their overall mean RT per session. These yielded lower reliability than the regression slope scores; full details are reported in the Appendix D.

value indicating procedural learning. Random slopes for each participant/session were obtained by running a linear mixed effects model with response time as a dependent variable and probability as a predictor, for the random structure participants were introduced as a random intercept and probability as a random slope (Lammertink, Boersma, Wijnen, et al., 2020; Llompart & Dąbrowska, 2020; Milin et al., 2017). Random slopes were computed as this measure better captures the learning trajectory for each participant and are less likely to be influenced by extreme scores.

To measure split-half reliability for both sessions, trials were separated into probable and improbable trials. Consecutive trials were labelled as odd or even. Split-half reliability was calculated by correlating the overall mean difference in RTs for even and odd trials. The split-half and test-retest reliability were computed both for the entire task and the last 600 trials, following the suggestion that the later stages of procedural learning may be more stable (Conway et al., 2019). Agreement was examined using the Bland-Altman method (Bland & Altman, 1986, 1999, 2003, 2010). The Bland-Altman method involves plotting the mean of the measures for each participant (e.g., (Diff2 + Diff 1)/2) against the difference of the paired measurements in sessions 2 and 1 (e.g., Diff2 - Diff 1), with 95% of the data points being expected to lie within ± 1.96 standard deviations of the mean difference, referred to as the 95 % limits of agreements. According to Bland and Altman (1986, 1999, 2003, 2010), whilst a consistent tendency in the scores where performance is superior in one of the sessions than the other can be adjusted for by subtracting the difference between sessions from the one with higher scores (bias), wide limits of agreement pose a more serious problem. Determining whether the limits are adequate will depend on how precise the instrument must be for its use in clinical or research settings.

Results

Data were available for 100/103 participants for Session 1 and for 98/103 participants for Session 2. Data from five participants were lost due to computer malfunction and one due to a participant being unable to attend the second session. Four of these participants contributed data for one of the sessions, but two participants' data was lost for both sessions. Three participants were identified as outliers for each session. Data from 97 participants for session 1 and from 95 participants for session 2 were therefore included in the analysis.

H1: Procedural learning in the SRTT

As evidenced in Figure 2.1, RTs decreased with practice as observed by faster RTs with successive epochs. There was evidence of procedural learning, with RTs faster for probable than improbable trials. This 'procedural learning effect' increased over epochs, as shown by the significant interaction between *Epoch* x *Probability* for Epoch2-1, Epoch3-2 and Epoch4-3 (no longer significant after correction for multiple comparisons), but not for the last contrast, possibly indicating a plateau in learning after epoch 4. The significant interaction between *Probability* x *Session*, indicates that participants showed a larger procedural learning effect in session 2 than session 1, but this was not significant after correction for multiple comparisons. The absence of a 3-way interaction between *Epochs* x *Probability* x *Session* indicates that the within-session progression of procedural learning was similar for both sessions.

Figure 2.1

Mean response times for probable and improbable trials per epoch and session (session 1 on the left and session 2 on the right). Bars indicate 95 % CI

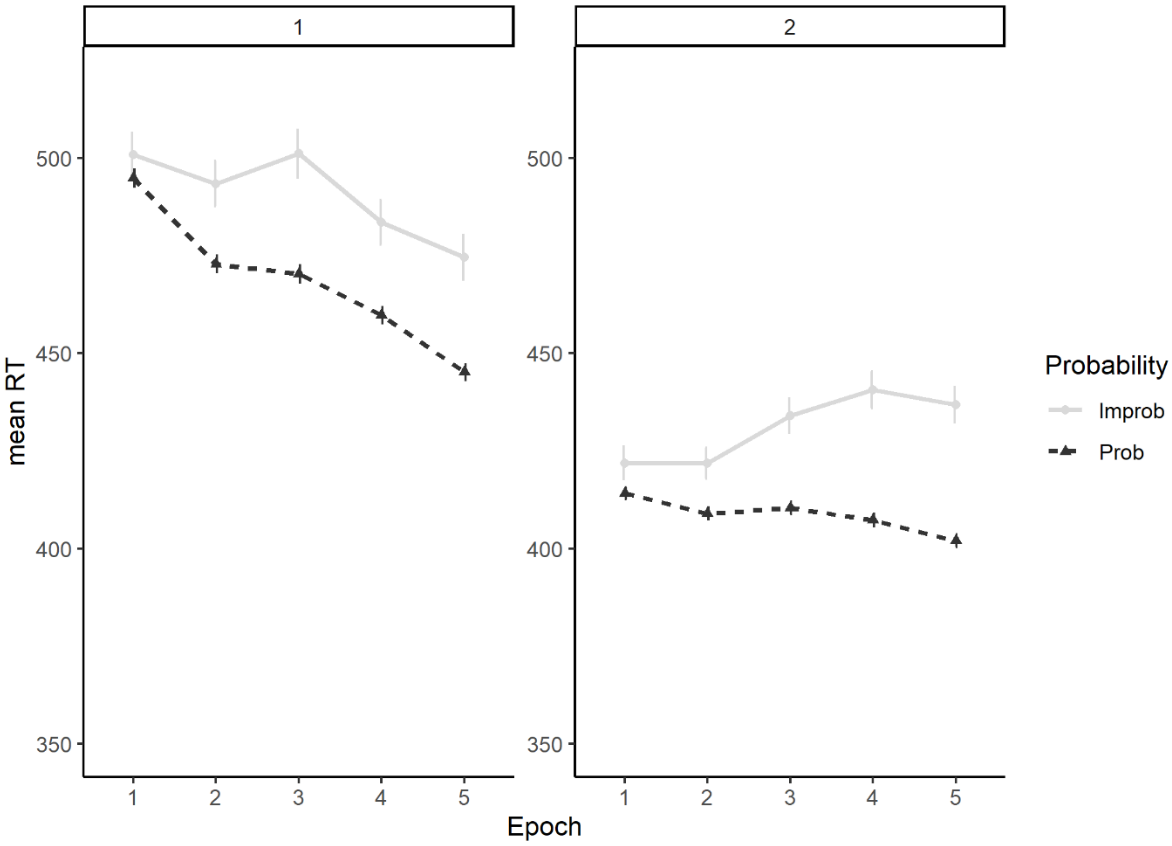


Table 2.1

Predictors of the magnitude of procedural learning

Fixed effects	*b*	*SE*	*t*	*p*	*CI*	
(Intercept)	**6.074**	**0.013**	**474.610**	**<.001**	**6.049**	**6.100**
Epoch2-1	**-0.019**	**0.004**	**-4.259**	**<.001**	**-0.028**	**-0.010**
Epoch3-2	0.008	0.004	2.130	.035	0.001	0.015
Epoch4-3	**-0.010**	**0.003**	**-3.001**	**.003**	**-0.017**	**-0.004**
Epoch5-4	**-0.019**	**0.004**	**-4.921**	**<.001**	**-0.026**	**-0.011**
Probability	**0.024**	**0.001**	**15.806**	**<.001**	**0.021**	**0.027**
Session	**0.061**	**0.004**	**16.080**	**<.001**	**0.054**	**0.069**
Epoch2-1 x Probability	**0.011**	**0.002**	**5.020**	**<.001**	**0.007**	**0.016**
Epoch3-2 x Probability	**0.012**	**0.002**	**5.018**	**<.001**	**0.007**	**0.016**
Epoch4-3 x Probability	0.005	0.002	2.309	.021	0.001	0.010
Epoch5-4 x Probability	0.004	0.002	1.653	.098	-0.001	0.008
Epoch2-1 x Session	**-0.017**	**0.004**	**-3.816**	**<.001**	**-0.025**	**-0.008**
Epoch3-2 x Session	-0.004	0.003	-1.392	.166	-0.010	0.002
Epoch4-3 x Session	**-0.012**	**0.003**	**-3.398**	**<.001**	**-0.019**	**-0.005**
Epoch5-4 x Session	-0.008	0.003	-2.328	.021	-0.015	-0.001
Probability x Session	-0.002	0.001	-2.297	.022	-0.003	0.000
Epoch2-1 x Probability x Session	0.001	0.002	0.618	.537	-0.003	0.006
Epoch3-2 x Probability x Session	-0.001	0.002	-0.396	.692	-0.005	0.004
Epoch4-3 x Probability x Session	-0.004	0.002	-1.638	.102	-0.008	0.001
Epoch5-4 x Probability x Session	0.000	0.002	0.079	.937	-0.004	0.005

Random effects	Variance	SD
Participant: (Intercept)	0.016	0.125
Participant: Session (Slope)	0.001	0.036
Participant: Epoch2-1 (Slope)	0.001	0.037

Participant: Epoch3-2 (Slope)	0.001	0.029
Participant: Epoch4-3 (Slope)	0.001	0.025
Participant: Epoch5-4 (Slope)	0.001	0.029
Participant: Probability (Slope)	0.000	0.013
Participant: Session x Epoch2-1 (Slope)	0.001	0.036
Participant: Session x Epoch3-2 (Slope)	0.000	0.020
Participant: Session x Epoch4-3 (Slope)	0.001	0.025
Participant: Session x Epoch5-4 (Slope)	0.001	0.025

H2: The effect of similarity on procedural learning

In the model incorporating sequence similarity, a similar pattern of results was obtained in terms of significant effects of *probability* and *session*. Turning to the effect of similarity, in line with our predictions, *Similarity and Similarity x Probability* were not significant predictors of RT, but there were *Probability* x *Session* x *Similarity* interactions. This indicates that greater similarity was associated with larger procedural learning effects in session 2. This was further examined by Pearson correlations between the similarity scores (Levenshtein Distance) for each participant and their procedural learning effect (for each session separately). As expected, similarity and procedural learning were not significantly correlated in Session 1 (given sequence similarity between the two sessions should have no effect on session 1) (overall: *r*(91) = .09, *p* = .40, 95% CI [-.12, .29]; last 600 trials: *r*(92)= .11, *p* = .271, 95% CI [-.09, .31]); but were moderately negatively correlated in Session 2 (overall: *r*(91) = -.34, *p* <.001, 95% CI [-.51., -.14]; last 600 trials: *r*(91) = -.34, *p* <.001, 95% CI [-.51, -.15]). This further confirms that participants who were exposed to more similar sequences in session 1 and 2 demonstrated larger procedural learning effects in session 2 (Figure 2.2).

Table 2.2

Predictors of the similarity effect on the magnitude of procedural learning

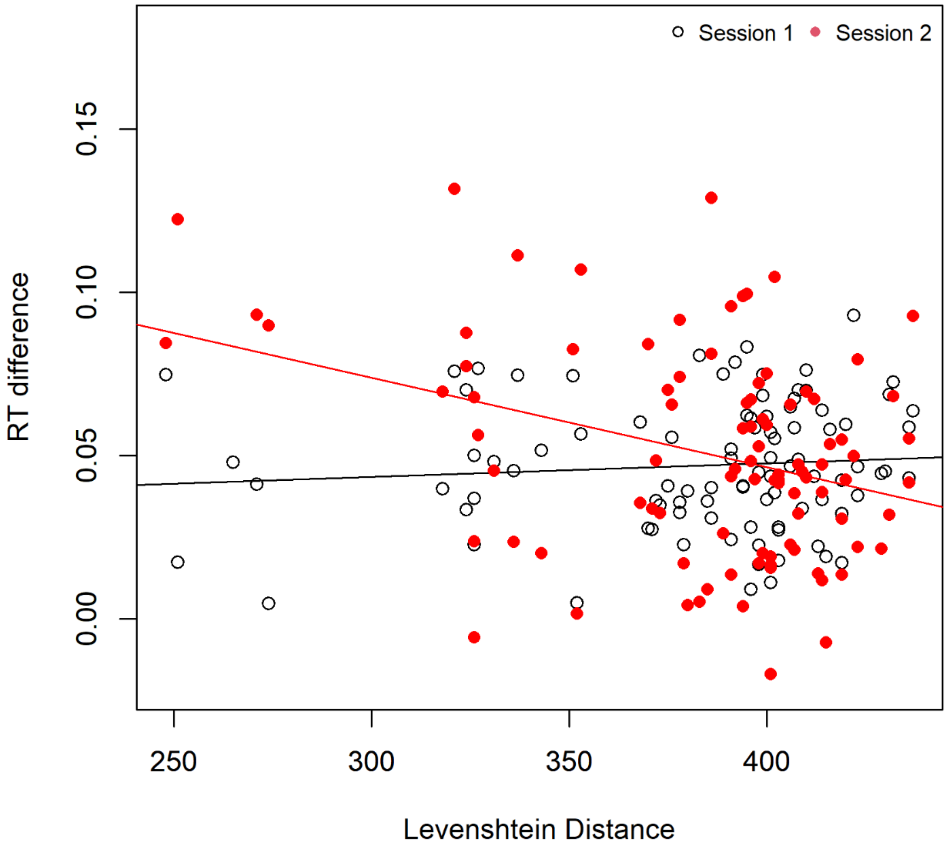
Fixed effects	*b*	*SE*	*t*	*p*	*CI*	
(Intercept)	**6.068**	**0.013**	**474.508**	**<.001**	**6.042**	**6.068**
Probability	**0.033**	**0.002**	**19.253**	**<.001**	**0.030**	**0.033**
Session	**0.051**	**0.003**	**15.181**	**<.001**	**0.044**	**0.051**

Similarity	-0.016	0.015	-1.060	.292	-0.046	-0.016
Probability x Session	-0.002	0.002	-1.524	.131	-0.006	-0.002
Probability x Similarity	-0.003	0.002	-1.518	.133	-0.007	-0.003
Session x Similarity	-0.005	0.004	-1.204	.232	-0.013	-0.005
Probability x Session x Similarity	**0.006**	**0.002**	**3.178**	**.002**	**0.002**	**0.006**

Random effects	Variance	SD
Participant (Intercept)	0.015	0.120
Participant: Session (Slope)	0.001	0.030
Participant: Probability (Slope)	0.000	0.013
Participant: Session x Probability (Slope)	0.000	0.012

Figure 2.2

Relationship between Levenshtein Distance and difference scores for both sessions for the last 600 trials



H3: Reliability

Split-half reliability (see Table 2.3) was very similar at both sessions for the overall task and last 600 trials; using random slopes rather than raw difference scores as the metric of learning yielded numerically higher estimates of reliability. The split-half coefficients ranged from .55 to .73 (>.70 is considered adequate; (Furr & Bacharach, 2008).

Test-retest reliability of the RTs themselves (e.g., the RT for probable trials in Session 1 with the RT for probable trials in Session 2) was high with a value equal or superior to .80. However, test-retest reliability of procedural learning effect was poor (*r* = .08-.17) irrespective of which measure was used and whether all RTs were included or just the final 600 trials (Table 2.3).

Table 2.3

Split-half and test-retest reliability of the procedural learning measures for overall and last 600 trials of the SRTT for session 1 (SRTT1) and session 2 (SRTT2)

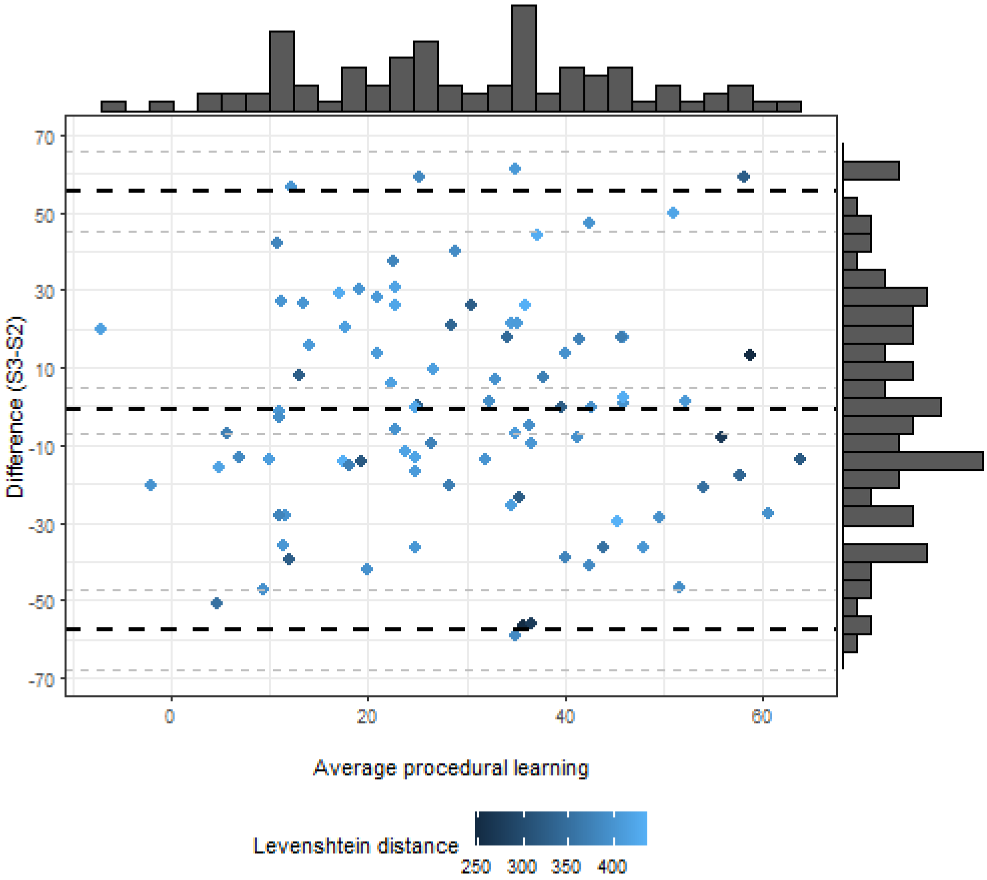
Session	***Trials***	***Split-half reliability***				***Test-retest reliability***			
		N	Difference scores	N	Random slope	N	Difference scores	N	Random slopes
SRTT1	1000	95	.55	95	.68	91	.14	91	.17
	Last 600	94	.50	94	.71	91	.08	91	.17
SRTT2	1000	91	.62	94	.70	91	.14	91	.17
	Last 600	93	.55	93	.63	91	.08	91	.17

Note. Split-half reliability correlations are significant (p <.05); test-retest reliability correlations are non-significant (*p*s > .05)

The levels of agreement between difference scores were explored via Bland-Altman plots (Figure 2.3). The Bland-Altman plots for the difference scores reveal that very few data points lie outside the limits of agreement (-57.53, 55.47), with a mean difference of -1.03; CI = [-7.03; 4.98]. However, although most data points lie within the limits of agreement, there are still considerable discrepancies between time points as evidenced by the poor precision of these limits, indicating a high degree of variance between sessions compared to between-subject variance, thus suggesting that the degree of agreement is not acceptable (Bland & Altman, 1986, 1999, 2003, 2010).

Figure 2.3

Plot of the procedural learning mean in session 1 and session 2 (x axis) against the differences between these measures (y axis). Black dashed line in the centre indicates the overall mean and the blues lines at the top and bottom represent 95% limits of agreement. Grey dashed lines represent CI around each measure.



H4: Similarity and test-retest reliability

Following the significant interaction between similarity and procedural learning, test-retest reliability was compared for participants with low and high sequence similarity scores (achieved by performing a median split). Test-retest reliability was poor for both the high- and low-similarity groups, with no significant differences between groups (overall task: z = .83, p = .41; last 600 trials: z = .15, p = .88) (Table 2.4).

Table 2.4

Test-retest reliability of the procedural learning measures for high and low similarity groups measured for overall and last 600 trials of the SRTT

Similarity	***Random slopes***		***Test-retest reliability***
	Trials	**N**	Random Slopes
low	1000	46	.30
	Last 600	46	.22
high	1000	47	.13
	Last 600	48	.20

Note: all correlations are non-significant (ps > .05)

Discussion

Experiment 1 examined the reliability of the procedural learning effect, as captured by a probabilistic SRTT, and examined the impact of the similarity of the sequences on the magnitude and stability of procedural learning. As expected, robust procedural learning effects (i.e., faster responses to probable than improbable trials) were observed. However, the level of procedural learning in a subsequent session was substantially influenced by how similar the new sequence was to a previously learned sequence. That is, greater similarity between sequences was associated with larger procedural learning effects for the new sequences. Furthermore, despite observing adequate levels of split-half reliability within each session (random slopes: .68 - .72), test-retest reliability was very poor, regardless of the level of similarity between sequences (r < .18).

The positive correlation between the procedural learning effect and sequence similarity aligns with previous results (e.g., Siegelman & Frost, 2015). West, Shanks, et al. (2021) tested participants on a probabilistic SRTT, with a 3–4-day interval between sessions and found no significant differences in performance between sessions. However, West al. used distinct sequences at test and retest with the aim of reducing practice effects. Together with the present results, these studies suggest that the SRTT is prone to practice-effects when subsequent sessions use similar sequences. The present study cannot speak to the mechanism/s that underlie the benefit of similarity on procedural learning. However, in light of the lack of evidence for a relationship between explicit awareness and the level of similarity between sequences (see Appendix E), one possibility is that consolidated knowledge of the first-learned sequence aids the acquisition of the second-learnt sequence (R. M. Brown et al., 2009; Press et al., 2005; Robertson, Pascual-Leone, & Miall, 2004) or that knowledge of the first-learned sequence proactively interferes with the acquisition of the second-learnt sequence (Desmottes, Maillart, et al., 2017).

The suboptimal test-retest reliability of the SRTT observed here is also generally consistent with previous findings. However, our test-retest coefficients were considerably lower than Siegelman and Frost (2015) (r = .47) and West, Shanks, et al. (2021) (*r* = .70), irrespective of similarity between sequences at both time points. Our coefficients are more akin to those obtained by West et al. (2018, 2021) in children (*r* = .21; *r* = .26, respectively). This low stability of the procedural learning effect is striking, particularly in the context of robust group-level procedural learning effects and despite high stability of overall RTs. One possibility is that difference scores, in general, are intrinsically less reliable than their component parts. This has been suggested by Hedge et al. (2018) as difference scores contain measurement error from both measures which leads to an increase in the proportion of measurement error relative to between-subject variance. Yet, the limitations of using difference scores does not seem to pose an issue when analysing the split-half reliability, nor does it explain the better test-retest reliability observed by Siegelman and Frost (2015) and West et al. (2018) despite also analysing difference scores. Furthermore, if difference scores were solely responsible for poor reliability, one would expect better outcomes for the random slopes. Unfortunately, that was not the case. Thus, other factors must contribute to the pattern of lower stability than split-half reliability.

It is possible that specific differences in design between our experiment and West, Shanks, et al. (2021) can account for the divergent findings. Firstly, West, Shanks, et al. (2021) recruited older participants (18 - 61 years, mean = 25.33 years; SD = 10.33 years) than in Experiment 2 (17 - 34 years, mean = 20.09 years, SD = 2.09 years). This could have contributed to increasing the stability of the SRTT as test-retest reliability has been found to increase with age in intelligence measures (Schuerger & Witt, 1989). Whilst presentation rates and age of participants have been shown to affect the procedural learning effect in the SRTT (presentation rates: e.g., Arciuli & Simpson, 2011; Emberson et al., 2011; Frensch & Miner, 1994; Soetens et al., 2004; Willingham et al., 1997; age: e.g., R. M. Brown et al., 2009; Juhasz et al., 2019) there is no evidence, to our knowledge, of its impact on the test-retest reliability of the task. Secondly, West, Shanks, et al. (2021) included a 250ms interstimulus interval (ISI) between trials which was absent in our experiment with the aim of reducing explicit awareness (Destrebecqz & Cleeremans, 2001). The inclusion of an ISI, however, could have contributed to the higher test-retest reliability by inducing stronger representations of the sequence (Cleeremans & Sarrazin, 2007; Gaillard et al., 2009), with explicit awareness possibly emerging as a consequence of the increased signal strength (Cleeremans, 2011; Timmermans et al., 2012). However, our data did not show indication that the magnitude of procedural learning was associated with explicit awareness (for more details see Appendix C). Furthermore, a follow-up experiment (fully described in Appendix E) replicated more closely the design adopted by West, Shanks, et al. (2021) by including a 250ms ISI and participants with ages between 18 - 60. Yet, this experiment still revealed suboptimal test-retest reliability (*r* < 21). Explicit awareness levels were also similar between groups with and without an ISI. Taken together, this suggests that the superior reliability observed by West and colleagues (West, Shanks, et al., 2021) may be explained by other design or sampling factors.

In sum, Experiment 1 obtained clear evidence of procedural learning, which was larger in the second session, particularly when the second-learned sequences were more similar to the first-learned sequences. However, test-retest reliability of procedural learning was very poor regardless of the level of similarity between sequences. Another possibility, examined in Experiment 2, is whether this variability in the procedural learning effect across sessions will diminish with further training - that is, that individuals will eventually reach a 'plateau' which more accurately reflects their intrinsic procedural learning capacity. Given the lack of evidence for any impact of sequence similarity on reliability of the SRTT, and the larger procedural learning effect for those learning sequences with higher similarity, sequences with high similarity were adopted in Experiment 2 in order to maximise the chances of participants reaching a "plateau" at an earlier stage of learning.

Experiment 2

Experiment 2 examined whether the inclusion of three sessions would increase the test-retest reliability of the SRTT, since, as suggested by Conway et al. (2019) the poor reliability of probabilistic procedural learning may be related to the measurement of earlier stages when learning might not be as robust. Palmer et al. (2018) have demonstrated patterns of increased stability on a variety of measures of cognitive ability commonly used to assess striatal dysfunction by increasing the number of training sessions. They reported that practice effects diminished in patients with striatal impairments by the third session, thus increasing the stability of the measures. Although Palmer et al. (2018) did not consider the SRTT, it is possible that it would follow a similar stabilisation trajectory, since the striatum has also been strongly implicated in performance on this task (Robertson et al., 2001; Torriero et al., 2004).

Experiment 2 also carried out a preliminary examination of the relationship between procedural learning and language and literacy. According to the procedural/declarative model (Ullman et al., 2020; Ullman & Pullman, 2015), performance on language measures (particularly grammar and phonology) and literacy measures (e.g., spelling, which requires procedural learning) should be associated with procedural learning. However, such correlations have not been consistently found in previous studies. If these correlations are masked by the low stability of the SRTT and if incorporating multiple sessions increases stability, then stronger correlations would be expected with procedural learning effects measured at later sessions. This hypothesis is supported by West, Shanks, et al. (2021), who found, in their children's sample, small to moderate correlations between linguistic/literacy measures and procedural learning captured in a second session, but not a first session.

Finally, Experiment 2 considered the role of attention in relation to procedural learning stability. An extensive literature has considered the role of attention in procedural learning in the context of dual task paradigms. Such studies demonstrate a detrimental effect on procedural learning when participants simultaneously perform the SRTT alongside a secondary task (deterministic sequences: Coomans et al., 2014; Schumacher & Schwarb, 2009; D. Shanks & Channon, 2002; probabilistic sequences: D. R. Shanks et al., 2005). In line with this, a positive correlation between sustained attention and procedural learning in children has been found by Sengottuvel and Rao (2013a) and West, Shanks, et al. (2021). In the latter, it was also observed that the attentional demands of the SRTT may vary depending on the session: although attention was found to positively correlate with procedural learning at both sessions, stronger correlations were observed for session 2. Furthermore, when attention was entered as a predictor of children's attainment (on measures of reading, grammar, and arithmetic), in a latent variable path model which also included the SRTT, measures of declarative learning and attention, attention and declarative memory contributed unique variance, but the SRTT did not. This suggests that while the SRT may be a weak correlate of language and related skills, this may be the result of overlapping variance with other variables, such as attention. This is further supported by the strong correlation between attention and procedural memory (r = .56) observed in West, Shanks, et al. (2021)**.**

However, in West, Shanks, et al. (2021), a 9-point observational rating scale was used to estimate the levels of attention throughout the SRTT, whilst Sengottuvel and Rao (2013a) assessed the offline attention skills through a Two Choice Reaction Time task. For both attentional tasks information regarding their psychometric properties is lacking, with the operationalisation of attention used by West, Shanks, et al. (2021) potentially tapping into other constructs such as motivation/boredom required for children to remain focussed on the task (e.g., R. S. J. d. Baker et al., 2010; Godwin et al., 2016). Here, a direct measure of attention (i.e., a psychomotor vigilance task) was adopted to further explore the relationship between procedural learning and attention.

Experiment 2 used the same SRTT as in Experiment 1 but on three separate sessions, to address the following pre-registered hypotheses (https://osf.io/yb3sv):

- H1: Participants are expected to demonstrate evidence of procedural learning in all three sessions.
- H2: Moderate to low test-retest reliability levels are expected between Sessions 1 and 2;
- H3: If stability of performance increases with the number of sessions, test-retest reliability will be higher between session 2 and 3 than between session 1 and 2;
- H3: Split-half reliability will be higher for later sessions when compared to session 1;
- H4: Procedural learning is expected to correlate with language and literacy performance/scores in all sessions;
- H5: Higher associations between language and procedural learning will be expected in later sessions if the procedural learning effects are more reliable at later sessions;
- H6: Participants with better attention skills will be expected to show more procedural learning;
- H7: Higher correlations between procedural learning and attention are expected for later sessions.

No hypotheses were pre-registered regarding how attention influences stability between sessions as, to our knowledge, this has not been previously tested using the SRTT. Exploratory analyses were therefore performed to examine relationships between attention and stability.

Methods

Participants

Forty-seven young healthy adults aged 17 and 34 years (M = 20.11 years, SD = 2.87 years) with language, literacy, and non-verbal intelligence within the average range (see Appendix F) were recruited from the University of York. All participants were native English speakers based in the UK with normal or corrected-to-normal hearing, vision and without motor impairments that may impede task performance. Participants received payment or course credit as compensation. The experiment was approved by the Ethics Committee of the Psychology Department in the University of York and each participant gave written informed consent.

Measures

Serial Reaction Time task

The SRTT used in Experiment 1 was used here, with the exception that the 1000 trials were distributed over five blocks rather than 20 to replicate the number of blocks adopted by West et al. (West, Shanks, et al., 2021). The first two sequences adopted were the ones included in Experiment 1. A new pair of sequences was selected for the additional session. The sequences were taken from (Kaufman et al., 2010): probable sequence E - 121432413423; improbable sequence F - 323412431421. These sequences were selected to have equivalent levels of similarity (as captured by LD) and the similarity was comparable to West et al. (2018, 2020) (Sequence 1 – Sequence 2: LD = 338; Sequence 1 – Sequence 3: LD = 342; Sequence 2 – Sequence 3: LD = 374).

Sustained attention

A computerised 10-min Psychomotor vigilance task (PVT) (based on (Reifman et al., 2018) was used to measure sustained or vigilant attention by recording response times (RT) to visual stimuli presented at random interstimulus intervals between 2 and 10 seconds (ISI). When performing the PVT, participants are asked to press the spacebar as soon as a red counter appears on screen, which stops the counter and displays the RT in milliseconds for a 1-s period. Based on the study by Basner and Dinges (2011), the mean reciprocal response time (mean 1/RT) and the number of lapses, defined as response times ≥ 500 ms, were selected as primary outcome measures. Whilst the reciprocal response time shows the most superior statistical properties, i.e., being sensitive to small changes in fast RTs and robust to extreme values; the lapses, which reflect state instability, are the most commonly used measure and have high ecological validity (Basner & Dinges, 2011). The reliability of the 10-min PVT is high, with test-retest reliability for lapses and median > 0.8 in adults (Dorrian et al., 2005).

Beyond these measures on the PVT, performance variability, which may be masked by analyses based on mean performance, has been explored as a valuable source of information to better understand individual differences in learning (Henríquez-Henríquez et al., 2015). The Ex-Gaussian method allows the examination of the response time distribution both for the 'mu' and 'sigma' parameters of the Gaussian distribution, which represent the mean and standard deviation of the normal component of the distribution, but also 'tau', which represents the exponential component reflecting the slower response times and is the tail of the distribution. Previous research has found that high indices of intra-individual variability, usually higher tau values, are characteristic of populations with ADHD (Borella et al., 2011; Gooch et al., 2012). Thus, the "tau" measure was also computed since it has been proposed as a stronger marker of attention difficulties than basic RT/lapses (Castellanos et al., 2006). Hence, the "tau" metric would potentially better capture the association between procedural learning and attention.

Standardised measures

All cognitive measures were delivered and scored in accordance with manual instructions.

Nonverbal intelligence was assessed by the Matrix Reasoning subtest of the Wechsler Abbreviated Scale of Intelligence - Second Edition (WASI-II; test-retest reliability r = .82; Wechsler, 2011). This task consists of 30 incomplete visual matrices and the participants are required to choose the item from a selection of five that correctly completes the matrix.

Expressive vocabulary was assessed using the Vocabulary subtest of the Wechsler Abbreviated Scale of Intelligence - Second Edition (WASI-II; test-retest reliability r = .90; Wechsler, 2011). This task requires participants to provide a definition for a series of words that increase in difficulty, presented both verbally and orthographically. Each answer is given a score of 0, 1, or 2 points depending on the quality of the description.

Nonword repetition was assessed with the Comprehensive Test of Phonological Processing - 2 (CTOPP-2; internal consistency alpha coefficient r = .77; (Wagner et al., 2013), providing a measure of phonological memory. Participants were told that they would hear nonwords (that increased in phonological complexity) via headphones and that they should repeat the nonword exactly.

Sentence recall was measured with the Recalling Sentences task from the Clinical Evaluation of Language Fundamentals - Fifth Edition (CELF-5, test-retest reliability r = .94; (Wiig et al., 2013) was used to assess individuals' ability to repeat sentences of increasing length and complexity.

Reading and spelling were assessed with the Wechsler Individual Achievement Test, third edition UK (WIAT-III^{UK}; internal consistency coefficients r ≥.90; Wechsler, 2009). For Word Reading, participants were asked to read aloud words and nonwords ordered in increasing difficulty. Participants' responses were audio-recorded and later scored. The Spelling subtest consists of a spelling-to-dictation task containing regular and irregular words. Participants first heard the target word in isolation, then in the context of a sentence, and finally in isolation again. Dictation was conducted using a recording of a native female speaker.

Procedure

A within-subjects design was used, with each participant performing the SRTT at three time points each separated by roughly one week (session 1: mean = 7 days, SD = 0; session 2: mean = 7.02 days, SD = .15; session 3: mean = 7.09 days, SD = .58). The three underlying sequences were counterbalanced across participants and sessions to avoid order effects.

All sessions started with the administration of the SRTT (duration of approximately 15 minutes). Standardised tests were administered after the SRTT in each session (i.e., literacy and attention tests in session 1; language measures in session 2; and nonverbal measure in session 3). A generation task was completed at the end of the final session, to capture explicit knowledge of the sequence learned in session 3. Session 1 lasted roughly one hour; sessions 2 and 3 were approximately 30 minutes.

Statistical analyses

H1: Mixed effects model

The same procedures adopted in experiment 1 were adopted for data treatment and analyses in experiment 2. The additional session allowed the exploration of its effects on the stability of procedural learning. For the three-level factor of session two orthogonal contrasts were set: delay1 which contrasts session 1 with sessions 2 and 3 (S1 vs S2 & S3) and delay2 contrasted the performance in sessions 2 and 3. After model selection, three participants were identified as influential cases. The analyses reported include the influential cases as this led to no differences in result interpretation with only minor changes in the degree of significance.

H1: Reliability and agreement

As in Experiment 1, test-retest reliability was calculated between sessions 1 and 2 and sessions 2 and 3 using difference scores and random slopes as measures of procedural learning. Agreement was assessed through Bland-Altman plots.

H2 and H3: Relationship between procedural learning and cognitive measures

Pearson correlations were conducted to explore the relationship between written and oral language measures and procedural learning. The Holm-Bonferroni method was used to correct for multiple comparisons (Holm, 1979). Based on the sensitivity analysis, this study has 80% power to detect correlations equal and above .35. Since non-significant results may represent either lack of evidence for a correlation or lack of power, Bayesian Pearson correlations will be computed alongside using the BayesFactor package (Morey & Rouder, 2022), with a beta distribution with a scale of 0.333 as a prior for the correlation coefficients. Bayes factors above 3 or below ⅓ will be taken as support for the alternative or null, respectively, yet we recognise that Bayes factors should be interpreted in a continuum (Jeffreys, 1961).

Exploratory analysis of attention

Ex-Gaussian analysis was performed on the PVT and the parameters were extracted using the package Retimes (Massidda, 2013). The Ex-Gaussian distribution is characterised by a mean mu, standard deviation sigma and exponential distribution with mean tau. In this analysis we focus on the measure tau as it represents the skewness or variability of the slow responses. This measure has been shown to be a better predictor of performance than traditional response times measures on attention and inhibition tasks (Gooch et al., 2012; Henríquez-Henríquez et al., 2015; van Belle et al., 2015).

Results

All participants completed the three sessions each separated by one-week, with the exception that one participant completed session 3 11 days after session 2 and another completed session 2 8 days after Session 1. Data from all participants was available for all sessions except for one participant who missed session 3. The remaining data were included in the analyses. The performance of two other participants was identified as an outlier, one for session 1 and another for session 3.

H1: Procedural learning in the SRTT - effect of session

Participants' RTs decreased with practice (Figure 2.5) as evidenced by significant main effects of *Epoch* for contrasts Epoch2-1 (no longer significant after correction for multiple comparisons) and Epoch5-4 and *Session* for both contrasts (Delay1: Session 1 vs Session 2 and 3; Delay2: Session 2 vs Session 3). Importantly, there was a main effect of *Probability*, as response times were faster for probable than improbable trials. This difference in probable and improbable response times increased with practice as evidenced by a significant *Epoch* x *Probability* interaction, as well as a significant *Session x Probability* interaction. Yet, the interaction between *Session x Probability* for Delay2 was no longer significant after correction for multiple comparisons.

Despite this improvement in procedural learning with practice, the three-way interaction between *Epochs x Probability x Session* was only significant for *Delay1* for Epoch3-2 and Epoch4-3 (also no longer significant after correction for multiple comparisons), thus indicating a significant increase in procedural learning in Sessions 2 and 3 for Epoch3-2 relative to session 1. This difference between Session 1 and Sessions 2/3 for Epoch3-2 is apparent in Figure 2.5. The non-significant interaction for *Delay2* (session 2 vs session 3) indicates that, despite the overall gains in procedural learning from session 2 to 3, the difference between sessions was not observed at the epoch level.

Figure 2.4

Mean and 95% CI response times for probable and improbable trials per Epoch and Session (Session 1 on the left, Session 2 in the centre and session 3 on the right).

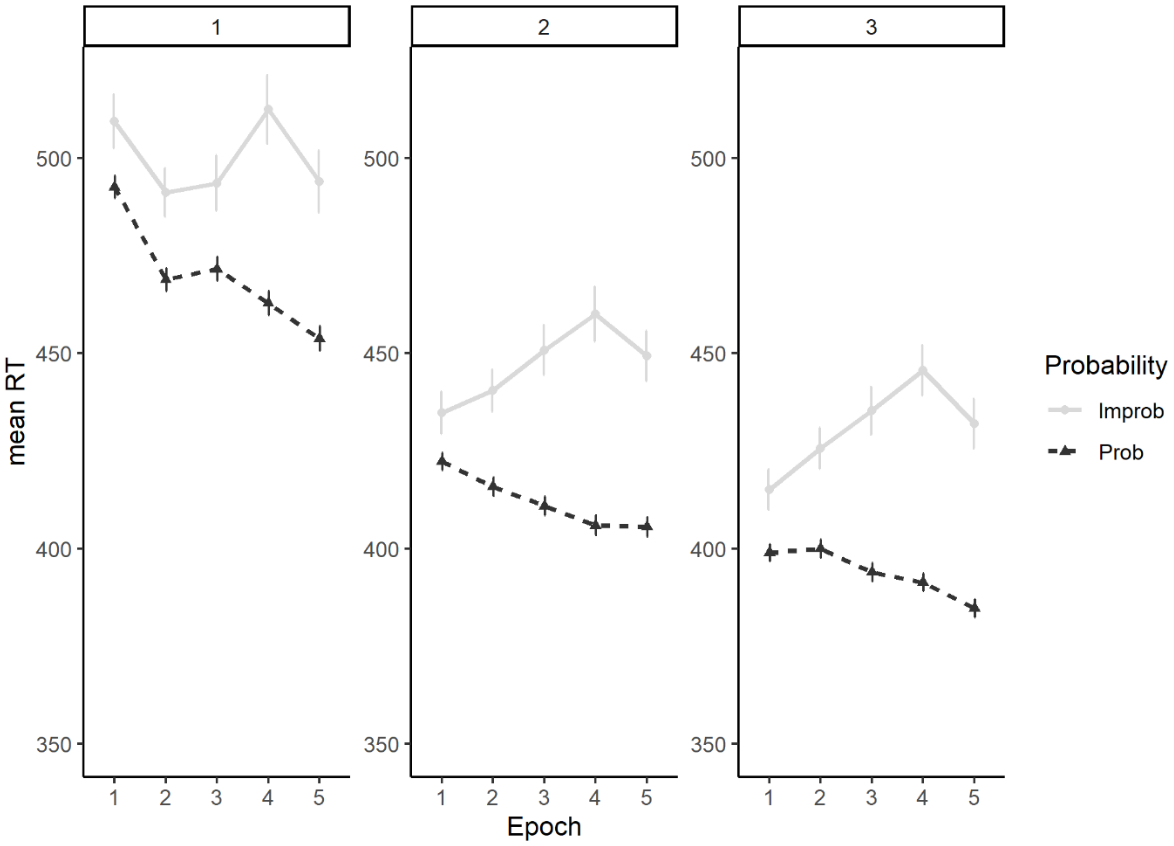


Table 2.5

The effect of an additional session on the magnitude of procedural learning

Fixed effects	*b*	*SE*	*t*	*p*	*CI*	
(Intercept)	**6.051**	**0.017**	**347.667**	**<.001**	**6.016**	**6.087**
Epoch2-1	-0.009	0.004	-2.135	.037	-0.018	-0.001
Epoch3-2	-0.002	0.004	-0.620	.538	-0.010	0.005
Epoch4-3	0.006	0.004	1.397	.167	-0.002	0.014
Epoch5-4	**-0.021**	**0.004**	**-5.099**	**<.001**	**-0.029**	**-0.012**
Probability	**0.039**	**0.002**	**21.674**	**<.001**	**0.035**	**0.043**
Delay1 (S1 vs S2 and S3)	**-0.045**	**0.003**	**-14.568**	**<.001**	**-0.051**	**-0.039**
Delay2 (S2 vs S3)	**-0.020**	**0.004**	**-5.271**	**<.001**	**-0.028**	**-0.012**
Epoch2-1 x Probability	**0.014**	**0.003**	**5.438**	**<.001**	**0.009**	**0.019**
Epoch3-2 x Probability	**0.010**	**0.003**	**4.042**	**<.001**	**0.005**	**0.015**
Epoch4-3 x Probability	**0.021**	**0.003**	**8.047**	**<.001**	**0.016**	**0.027**
Epoch5-4 x Probability	-0.008	0.003	-2.763	.006	-0.013	-0.002
Epoch2-1 x Delay1	**0.013**	**0.002**	**7.406**	**<.001**	**0.010**	**0.017**
Epoch3-2 x Delay1	0.001	0.002	0.828	.408	-0.002	0.005
Epoch4-3 x Delay1	-0.001	0.002	-0.442	.659	-0.005	0.003
Epoch5-4 x Delay1	0.003	0.002	1.524	.128	-0.001	0.007
Epoch2-1 x Delay2	0.004	0.003	1.404	.160	-0.002	0.010
Epoch3-2 x Delay2	0.003	0.003	0.954	.340	-0.003	0.009
Epoch4-3 x Delay2	0.001	0.003	0.257	.797	-0.006	0.007
Epoch5-4 x Delay2	-0.007	0.003	-2.185	.029	-0.014	-0.001
Probability1 x Delay1	**0.004**	**0.001**	**6.757**	**<.001**	**0.003**	**0.005**
Probability1 x Delay2	0.003	0.001	2.556	.011	0.001	0.005
Epoch2-1 x Probability x Delay1	0.001	0.002	0.367	.714	-0.003	0.004
Epoch3-2 x Probability x Delay1	0.005	0.002	2.644	.008	0.001	0.008
Epoch4-3 x Probability x Delay1	-0.005	0.002	-2.427	.015	-0.008	-0.001

Epoch5-4 x Probability x Delay1	0.001	0.002	0.331	.741	-0.003	0.004
Epoch2-1 x Probability x Delay2	-0.002	0.003	-0.668	.504	-0.008	0.004
Epoch3-2 x Probability x Delay2	0.002	0.003	0.494	.621	-0.005	0.008
Epoch4-3 x Probability x Delay2	0.002	0.003	0.516	.606	-0.005	0.008
Epoch5-4 x Probability x Delay2	-0.001	0.003	-0.403	.687	-0.008	0.005

Random effects	Variance	SD
Participant (Intercept)	0.000	0.113
Participant: Delay1 (Slope)	0.001	0.019
Participant: Delay2 (Slope)	0.001	0.024
Participant: Block2-1 (Slope)	0.000	0.024
Participant: Block3-2 (Slope)	0.000	0.020
Participant: Block4-3 (Slope)	0.000	0.020
Participant: Block5-4 (Slope)	0.000	0.019
Participant: Probability (Slope)	0.041	0.010

H2, H3 and H4: Reliability

As shown in Tables 2.6 and 2.7 and similar to Experiment 1, split-half reliability for the SRTT was numerically higher when using slope coefficients compared to raw difference scores and ranged from low (*r* = .34) to excellent (*r* = .91; (Cicchetti, 1994a, 1994b; Cicchetti & Sparrow, 1990). This difference reached significance in the third session for both contrasts (*p* < .001).

As in Experiment 1, overall response times were highly stable across sessions (probable trials, *r*s = .82-.89; improbable trials, *r*s = .79-.83) but the procedural learning effect showed poor stability between Sessions 1 and 2, as reported in Table 2.7. Although there was a numerical improvement in stability between sessions 2 and 3 which was most evident for the random slope metric, this numerical increase in stability was not statistically significant (overall: *z* = - 0.38, *p* = .70; last 600 trials: *z* = -1.08, *p* = .28).

Table 2.6

Split-half reliability for the procedural learning measures per session (SRTT1, Session 1; SRTT2, Session 2; SRTT3, Session 3)

Session	***Trials***	***Split-half reliability***			
		N	Difference scores	N	Random slopes
SRTT1	1000	45	.60***	45	.77***
	Last 600	44	.56***	45	.66***
SRTT2	1000	45	.55***	47	.55***
	Last 600	46	.36*	47	.56***
SRTT3	1000	43	.23	44	.81***
	Last 600	45	.29	45	.91***

Note. *p <.05, *** p<.001

Table 2.7

Pairwise test-retest reliability of the procedural learning measures

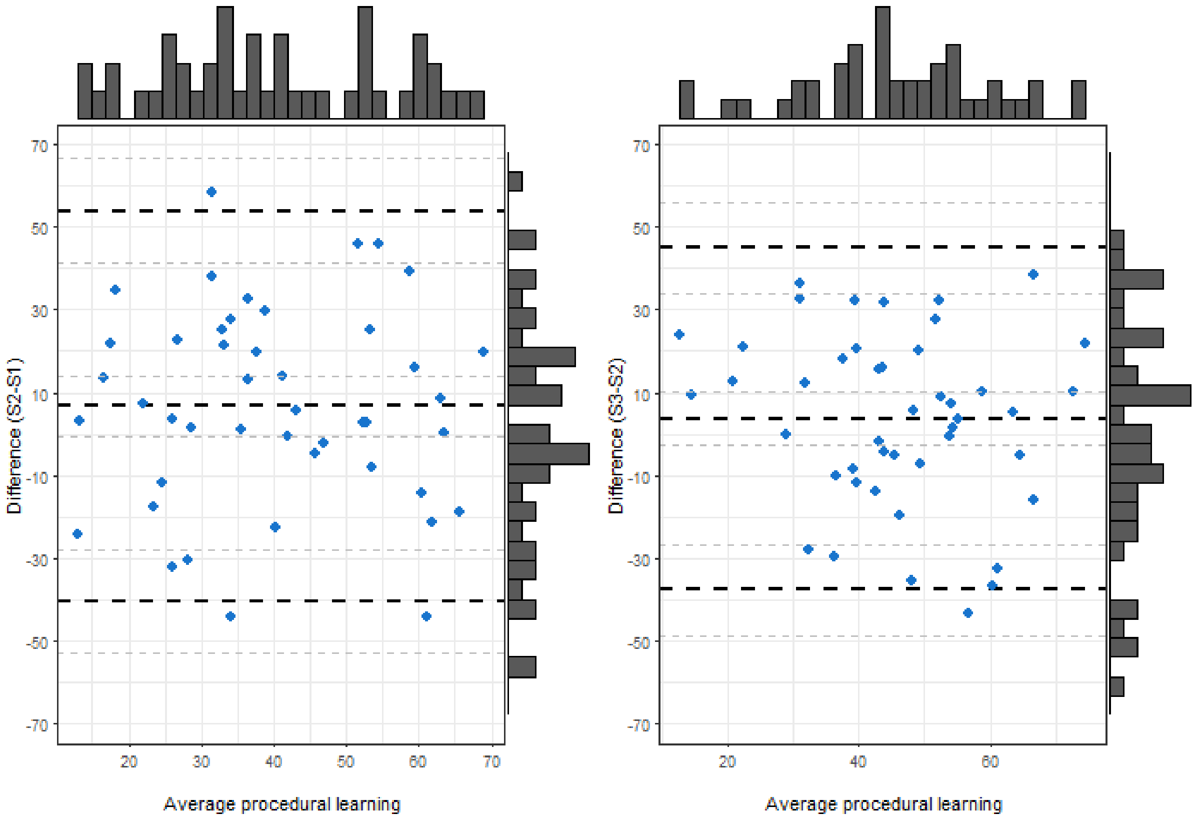
Session	***Trials***	***Test-retest reliability***	
		Difference scores	Random slopes
SRTT1 – SRTT2	1000	.22	.28
	Last 600	.25	.42**
SRTT2 – SRTT3	1000	.15	.41**
	Last 600	.30*	.60***

Note. * p < .05, ** p < .01, *** p < .001

The Bland–Altman's 95% limits of agreement range between -40.47 to 54.03 for sessions 1 and 2 and between -37.62 to 45.03 for sessions 2 and 3. Almost all participants fell within the limits of agreement, however the limits of agreement lacked precision (i.e., the magnitude of the procedural learning effect lacks consistency whereby performance on one session is not necessarily replicated in another possibly reflecting a high degree of measurement), thus revealing poor agreement between measures. Yet, the Bland-Altman plot for sessions 2 and 3 shows narrower limits of agreement, indicating an improvement in agreement for later sessions.

Figure 2.5

Plot of the mean of the two measurements against the differences between procedural learning in sessions 1 and 2 (right) and sessions 2 and session 3 (right)



H4 and H7: Relationship between procedural learning and cognitive measures

The random slopes were used as a measure of procedural learning for analyses of individual differences as this method of calculation demonstrated the highest split-half and test-retest reliability, especially between sessions 2 and 3 (see Appendix G for the Bayes factors and credible intervals for the bivariate correlations between procedural learning and cognitive measures).

Nonverbal IQ. Procedural learning was not significantly correlated with nonverbal IQ (session 1: r = -.08, BF_{10} = 0.38; session 2: r = .09, BF_{10} = .39; session 3: r = .22, BF_{10} = .87); thus, nonverbal IQ was not used as a covariate in subsequent analyses.

Language and Literacy. Vocabulary (r = .39) was the only significant language or literacy correlate of procedural learning and only in session 3, indicating that participants with higher vocabulary skills also demonstrated greater procedural learning. However, this correlation did not survive Holm-Bonferroni correction (see Table 2.8). Nonetheless, Bayesian correlations revealed that there was evidence against the null hypothesis (BF_{10} = 7.55).

Attention. A positive and significant correlation was observed between procedural learning and sustained attention for session 1 (median: *r* = -.28; BF_{10} = 1.46, reciprocal: *r* = .30, BF_{10} = 1.90) and 2 (median: *r* = -.45, BF_{10} = 29.88; reciprocal: *r* = .49, BF_{10} = 64.40); this association was smaller and non-significant for session 3 (median: *r* = -.25, BF_{10} = 1.11; reciprocal: *r* = .25, BF_{10} = 1.15). As shown in Table 3.8, there were negative and non-significant correlations for the tau parameter, which indexes intra-individual variability (M = 63.84, SD = 28.73) for all sessions (SRTT1: *r* = -.18, BF_{10} = .63; SRTT2: *r* = -.14, BF_{10} = .48; SRTT 3: *r* = -.19, BF_{10} = .68).

Table 2.8

Correlation matrix between procedural learning and cognitive measures

Measures		***Procedural learning Session 1***	***Procedural learning Session 2***	***Procedural learning Session 3***
Age		.22	-.04	-.05
Literacy	Word Reading	.04	.014	-.002
	Nonword Reading	.02	.08	.06
	Spelling	.20	.25†	.03
Language	Vocabulary	-.08	.11	.39**[a]
	Nonword Repetition	-.04	-.10	-.24
	Recalling	-.16	-.15	-.29†
Nonverbal IQ	Matrix Reasoning	-.07	.09	.22
Attention	PVT median	-.28†	**-.45****[a]	-.25
	PVT Reciprocal	.30*	**.49*****[a]	.25†
	PVT tau	-.18	-.14	-.19

Note. †p < .10; *p < .05; **p < .01; ***p < .001; bold – correlations that survived correction for multiple comparisons; [a] correlations with Bayes factor equal or bigger than 3

Given the positive relationship between attention and procedural learning, whereby individuals with better attentional skills showed better procedural learning, correlations between tau and SRTT stability were explored to examine whether individuals with high levels of intra-individual variability in attention would also show less stability in the SRTT. Using a medium split approach, the sample was divided into high- and low-tau groups. With respect to Sessions 1 and 2, moderate stability was found for both low-tau (*r* = .29) and high-tau (*r* = .42) groups (the numerical difference was non-significant: *z* = -.33, *p* = .74). However, there was a marked difference between low- and high-tau groups for test-retest stability across Sessions 2 and 3, with the low tau group showing higher test-retest stability (*r* = .73) than the high tau group (*r* = .26). Importantly, the difference between these correlations was statistically significant: *z* = 2.07, *p* = .04. That is, participants with lower intra-individual variability on the measure of sustained attention were also those with more stable procedural learning effects in the SRTT across Sessions 2 and 3.

Discussion

Experiment 2 examined the stability of procedural learning over three sessions, as well as the relationship between procedural learning and attention and language measures. As in Experiment 1, the procedural learning effect was robust in all sessions. Whilst there was some evidence of a numerical increase in reliability for the later sessions for both split-half and test-retest reliability, these improvements were not statistically significant, and stability remained suboptimal. Procedural learning positively and significantly correlated with sustained attention and, to a lesser extent, vocabulary, with the latter not surviving correction for multiple comparisons.

As predicted, the test-retest reliability of the SRTT showed numerical improvements across sessions, with stability slightly higher for later sessions. Indeed, the highest level of stability in the current experiment was between Sessions 2 and 3 when using random slopes as the index of learning (*r* = .60). This is more akin to the stability reported by Siegelman and Frost (2015) (*r* = .47) and West, Shanks, et al. (2021) (*r* = .70), although in these studies this level of stability was found across two sessions rather than three. Overall, the highest stability was observed when focusing on the procedural learning effect on the last 3 epochs, which aligns with Conway and colleagues' (2019) suggestion that the inclusion of earlier stages of procedural learning, when learning is not yet robust, may reduce test-retest reliability. Nonetheless, the linear mixed-effects model and the Bland-Altman plots indicate that, even though increasing the number of sessions reduced practice effects, there was still a significant procedural learning improvement between sessions 2 and 3. This may indicate that additional sessions may be required to reach a plateau in procedural learning; while this would be theoretically important to ascertain, it would limit the practical utility of using the SRTT in clinical or developmental research, thus placing doubt over the practical utility of this task. This pattern was observed despite adopting distinct, though similar, sequences at each session, with the aim of reducing practice effects (Palmer et al., 2018). In a recent meta-analysis on retest effects in working memory tasks, improvements in performance were observed until the 7th session, yet they were no longer significant after the 4th administration (Scharfen et al., 2018). Trial variability (i.e., the variance in the RTs for probable and improbable trials) also decreased across sessions, further suggesting that measurement error decreased across sessions, with an increase in the signal to noise ratio (Chen et al., 2021; Rouder et al., 2019).

Counter to our hypotheses, there was minimal evidence of an association between procedural learning and language. We found only a moderate correlation, that did not survive correction for multiple comparisons, between procedural learning and vocabulary Session 3. It is worth noting that this aspect of language is proposed to be more highly associated with declarative than procedural memory (Ullman, 2004). Notably, and also counter to Ullman (2004), there were no associations between procedural learning and measures of grammar, phonology and decoding. As with previous studies that have failed to find robust associations, it may be that the sub-optimal test-retest reliability of the SRTT results in an underestimation of the true effect size (Rouder et al., 2019).

The most robust association in the present experiment was between attention and procedural learning, particularly in Sessions 1 and 2. This finding is consistent with the results obtained by Sengottuvel and Rao (2013a) and West, Shanks, et al. (2021), and points to attentional resources playing a facilitatory effect in the magnitude and stability of procedural learning in the SRTT as individuals with lower intraindividual variability (as indexed by tau) showed better stability, particularly for later sessions. The decrease in the magnitude of the correlation between attention and procedural learning in Session 3 may be related to the findings obtained by Thomas and colleagues (2004) which demonstrated that a decrease in parietal activity, a brain region which plays a role in visual attention and spatial orienting, occurred once the sequence became more predictable. Thus, tentatively, the smaller correlation in Session 3 may indicate that as the sequence became more predictable with increasing practice this worked to reduce reliance on attentional resources (Thomas et al., 2004). However, it remains for future research to test this hypothesis directly.

General discussion

Procedural learning is thought to be a fundamental component of the memory system, crucial for encoding, storing, and retrieving rule-governed knowledge that underlies motor and cognitive abilities (N. J. Cohen & Squire, 1980). Research into this vital memory system is often reliant on the SRTT; however, questions have been raised about the reliability of this task. Here, we present a systematic examination of the reliability of procedural learning as measured by the SRTT, with the important aim of identifying extrinsic design features (i.e., similarity of sequences learned over sessions, number of sessions, stimulus presentation rate) and participant characteristics (i.e., attention, age, see Appendix F) that could influence reliability. In Experiment 1, manipulation of the levels of similarity between sequences learned at session 1 and 2 revealed a positive relationship between similarity and the procedural learning effect, yet the participant-level stability of the effect was low irrespective of similarity. A follow-up to this found that despite further manipulations of sample (age) and task (interstimulus interval) characteristics (see Appendix E) the test-retest reliability of the SRT remained low. Experiment 2 examined the effect of training over three sessions. However, irrespective of experimental manipulations and participant characteristics, the test-retest reliability of the SRTT remained persistently suboptimal (*r*s < .70). This may be a genuine characteristic of the SRTT, such that procedural learning is inherently unstable across time, but it may also reflect the limitations of the analytical methods used to analyse the psychometric properties of the SRTT. We now turn to discuss these possibilities, as well as the suitability of the SRTT as a marker of procedural learning for use in individual differences research.

Similar to other experimental tasks that tend to show poor test-retest reliability despite robust effects at the group level - a phenomenon termed by Hedge and colleagues as the “reliability paradox” (Hedge et al., 2018), the SRTT having been created for experimental purposes may not be suitable to produce sufficient inter-individual variability in the learning effect (beyond measurement error) that would allow for adequate differentiation between individuals (Hedge et al., 2018; Rouder et al., 2019). The SRTT also shows poor agreement between scores at different time points. Crucially, this indicates that not only the ranking order of participants’ scores, but also the scores themselves, lack stability across sessions. However, despite poor test-retest reliability and agreement, the SRTT often showed adequate split-half reliability, suggesting that individuals’ performance shows within-session consistency.

The use of difference scores has been suggested as a contributing factor to poor reliability as such scores can reduce the signal-to-noise ratio (Hedge et al., 2018). Despite the debates surrounding the limitations of adopting difference scores as indices of the construct of interest (Hedge et al., 2018), differences scores were used in this experiment to estimate split-half reliability and produced good within-session stability. Thus, revealing adequate internal consistency in participants’ performance between halves (odd-numbered and even-numbered trials). Furthermore, the use of random slopes as an index of procedural learning did not significantly improve reliability. Importantly, this suggests that one should not dismiss difference scores as being intrinsically unreliable. This also raises a clear distinction between within-session and across-session stability in the SRTT. Higher within- than across-session stability of the SRTT has been found in previous studies of children and adults (e.g., West et al., 2018; West, Shanks, et al., 2021), with this pattern mirrored in studies using other measures of sequential learning (Hebb task - e.g., (Bogaerts et al., 2018; West et al., 2018); statistical learning e.g., Arnon, 2019 - although this pattern was only found for a visual version of the task and not for linguistic/non-linguistic versions). One simple explanation for why we observe higher within-session than across-session reliability could be due to temporal differences, such that there is a decrease in the magnitude of correlations between trials as the number of intervening trials increases (Wagenmakers et al., 2004). More specifically, whilst short-scale fluctuations are present when computing split-half reliability where even-odd trials are compared, more distant points are compared for the test-retest reliability which, in the present studies, occurred one week apart.

However, this explanation does not account for why we do not see the same disparity between within- and across-session stability for declarative tasks (Buchner & Wippich, 2000; LeBel & Paunonen, 2011; Ward et al., 2013). Kalra et al. (2019) and West et al. (2018) observed that the test-retest reliability of all procedural learning measures was inferior to those of declarative measures. In West et al. (2018), for example, test-retest reliability for the nonverbal immediate serial recall and dot locations tasks test-retest .71 and .57 and split-half reliability was .68 and .76, respectively. This is perhaps in part due to the complex nature of procedural learning itself and the multifaceted nature of the tasks used to measure this poorly defined construct (Bogaerts et al., 2021). Addressing this issue is made even more complex by the interchangeable use of tasks (e.g., Artificial Grammar Learning, Weather Prediction task) that are claimed to tap into procedural memory as a unified ability, despite their computational and modality differences.

Recently, it has been argued that poor test-retest reliability of some tasks (e.g., Stroop task, Flanker test), well-known for producing robust effects at the group level, may be related to the methods adopted to analyse their psychometric properties. (Haines et al., 2020) shows adequate test-retest reliability when using Bayesian hierarchical modelling which more closely captures individuals' performance and accounts for within-subject variability, but suboptimal test-retest reliability when using difference scores. In these models, instead of ignoring uncertainty, as is the case when using point estimates (e.g., mean), which may underestimate test-retest reliability, hierarchical Bayesian models aim to closely represent the data generating process. By using generative modelling, a single model is able to integrate information at person and group level when estimating parameters, accounting for our assumptions and hypotheses from the trial-by-trial response times at the individual level to the overall distribution of individual differences across people (see Haines et al., 2020). Yet here we aimed to explore the impact of experimental manipulations on reliability using statistical methods/measures comparable to previous research (i.e., by estimating the procedural learning effect separately for each session). Future studies may aim to apply the methods applied by Haines et al.

(2020) to the SRTT to determine whether it would better capture the stability of the procedural learning effect across sessions.

Previous studies have noted an association between attention and procedural learning (Arciuli, 2017; Sengottuvel & Rao, 2013a; D. R. Shanks & St. John, 1994; West, Shanks, et al., 2021); however, here, we carried out the first investigation of whether attention influences the stability of procedural learning. Exploratory analyses in Experiments 2 and 3 (see Appendix E) revealed that participants with better attention skills (lower tau) showed more stable procedural learning across sessions than those with worse attention. Thus, these results may lend support to the hypothesis that fluctuations in attention during the task could lead to lower test-retest reliability. One interesting prediction that arises here is that fluctuations of attention may exert lower impact on split-half reliability as this type of stability would be captured by both halves of the task due to the time proximity between even and odd trials. This warrants a systematic assessment of the attention skills during the SRTT using online measures of attention such as pupillometry to better determine its relationship with procedural learning both within and across sessions. A second interesting prediction here is that if attentional skills influence the stability of procedural learning in the SRTT task, then children would be expected to show poorer test-retest reliability than adults as their attentional skills are under development (Levy, 1980). Indeed, this pattern of lower retest reliability has been observed in children by West et al. (2018, 2021), despite somewhat comparable split-half reliability to adults (children: SRT1 *r* = .75; SRT2 r = .49 (500 trials), West et al., 2018; SRT1 *r* = .51; SRT2 *r* = .62 (1000 trials), West, Shanks, et al., 2021; adults: SRT1 *r* = .84; SRT2 *r* = .92 (1000 trials), West, Shanks, et al., 2021).

Fluctuations in procedural learning over time may also be related to changes in performance between measurement points due to individual differences in consolidation and other learning-related strategies adopted at test and retest. This could also account for the higher within- than across-session stability. In line with this, Scharfen et al. (2018), in a recent meta-analysis observed that participants reached a plateau later in working memory tasks compared to other cognitive ability tests. Authors argued that more complex tasks lead to larger retest effects because more test-specific strategies can be developed compared to easier tasks for which strategies do not apply. In the SRTT, this may be accompanied by, or occur due to the development of explicit awareness, as suggested by Stark-Inbar et al. (2017). Thus, the numerically higher test-retest reliability for later sessions observed in Experiment 2 would be expected given that participants' may be reaching a plateau in their learning effect - seen as a reduction in the practice effects for later sessions. Additionally, the strategies adopted for later sessions would potentially be more similar as most participants would already possess some awareness of the presence of an underlying sequence. Future research may aim to explore the trajectory of learning in the SRTT across sessions until no practice effects are observed and its impact on reliability. Alternatively, participants could be asked to perform the SRTT in an initial practice session until each reaches a plateau in performance, only then reliability would be assessed in two separate sessions.

It is important to consider the extent to which poor across-session reliability of procedural learning in the SRTT may impact our ability to adequately test the predictions of models of language and literacy acquisition, namely the declarative/procedural model (Ullman, 2004). This model predicts that the procedural memory system is involved in the development of language and literacy abilities that underlie aspects of rule-based learning. Yet, given that procedural learning tasks may fail to capture participants' true procedural learning abilities, attenuation of the correlation between the constructs of interest is likely to occur. Thus, unsurprisingly, Experiment 2 provided no support for the declarative/procedural model (Ullman, 2004). Whilst there was a weak positive correlation between procedural learning and vocabulary (which would not necessarily be a firm prediction of the declarative/procedural model), there were no other significant correlations with other language/literacy measures that have been claimed by this model to be associated with procedural learning (i.e., grammar, phonological skills). Nevertheless, a positive relationship between procedural learning and attention was observed in Experiment 2 (and also in the experiment presented in Appendix E) irrespective of the reliability issues and possible attenuation of correlations between measures. Thus, it is also possible that this result reflects a genuine lack of support for the declarative/procedural model (Ullman, 2004) and/or poor measurement of procedural learning (Enkavi et al., 2019).

Whilst the various experimental attempts to improve the test-retest reliability of the SRTT were not effective here, there are other potential manipulations to explore. For instance, a critical design element of SRTTs is the number of trials. We carried out a preliminary exploration of this factor with simulation work presented in Appendix A and demonstrated that the ratio of probable to improbable trials can influence test-retest reliability. Whilst researchers have considered the number of trials in the SRTT (e.g., West, Shanks, et al., 2021), the focus tends to be on the overall number of trials, rather than the number of trials per condition as recommended by Rouder et al. (2019). Further experimental work is necessary to determine whether increasing the number of trials in the improbable condition could reduce measurement error, whilst considering the potential consequences for the size of procedural learning effect. Furthermore, considering the findings by West, Shanks, et al. (2021), which suggest that attention during the SRTT, but not procedural learning, predict children's reading, grammatical or arithmetic skills, it is crucial to determine if attention mediates the relationship between procedural learning and language/literacy measures or whether poorer attentional skills represent an additional risk factor for procedural learning deficits in children/adults with dyslexia.

Finally, Bayesian hierarchical models have been shown to be useful in estimating the degree of attenuation in correlations between measures (e.g., attentional control: Rouder & Haaf, 2019; von Bastian et al., 2020), with trial noise and true variability being estimated separately (Rouder & Haaf, 2019). Future research would benefit from exploring the use of these approaches for procedural learning. Regardless of the consistent sub-optimal test-retest reliability of the procedural learning effect, the SRTT has reliably produced robust evidence of learning across populations and settings. Thus, whilst the current set of experiments challenges its suitability for individual differences research (Enkavi et al., 2019), there is little doubt that the SRTT is still a valuable paradigm for group-level experiments.

In conclusion, the probabilistic SRTT used here produced robust procedural learning effects across three experiments, irrespective of samples and testing conditions. Yet, despite some weak evidence of improvement in stability due to the experimental manipulations here presented, it remains suboptimal. Future research should focus on understanding a) the discrepancy between within and across session reliability (e.g., temporal dynamics, consolidation processes) and b) whether there are more sensitive analytical methods that can be used to assess across-session reliability (e.g., Haines et al., 2020). It will also be important to further investigate the potential role of attention in procedural learning, particularly in individuals vulnerable to poor attention (e.g., including those with dyslexia/DLD). Thus, until these questions are answered, it is not possible to use the SRTT to test the boundaries of the procedural/declarative model.

Chapter 3. The Reliability of The Serial Reaction Time Task: Meta-Analysis Of Test–Retest Correlations

Abstract

The Serial Reaction Time task, one of the most widely used tasks to index procedural memory, has been increasingly employed in individual differences research examining the role of procedural memory in language and other cognitive abilities. Yet, despite consistently producing robust procedural learning effects at the group level (i.e., faster responses to sequenced/probable trials versus random/improbable trials), these effects have recently been found to have poor reliability. In this meta-analysis (N datasets = 7), comprising 719 participants (M = 20.81, SD = 7.13), we confirm this "reliability paradox". The overall retest reliability of the robust procedural learning effect elicited by the SRTT was found to be well below acceptable psychometric standards ($r < .40$). However, split-half reliability within a session is better, with an overall estimate of .68. There were no significant effects of sampling (participants' age), methodology (number of trials, sequence type, inclusion of an interstimulus interval, version of the SRTT) and analytical decisions (whether all trials were included when computing the procedural learning scores; using difference scores, ratio scores, or random slopes as an index of learning). Future research will be better equipped to examine the influence of these factors on the reliability of the SRTT once reporting of the psychometric properties of experimental tasks becomes the norm instead of the exception. Thus, despite producing robust effects at the group-level, until we have a better understanding of the factors that improve the reliability of this task and/or have an improved understanding of how best to quantify retest reliability, using the SRTT for individual differences research should be done with caution.

Introduction

The attempt to understand the role of individual differences in cognitive abilities has often led researchers to rely on well-established experimental tasks to capture effects of interest (e.g., Rouder et al., 2019; von Bastian et al., 2020). However, such tasks were typically not designed to be sensitive at the individual level; in fact, it was desirable to reduce individual variability in order to better capture the phenomenon of interest at the group level. Consequently, an increasing number of "mainstream" experimental tasks (e.g., Stroop task: MacLeod, 1991; Flanker task: B. A. Eriksen & Eriksen, 1974; Navon task: Navon, 1977) are being reported to have poor reliability (Hedge et al., 2018; von Bastian et al., 2020). This phenomenon is now referred to as the "reliability paradox" (Hedge et al., 2018): specifically, tasks that produce robust group-level effects but fail to capture reliable individual differences, such that the rank order of participants' performance on the same measure lacks stability between test and retest (i.e., test-retest reliability) or within a session (i.e., split-half reliability) (Berchtold, 2016; Hedge et al., 2018; Nunnally & Bernstein, 1994). Low reliability of these experimental paradigms has often been attributed to the use of difference scores to isolate the effect of interest, whereby subtracting the experimental and control conditions reduces the variance between subjects (Enkavi et al., 2019). Measurement error may also contribute to the poor stability of participant scores at test and retest where participants' scores will change non-systematically between sessions. Thus, if these tasks are mainly capturing measurement error, instead of stable effects, individual difference studies based on these measures will fail to reflect real variation in individuals' performance; and thus are also likely to produce attenuated correlations with other variables (Enkavi et al., 2019; Hedge et al., 2018; Rouder et al., 2019). However, the opposite effect, whereby effect sizes are overestimated, may also occur by chance in small samples due to measurement error as demonstrated by Loken and Gelman (2017). Relatedly, and potentially partly a consequence of poor psychometric properties, experimental tasks that are thought to measure the same construct (e.g., Flanker and Stroop tasks) often fail to correlate with each other (Rouder et al., 2019; Rouder & Haaf, 2020). Thus, the "reliability paradox" is an important problem to tackle if we are to advance our understanding of key cognitive constructs, particularly in the context of individual differences.

The SRTT is one such tasks, known for producing robust procedural learning effects across settings, populations and task manipulations, in the face of poor psychometric properties (Kalra et al., 2019; Siegelman & Frost, 2015; Stark-Inbar et al., 2017; West, 2018; West, Shanks, et al., 2021). In the SRTT (Nissen & Bullemer, 1987), a stimulus appears on screen in one of four rectangles and participants are asked to respond as soon as possible to the position of the stimulus by pressing the corresponding key on the keyboard. Unbeknownst to the participants, the position of the stimulus follows a pattern of either deterministic or probabilistic nature. In deterministic SRTT, blocks of patterned and random trials are alternated, with a higher number of patterned than random blocks. In probabilistic sequences, on the other hand, patterned trials are interspersed with random trials, with varying degrees of signal to noise ratio across tasks (e.g., alternating SRTTs have the same number of sequenced and random trials). Irrespective of which task is used, as participants learn the sequence, response times are expected to become faster for sequenced trials when compared to random ones, as participants are thought to be able to extract regularities from sensory input and anticipate the position of the next stimulus.

Procedural memory is thought to underlie the acquisition and use of complex, sequence-based motor, perceptual and cognitive skills, including - according to an influential account - language acquisition (Ullman, 2016; Ullman et al., 2020). This has led to an increased interest in the role that procedural memory plays in driving individual variability in language development, whereby those with better procedural memory skills are expected to show better language proficiency and deficits in procedural memory can result in deficits in language and literacy development. However, despite the increasing interest in individual differences in procedural memory, researchers have used SRTTs without much consideration for their psychometric properties (Siegelman, Bogaerts, & Frost, 2017). Only more recently has the reliability of these tasks been examined and found to be suboptimal (*r*s <.70; e.g., Siegelman & Frost, 2015; Stark-Inbar et al., 2017; West et al., 2018; West, Shanks, et al., 2021), a finding that is mirrored in other measures of procedural learning (e.g., the Hebb task: West, 2018; contextual cueing: West, 2018; artificial grammar learning: Erickson et al., 2016, word segmentation tasks: Arnon, 2020; Erickson et al., 2016; Siegelman & Frost, 2015). Indeed, Kalra et al. (2019) found that a number of implicit learning tasks (i.e., SRTT, artificial grammar learning, probabilistic classification task), all of which are thought to index procedural learning showed below optimal test-retest reliability (*r*s <.45), and that inter-correlations among these measures were low, ranging from -.18 to .32. Similarly, in children, West et al. (2018) observed poor test-retest reliability (*r* <.30) for the verbal and nonverbal versions of the SRTT and Hebb tasks, and correlations between measures ranging from -.18 to .24. Whilst the small inter-correlations between procedural learning measures may be a consequence of the poor reliability of the individual tasks, it is also possible that these reflect a componential, rather than unitary, underlying construct (Bogaerts et al., 2021).

Unsurprisingly then, the large body of research over the last decade that has used the SRTT to explore relationships between procedural memory and language and literacy abilities has produced inconsistent findings. Specifically, whilst some studies have shown that individuals with greater procedural learning ability also have better language and literacy skills (e.g., Conti-Ramsden et al., 2015; Hamrick et al., 2018; Kidd, 2012), this finding is not ubiquitous (e.g., Henderson & Warmington, 2017; Siegelman & Frost, 2015; West et al., 2018). However, it is not clear whether differences in methods, reliability, or a combination of both, account for the inconsistent association between procedural learning and language. Namely, the above-mentioned studies vary on sample characteristics (e.g., age), design (e.g., interval between test and retest, use of the same or different SRTT sequences at different testing points), and the analytic method by which procedural learning is calculated (e.g., a simple difference between conditions, ratio scores, random slopes). However, the impact of these methodological variations on reliability is unknown. Thus, a systematic examination of the influence of such factors on reliability of the procedural learning effect is required, to better understand how to optimise the psychometric properties of the SRTT.

Whilst many factors likely influence the test-retest reliability of the SRTT, here we focus on (i) sample characteristics (ii) task design, including the type of SRTT and the interval between test and retest, (iii) use of same or alternate forms of the SRTT and analytical decisions (e.g., measure of procedural memory, whether to include all trials). Regarding sample characteristics, there is some evidence that reliability of the SRTT is moderated by age, with lower test-retest reliability for children than adults (West et al., 2018; West, Shanks, et al., 2021). One possibility is that lower retest reliability of procedural learning may be seen for children owing to age-related differences in the attentional and motivational demands of the SRTT (West, Shanks, et al., 2021). There also appears to be a tendency for higher test-retest reliability for probabilistic (including both probabilistic and alternating SRTTs) than deterministic tasks as observed by Stark-Inbar et al. (2017), where alternating sequences showed a test-retest reliability of .46 whilst for the deterministic SRTT the reliability coefficient was nearly zero (.07). Furthermore, the study reporting the highest reliability thus far (West, Shanks, et al., 2021) used a probabilistic SRTT. The superior psychometric properties of probabilistic SRTT tend to be attributed to the lower likelihood of eliciting explicit awareness (Stark-Inbar et al., 2017), since in deterministic SRTT the continuous repetition of the same elements of the sequence may contribute to its higher salience compared to sequences learnt in a noisier context (Vandenberghe et al., 2006). In Stark-Inbar et al. (2017), despite longer practice sessions in the alternating SRTT, participants who learnt the deterministic SRTT still demonstrated more evidence of explicit awareness. Furthermore, it is possible that procedural learning in probabilistic SRTTs is less confounded with fatigue, given that probabilistic sequences allow for the tracking of procedural learning throughout the task, instead of only in the last blocks as is common practice in deterministic sequences (Pan & Rickard, 2015). However, this is unlikely to be responsible for the differences in reliability between these variants of the SRTT, as Oliveira et al. (submitted) observed superior reliability when including only the last 3 blocks, instead of all trials.

Whilst the effect of the interval length between test and retest has not yet been examined as a factor that influences reliability of the SRTT, evidence suggests that in cognitive and neuropsychological tests shorter intervals tend to lead to higher retest reliability coefficients than longer intervals (Duff, 2012). This may be due to the possibility for true change in cognitive abilities to occur with longer intervals (Allen & Yen, 1979; Duff, 2012). However, shorter intervals are also associated with greater opportunity for practice effects than longer intervals (Calamia et al., 2012, 2013), where improvements across sessions may result from familiarity with the testing procedures, memory traces of the test items and the development of strategies (McCaffrey et al., 2000). The impact of practice effects on test-retest reliability may not be trivial, unless all participants show the same magnitude of improvement at retest, which is unlikely (e.g., R. M. Brown et al., 2009; Hodel et al., 2014), such effects are likely to change the rank order of participants from test to retest. To reduce practice effects, researchers often administer alternate forms of a test; in the case of the SRTT, that can be achieved by using different sequences at test and retest, whilst the remaining task characteristics are kept consistent across sessions. However, alternate forms are often not sufficient to prevent practice effects from occurring (Beglinger et al., 2005). This issue is relevant as practice effects in the SRTT were evident in Siegelman and Frost (2015) where 64 out of 75 participants showed better performance at retest when using the same sequences at test and retest. However, its impact on reliability has rarely been experimentally tested. In (Oliveira et al., submitted), a positive effect of similarity between sequences at test and retest on the magnitude of procedural learning was observed, but not on test-retest reliability. Similar results were obtained by West and colleagues in two separate experiments, where reliability was assessed using the same (*r* = .21, West et al., 2018) or alternate forms of the SRTT (*r* = .26, West, Shanks, et al., 2021), thus suggesting that the adoption of alternate forms did not lead to significant changes in the coefficients. This suggests that even when steps are taken to avoid practice effects, performance may be susceptible to change across sessions; whether such changes are greater as a function of interval length remains an open question.

Irrespective of task characteristics, a more recent debate has focussed on the methods used to capture reliability, whereby the “reliability paradox” may not reflect lack of stability of the underlying construct, but instead indicate that the use of point estimates to analyse reliability may fail to adequately model the data-generating process (Haines et al., 2020). More concretely, instead of relying on point estimates, the methods for assessing reliability should integrate information at the individual and group level, whilst accounting for trial-by-trial variability (Haines et al., 2020; Rouder et al., 2019; von Bastian et al., 2020). Unfortunately, most evidence on the procedural learning effect and its reliability has used difference scores, with only a few studies which have controlled for overall speed by using ratio scores (e.g., West et al., 2018; West, Shanks, et al., 2021) or adopted random slopes (Lammertink et al., 2020; Oliveira et al., submitted) thus integrating information at the individual and group level. The use of random slopes may be able to capture better reliability, if the construct is indeed stable, as according to Stein's paradox (Efron & Morris, 1977) the best predictor of participant's true ability is not their own performance across sessions, but instead their performance adjusted to be more in line with the group.

The aims of the present meta-analysis were threefold. First, we aimed to assess the frequency with which the reliability of the SRTT is reported. Second, we endeavoured to establish the test-retest reliability of the procedural learning effect as measured by the SRTT. Whilst our preregistered objective was to examine test-retest reliability, the search strategy also produced studies examining split-half reliability, and thus this was also examined. Third, we aimed to examine which, if any, methodological factors influence the psychometric properties of this task (i.e., sample characteristics, task design including the interval between test and retest and use of same or alternate forms of the SRTT, analytical method for calculating procedural learning). Regarding these methodological factors, we predicted that a) children would show poorer reliability than adults; b) longer intervals between test and retest would result in poorer reliability, c) that poorer reliability would be expected for difference scores than other measures of procedural learning (ratio scores and random slopes).

Study objectives

We describe a protocol for a meta-analysis of studies investigating the test-retest reliability of the SRTT. This investigation aims to assess the across-session stability of the SRTT, whilst considering possible moderating variables (e.g., age, length of interval between sessions). Suboptimal test-retest reliability was predicted ($r < .70$) (H1), especially for children (H2) (West et al., 2018; West, Shanks, et al., 2021) and for longer intervals between test and retest (H3) as observed in other neuropsychological tasks (e.g., Calamia et al., 2013). On the other hand, measures that take into account individuals' speed (e.g., ratio, random slopes) are expected to have higher test-retest reliability than those which do not (e.g., difference scores) (H4) (Oliveira et al., submitted; West, Shanks, et al., 2021). We also examined whether split-half reliability was closer to acceptable standards, as in previous studies (West et al., 2018, 2019; West, Shanks, et al., 2021).

Exploratory analyses were conducted to determine further methodological characteristics that may influence test-retest reliability, namely the number of trials, use of the same or different sequences at test and retest, or SRTT variant (e.g., deterministic vs probabilistic sequence).

Methods

The protocol containing hypotheses, methods and analysis plan for this review was prospectively registered on the Open Science Framework (https://osf.io/uyqvt). All materials for this meta-analysis are available (https://osf.io/a65hn/), including the dataset and scripts necessary to replicate all reported analyses and plotting.

Search strategy

To ensure a comprehensive search strategy, a university librarian was consulted when developing the terms for each database. Literature was compiled by performing a full-text search in July 2021 on Medline, PsycINFO and Embase, as well as on BASE - Bielefeld Academic Search Engine for grey literature. Citation searching was also conducted to ensure that all relevant papers were identified.

Specifically, the following search string was used for Medline, PsycINFO and Embase: (Procedural learning OR Procedural memory OR Sequence learning OR Implicit learning OR Statistical learning OR Procedural knowledge.sh OR Serial Reaction Time task (.tw for PsycINFO)) AND (Reliab* OR Consistency OR Stab* OR Individual differences OR Valid* OR Psychometr* OR Measurement). For grey literature on BASE the following search string was used instead: ("procedural learning" "procedural memory" "sequence learning" "implicit learning" "statistical learning" "procedural knowledge" "serial reaction time task") AND (reliab* consisten* stab* "individual differences" valid* psychometr* measurement*)

Selection of studies

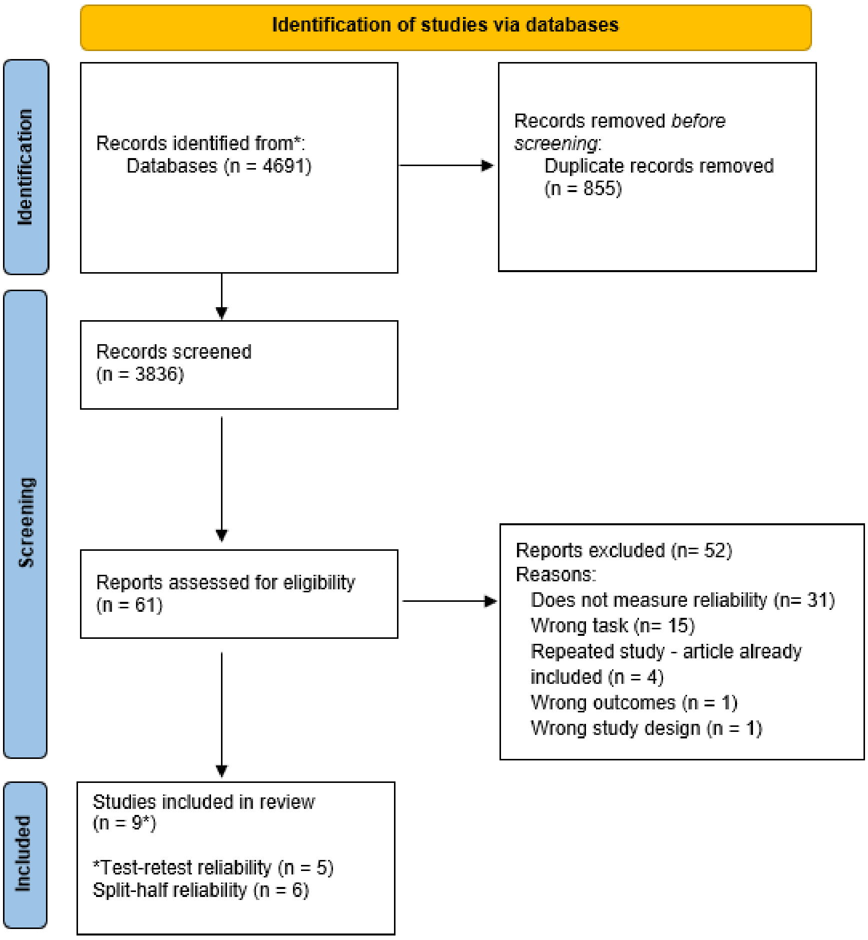
One reviewer independently screened all articles and identified those relevant for the meta-analysis. This screening was done at the title and abstract level. At the full-article level the list of papers was screened by the first author to determine whether they fitted the inclusion criteria. To assess full-text eligibility the following inclusion criteria: i) Used a strictly visual deterministic, probabilistic or alternating SRTT with procedural learning computed as the difference between sequenced/probable and random/improbable trials; ii) Reported Pearson's correlation (or equivalent) coefficients between procedural learning at two or more time points; iii) If the same results were published in multiple articles, these were only reported once in the meta-analysis; iv) Language of publication: English. Exclusion criteria were: i) Studies that used adaptations that considerably deviate from the task proposed by Nissen and Bullemer (1987); ii) Dual task paradigms; iii) Studies with active interventions or studying populations whose performance is expected to change over time (e.g., stroke patients).

Discrepancies in article inclusion were resolved by discussion between the three reviewers. Once the list of full articles was agreed upon, the first reviewer coded the data for the following items: a) Author/s; b) Publication year; c) Number of participants; d) Age of participants; e) Test-retest reliability; f) Split-half coefficient; g) Variant/version of SRTT (deterministic or probabilistic); h) Sequence complexity, i) Interval between sessions; j) Design: same or alternate version at 2nd test-point, k) Total number of trials completed, l) Number of trials included when computing the reliability measure: all trials or only last blocks, m) measure of procedural learning (difference scores or ratio/random slopes - ratio and random slopes were collapsed due to the small number of studies examining measures other than difference scores).

The PRISMA flow diagram (Page et al., 2021) in Figure 3.1 will be used to track the number of records identified, included, and excluded; as well as the reasons for exclusions.

Figure 3.1

PRISMA flowchart showing selection of studies for meta-analysis on the reliability of the SRTT


Identification of studies via databases
Identification
Screening
Included
Records identified from*:
Databases (n = 4691)
Records removed *before screening*:
Duplicate records removed (n = 855)
Records screened (n = 3836)
Reports assessed for eligibility (n = 61)
Reports excluded (n= 52)
Reasons:
Does not measure reliability (n= 31)
Wrong task (n= 15)
Repeated study - article already included (n = 4)
Wrong outcomes (n = 1)
Wrong study design (n = 1)
Studies included in review (n = 9*)
*Test-retest reliability (n = 5)
Split-half reliability (n = 6)

Statistical analyses

The analyses were carried out using R (version 4.1.1) (R Core Team, 2020). All continuous moderators were mean centred. Centering moderators does not change the random coefficients. All correlation coefficients were converted from Pearson's r to Fisher's z scale as Pearson's r is not normally distributed (Efron & Morris, 1977).

The *metafor* package (Viechtbauer, 2010) was used for model fitting. Random effects working models were fitted following the guidelines from J. E. Pustejovsky and Tipton (2022) for model specification to reflect the levels of dependency in the dataset. Since research groups contributed multiple correlation coefficients from the same samples, to deal with the lack of independence across effect sizes and avoid reducing power by calculating the average effect sizes for these studies, multilevel models were estimated using the function *rma.mv()* from the *metafor* package (Viechtbauer, 2010). As recommended by J. E. Pustejovsky and Tipton (2022), robust variance estimation standard errors, hypothesis tests, and confidence intervals for the meta-regression coefficient estimates were computed using the functions *coef_test()* or *conf_int()* from the *clubSandwich* (J. Pustejovsky, 2021) to guard against model misspecification (J. E. Pustejovsky & Tipton, 2022). Since these multilevel models represent working models which may fail to fully represent the dependence structure, robust variance estimation methods were used as these do not require exact knowledge of the dependence, thus even if the working model is misspecified, the estimates will be unbiased (J. E. Pustejovsky & Tipton, 2022). The correlation of the sampling errors within clusters (rho) was set at .80, and sensitivity analyses were conducted to determine whether this decision impacted the overall estimate.

An intercept-only model was fitted to estimate the overall test-retest reliability (i.e., the average correlation coefficient between test and retest) of the SRTT. Following the intercept-only model, separate meta-regression models were performed for each mediator variable (e.g., age, total number of trials) to determine whether any of these factors influence the test-retest reliability of the SRTT. After performing the meta-analytic calculations, Fisher's *z* effect sizes were converted back to Pearson's *r* for reporting the average correlation and 95% CI for each model. We first started by fitting a reduced model which included only one effect size per sample. When multiple reliability estimates were available for the same sample, difference scores were adopted as a default measure unless such measure was not available, as these better represent current practices in the field of procedural learning. Finally, a second model was fitted (full model) which includes all effect sizes, thus allowing for direct comparisons between analytical decisions across studies.

Bias analysis

Study heterogeneity was analysed using the Q-test for heterogeneity (Cochran, 1954) which reflects the ratio of observed variation to within-study variance. Cook's distances and studentized deleted residuals (or externally studentized residuals) were used to identify potential influential and outlier cases, respectively, as these may distort the conclusions of the meta-analysis (Viechtbauer & Cheung, 2010). Cook's distances provide information about the leverage of each effect size by excluding each study in turn and determining its impact on the overall estimate. Studentized deleted residuals, on the other hand, were used to identify potential outlier points, i.e., absolute studentized deleted residuals larger than 1.96 (Viechtbauer & Cheung, 2010).

To detect evidence of bias, funnel plots, contour-enhanced funnel plots, as well as Egger's regression test (Sterne & Egger, 2005), were used to check for funnel plot asymmetry. Contour-enhanced funnel plots are an extension of funnel plots, as the areas of statistical significance have been overlaid on the funnel plot. By adding these contours, it is possible to determine whether there are potential missing studies in areas of no significance, thus suggesting that the asymmetry may be due to publication bias (Peters et al., 2008). Following the recommendations by Sterne et al. (2011), these were interpreted with caution given the small number of studies (< 10 studies) included in the present meta-analysis.

Finally, given that our group (Oliveira et al., submitted, in prep) has comprehensively examined the reliability of the SRTT across settings (in lab and online), with various measures (e.g., difference scores and random slopes) and using different levels of similarity between sequences, follow-up models were fitted to the data after excluding the effect sizes from our group.

Results

In total, the meta-analysis includes 7 independent studies (Kalra et al., 2019; Oliveira et al., submitted, in prep; Siegelman & Frost, 2015; Stark-Inbar et al., 2017; West et al., 2018, 2021) (citations marked with a dagger) summarising 36 effect sizes and data from 719 participants (M = 20.81, SD = 7.13), comprising 199 children and 520 adults. Thus, it was observed that, despite the frequent adoption of the SRTT to analyse procedural memory (as of September 2021 a Google Scholar search of "Serial Reaction Time task" yields 13,300 results), only a small fraction of studies reported a test-retest reliability estimate. All studies were published between 2015 and 2021.

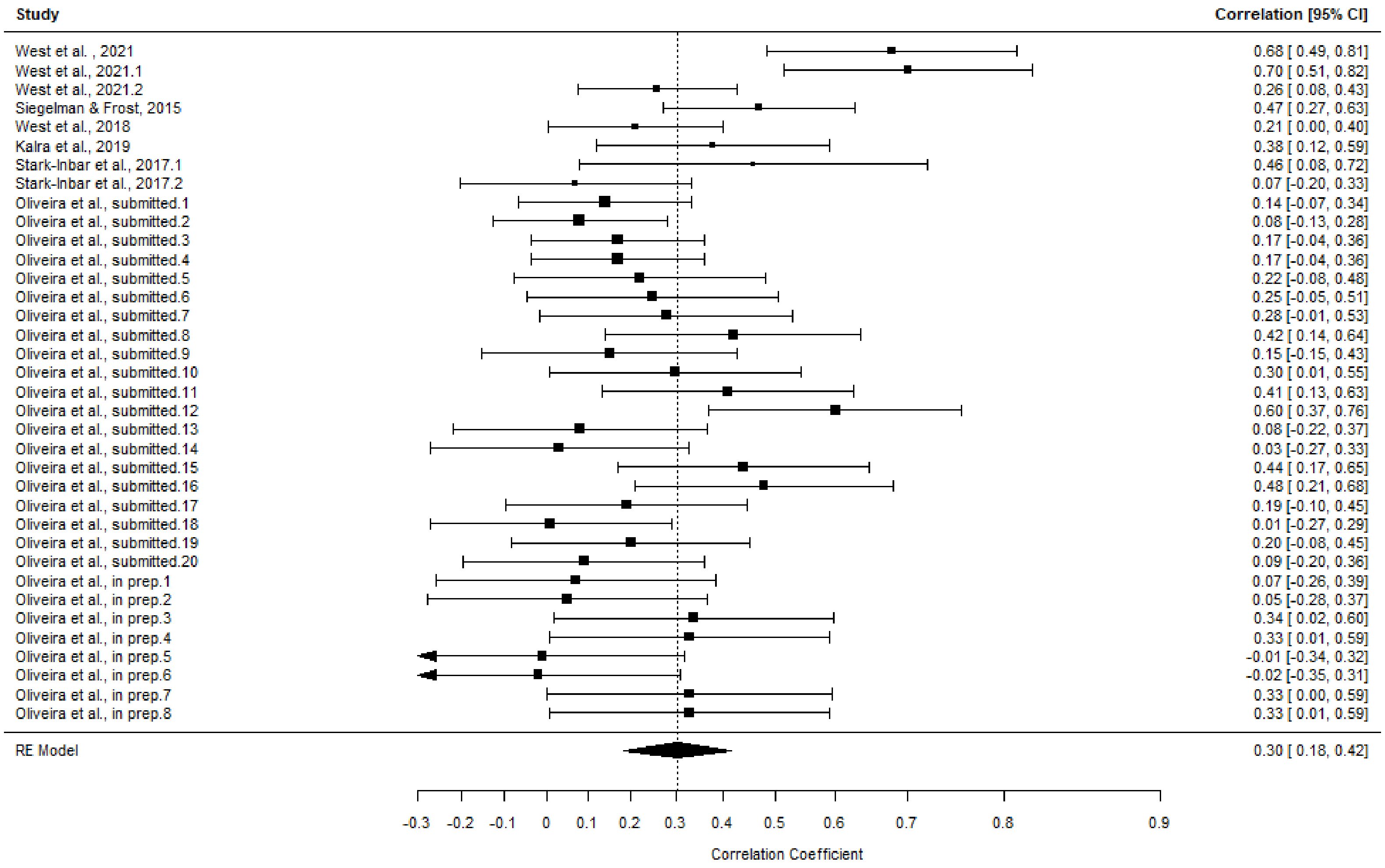
Test-retest reliability

A multilevel mixed effects model was fitted to the test-retest reliability data. In this first model, only a single reliability score was included per experiment, with difference scores being chosen when more measures were available. Only the study by Kalra et al. (2019) did not report difference scores, reporting instead ratio scores. This (reduced, i.e., only includes one effect size per experiment) model revealed a significant and suboptimal pooled test-retest reliability across studies and measures (Fisher's z = .29, 95% CI [.16, .43], SE = .07, z = 4.31, p < .001), with an equivalent test-retest reliability of r = .28, 95% CI = [.16, .40]. Follow-up RVE estimates were computed to guard against model misspecification and correct for the small sample size; these showed consistent results with the multilevel model (SE = .07, t(5.04) = 4.23, p = .008). A forest plot showing the observed outcomes and the estimate based on the random-effects model is shown in Figure 3.2.

Because the studies by Oliveira et al. (submitted, in prep) contributed a relatively large amount of data (5 effects and a total of N = 306 participants) to the meta-analysis, a second multilevel model was fitted after removal of the effect sizes from this lab to determine their impact on overall reliability. Suboptimal test-retest reliability was still observed, with only a small increase in the estimated reliability (Fisher's z = .38, 95% CI [.21, .56], SE = .09, z = 4.21, p < .001), which corresponds to r = .37, 95% CI = [.20, .51].

Figure 3.2

Forest plot showing the observed outcomes and the estimate of the multilevel model for test-retest reliability



According to the results of the Q-test, there is considerable heterogeneity in the estimation of the test-retest reliability of the SRTT (Q(df = 11) = 26.51, p = .005). Thus, this variability across studies was explored through meta-regressions. To achieve this, a second (full) model including all effect sizes was fitted to determine the impact of some frequent methodological decisions. This model revealed a similar average test-retest reliability (Fisher's z = .31, 95% CI [.19, .44], SE = .07, z = 4.80, p < .001), corresponding to a Pearson's correlation of .30, 95% CI = [.18, .42]. As before, a follow-up model was fitted which excluded the findings from Oliveira et al. (submitted); this again yielded a small improvement in reliability which was however still far from optimal (Fisher's z = .39, 95% CI [.20, .58], SE = .10, z = 4.08, p < .001; r = .37, 95% CI [.20, .52]).

Given the clustered nature of the models presented, influential and outlier effect sizes were identified at the various levels. At the study and experimental level, the study conducted by West, Shanks, et al. (2021) in adults was identified as an influential point which upwardly biased the overall estimate. In the opposite direction, Oliveira et al. (submitted) was identified as an influential point at the study level for both models, whilst the experiment conducted by West, Shanks, et al. (2021) in children was influential only at the experimental level. The effect sizes from West, Shanks, et al. (2021) were identified as outliers for both models (reduced: only adult data; full model: both child and adult effect sizes).

Moderator analyses revealed no evidence of a significant moderating effect of age, total number of trials, or test-retest interval on the magnitude of the test-retest reliability coefficient (*ps* > .05). For categorical variables (i.e., measure, type of SRTT, ISI, trials included when computing the reliability measure, SRTT version), the test-retest reliability was significantly different from zero for at least one level of the moderator variable. However, given the small sample size, only RVE estimates will be interpreted. These revealed an average test-retest reliability that was significantly different from zero across measures, irrespective of which measure was used. A numerical but non-significant advantage was observed for ratio and random slopes compared to difference scores (F(1,1.46) = 2.57, p = .293). A slight numerical, but not significant, advantage was also observed for SRTTs with an interstimulus interval versus those without (F(1, 2.85) = .266, p = .644), as well as for studies which computed procedural learning using the last blocks of the experiment (F(1,1.38) = .86, p = .487). Probabilistic SRTTs also yield slightly better test-retest reliability than deterministic tasks, but again this difference did not reach significance (F(1, 1.52) = .247, p = .682).

Table 3.1

Results of all separate meta-regressions with moderator variables for test-retest reliability

Moderator (bolded) and levels	*s*	*exp*	*ES*	*Test of Moderators*		*Meta regression*							*RVE*			
				QM	*p*	*Fisher's z*	*r*	*SE*	*z*	*p*	*95% CI*		*SE*	*t*	*df*	*p*
Age	7	12	36	0.59	.443	0.04	.04	0.05	0.77	.443	-0.06	0.13	0.05	0.72	2.46	.536
Measure	7	12	36	27.92	<.001	—	—	—	—	—	—	—	—	—	—	—
Difference scores	—	—	—	—	—	0.28	.27	0.07	3.92	<.001	0.14	0.42	0.08	3.33	4.73	.023
Ratio/Random slopes	—	—	—	—	—	0.40	.38	0.08	5.28	<.001	0.25	0.55	0.07	5.74	3.67	.006
# of Trials	7	12	36	1.72	.190	0.06	.06	0.05	1.31	.190	-0.03	0.15	0.08	0.80	1.25	.548
Type of SRTT	7	11	35	15.01	<.001	—	—	—	—	—	—	—	—	—	—	—
Deterministic	—	—	—	—	—	0.23	.23	0.18	1.29	.197	-0.12	0.59	0.17	1.42	1.00	.390
Probabilistic	—	—	—	—	—	0.33	.32	0.09	3.65	<.001	0.15	0.50	0.09	3.73	3.43	.027
ISI[1]	6	10	34	18.85	<.001	—	—	—	—	—	—	—	—	—	—	—
0	—	—	—	—	—	0.30	.29	0.10	2.93	.003	0.10	0.50	0.06	5.28	1.59	.054
250	—	—	—	—	—	0.36	.35	0.10	3.39	<.001	0.15	0.56	0.12	2.88	2.71	.072
Trials included	7	12	36	22.45	<.001	—	—	—	—	—	—	—	—	—	—	—
all trials	—	—	—	—	—	0.30	.29	0.07	4.09	<.001	0.15	0.44	0.07	4.26	3.82	.015
last blocks	—	—	—	—	—	0.35	.34	0.08	4.55	<.001	0.20	0.50	0.08	4.25	3.33	.019
Version	7	12	36	21.09	<.001	—	—	—	—	—	—	—	—	—	—	—
Alternate	—	—	—	—	—	0.31	.30	0.08	4.03	<.001	0.16	0.46	0.08	3.94	2.97	.030
Same	—	—	—	—	—	0.36	.35	0.16	2.21	.027	0.04	0.68	0.15	2.42	1.00	.250
Interval[2]	3	6	29	0.39	.531	0.03	.03	0.04	0.63	.531	-0.06	0.11	0.03	0.57	1.17	.659

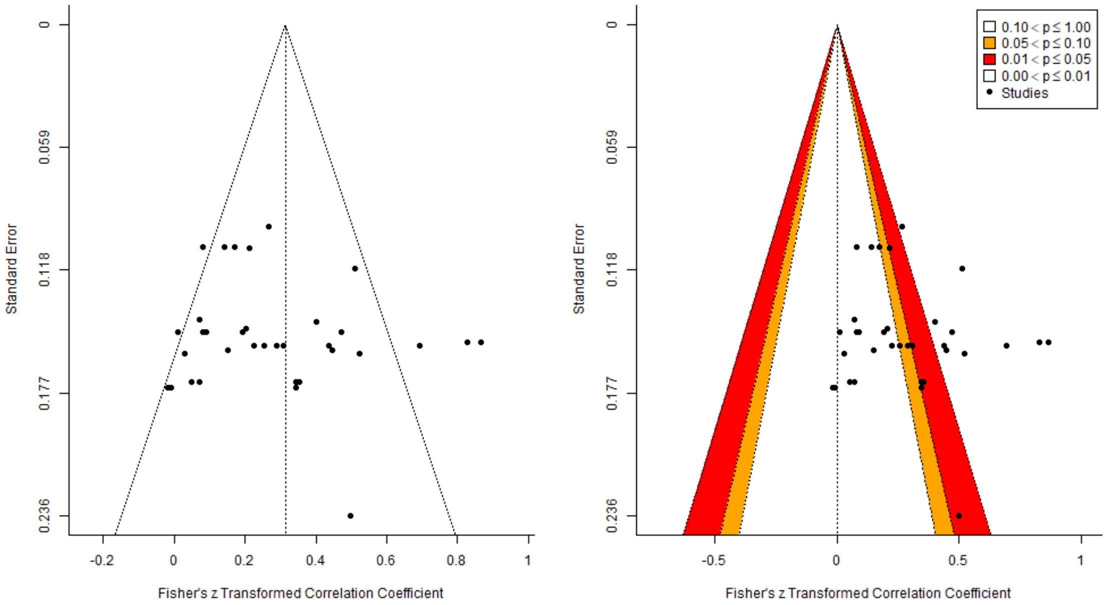
Note. s = number of studies; exp = number of experiments; ES = number of effect size estimates; z' = Fisher's z values; r = Pearson's R correlation; standard errors (SE) and z values for individual levels of a moderator; p values correspond to z or t - values; 95 % CI corresponds to the Fisher's z; [1] As only one effect size was available for ISIs of 100 and 120, both from Stark-Inbar et al. (2017), these were not included in the analysis; [1] Only a small number of experiments (n = 4; Kalra et al., 2019; Oliveira et al., submitted, prep), reported the mean interval between sessions.

Publication bias

Visual inspection of the funnel and contour plots shown in Figure 3.3 for all effect sizes does not reveal evidence of plot asymmetry or overrepresentation of studies in the significance contours, which is consistent with the non-significant Egger's test with standard error as a predictor (*b* = 3.23, *p* = .075).

Figure 3.3

Funnel plot showing effect sizes plotted against standard error for test-retest reliability. A: funnel plot (left panel) and B: contour-enhanced funnel plot (right panel).



Split-half reliability

Whilst the search strategy was aimed towards identifying studies which examined the test-retest reliability of the SRT, the search results also yielded eight studies reporting split-half reliability comprising 1200 participants (Mage = 19.22, SD = 9.21); an exploratory analysis examining this was carried out (J. Brown, 2010; Feldman et al., 1995; Lammertink, Boersma, Wijnen, et al., 2020; Oliveira et al., submitted, in prep; van Witteloostuijn et al., 2021; West et al., 2018; West, Shanks, et al., 2021)**.**

As for test-retest reliability, when studies computed split-half reliability using multiple measures, only difference scores were selected for this first model as this measure is the most commonly used in the field. This was only the case for the studies by Oliveira et al. (submitted). The overall split-half reliability of the SRTT was higher than its test-retest reliability, with a pooled effect size of r = .65, 95% CI [.53, .74] (Fisher's z = .77, 95% CI [.61, .93], SE = .08, z = 9.42, p < .001); this was unaffected when using RVE (SE = .08, t(4.33) = 9.40, p < .001). Sensitivity analyses revealed that the estimates were robust to distinct values of rho with the estimates ranging from .772 to .773. As before, when removing the effect sizes from Oliveira et al. (submitted) there was a slight improvement in the split-half estimate r = .69, 95% CI [.49, .81] (Fisher's z = .84, 95% CI [.54, .1.14], SE = .15, z = 5.51, p < .001).

Following the high degree of heterogeneity in the estimates of split-half reliability in the reduced model (only one effect size per experiment) (Q(21) = 163.86, p < .001), a full model with all effect sizes was performed to explore whether any of the sampling and/or methodological factors impact the split-half reliability of the SRTT.

When all effect sizes were included there was a slight increase in the split-half reliability, r = .68, 95% CI [.57, .76] (Fisher's z = .83, 95% CI [.65, 1.00], SE = .09, z = 9.38, p < .001), which was unchanged when using RVE (SE = .07, t(3.79) = 12.10, p < .001). Removal of the effect sizes by Oliveira et al. (submitted, in prep) did not change the findings, r = .69, 95% CI [.49, .81] (Fisher's z = .84, 95% CI [.54, 1.14], SE = .15, z = 5.51, p < .001; RVE: SE = .13, t(4.24) = 6.53, p = .002). The study by West, Shanks, et al. (2021) with adults was identified as an influential case at all levels upwardly biasing the estimate for both the full and reduced models. Finally, the effect size from West, Shanks, et al. (2021) with adults and the highest effect size from Oliveira et al. (submitted) were identified as outliers. A forest plot showing the observed outcomes and the estimate based on the multilevel model is shown in Figure 3.4.

Meta-regressions revealed results consistent with the findings for test-retest reliability (Table 3.2). There was no evidence that age and the total number of trials had a moderating effect on split-half reliability (ps > .05). For categorical variables, there was no statistical difference between any of the contrasts, although random slopes showed a numerically slightly higher split-half reliability compared to difference scores (F(1, 1.38) = 0.79, p = .503) and there appeared to be a slight numerical advantage of having an interstimulus interval (250ms), (F(1, 3.20) = 0.203, p = .681). No difference was observed in the split-half reliability of probabilistic and deterministic SRTTs, F(1, 2.87) = 0.06, p = .830.

Figure 3.4

Forest plot showing the observed outcomes and the estimate of the multilevel model for split-half reliability

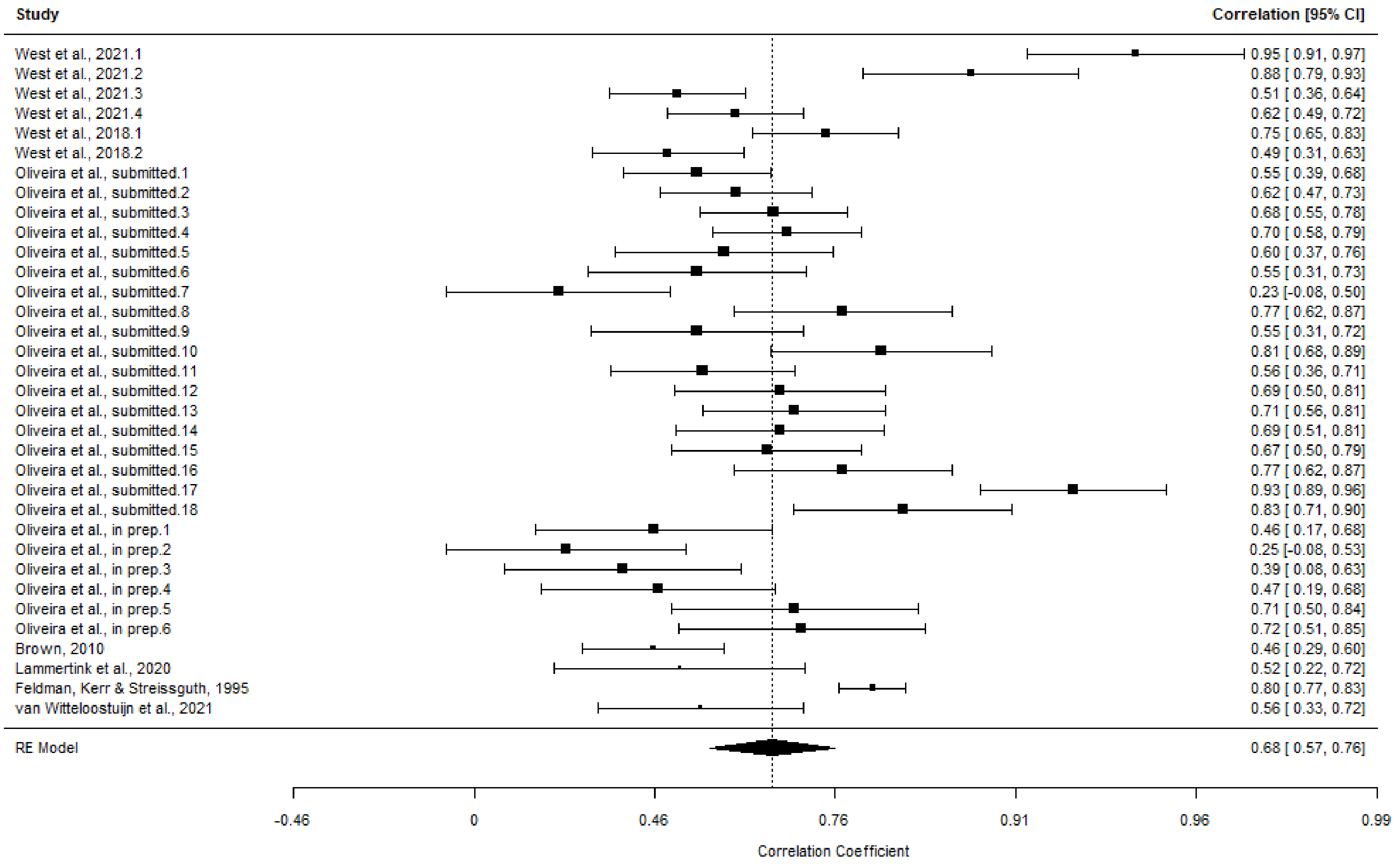


Table 3.2

Results of all separate meta-regressions with moderator variables for split-half reliability

Moderator (bolded) and levels	***s***	***exp***	***ES***	***Test of Moderators***		***Meta regression***							***RVE***			
				QM	*p*	*Fisher's z*	*r*	*SE*	*z*	*p*	*95% CI*		*SE*	*t*	*df*	*p*
Age	8	12	34	0.66	.416	0.06	.06	0.08	0.81	.416	-0.09	0.22	0.07	0.97	4.82	.380
Measure	8	12	34	77.49	<.001	__	__	__	__	__	__	__	__	__	__	__
Difference scores	__	__	__	__	__	0.76		0.10	7.49	<.001	0.56	.96	0.10	7.56	3.78	.002
Ratio/Random slopes	__	__	__	__	__	0.94		0.11	8.29	<.001	0.71	1.16	0.15	6.21	2.66	.012
Total # of trials	8	12	34	2.54	.111	0.11		0.07	1.59	.111	-0.02	0.24	0.09	1.24	3.90	.285
Type of SRTT	8	12	34	80.76	<.001	__		__	__	__	__	__	__	__	__	__
Deterministic	__	__	__	__	__	0.79		0.21	3.69	<.001	0.37	1.21	0.18	4.37	1.98	.049
Probabilistic	__	__	__	__	__	0.84		0.10	8.19	<.001	0.63	1.03	0.08	10.82	2.49	.004
ISI[1]	7	11	33	69.65	<.001	__		__	__	__	__	__	__	__	__	__
0 ms	__	__	__	__	__	0.76		0.14	5.55	<.001	0.49	1.03	0.14	5.42	1.48	.060
250 ms	__	__	__	__	__	0.85		0.14	6.23	<.001	0.58	1.12	0.12	6.85	3.33	.005

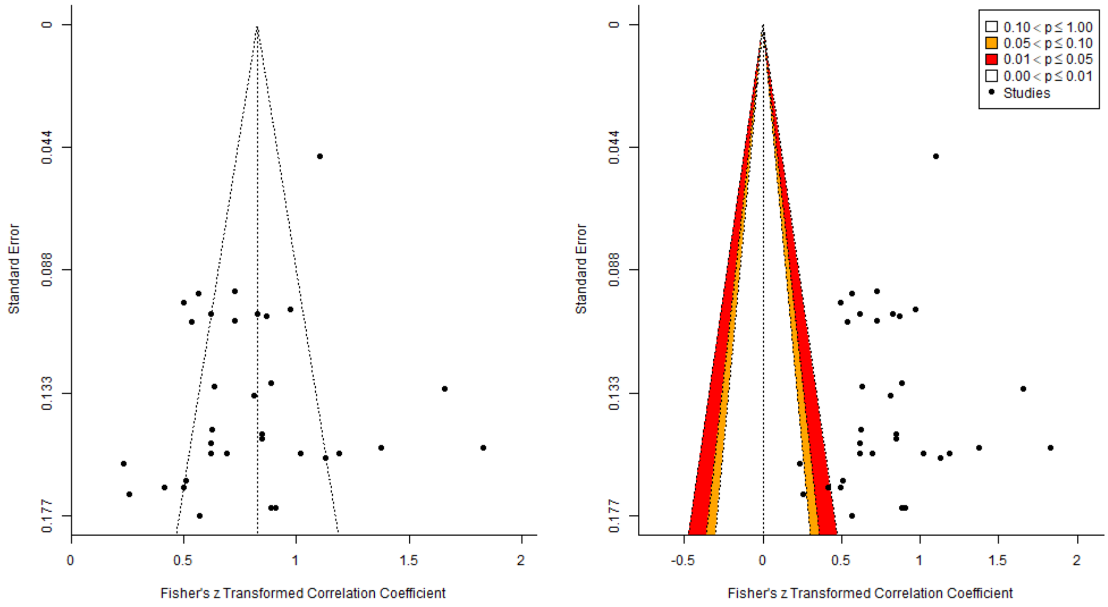
Note. s = number of studies; exp = number of experiments; ES = number of effect size estimates; z' = Fisher's z values; r = Pearson's R correlation; standard errors (SE) and z values for individual levels of a moderator; p values correspond to z or t - values; 95 % CI corresponds to the Fisher's z; [1] Only the study conducted by Feldman et al. (1995) included an ISI of 500ms, therefore it was not included in this analysis.

Publication bias

Visual inspection of the funnel and contour plots (Figure 3.5) shows no clear evidence of plot asymmetry or concentration of the effect sizes in the significance contours. This pattern is consistent with the non-significant Egger's test (*b* = -.33, *p* = .897).

Figure 3.5

Funnel plot showing effect sizes plotted against standard error for split-half reliability. A: funnel plot (left panel) and B: contour-enhanced funnel plot (right panel)



Discussion

The reliability of the SRTT is clearly understudied as, despite being one of the most commonly used experimental paradigms in procedural memory research, only 7 studies reported the psychometric properties of this task. Drawing on these studies, as expected (H1), the present meta-analysis shows evidence that the SRTT generally does not meet the standards for adequate test-retest reliability (i.e. *r* >.70, Burlingame et al., 1995; Nunnally & Bernstein, 1994), with an average test-retest reliability coefficient of between .28 and .30. This low test-retest reliability was observed irrespective of sampling and methodological considerations, which will be further discussed below. Split-half reliability, on the other hand, was better, with reliability coefficients varying between .66 and .69. Thus, this meta-analysis confirms poor across-session reliability procedural learning, in the context of near-acceptable within-session reliability as previously observed by (Oliveira et al., submitted)**.**

For test-retest reliability, no single sampling, methodological or analytical decision examined here appeared to be sufficient for reaching the threshold of adequate retest reliability. Although there were small numerical improvements in reliability for measures of procedural learning that account for participants' speed (i.e., ratio and random slopes) as opposed to difference scores, and for the probabilistic version of the SRTT when compared to deterministic tasks, neither of these factors significantly influenced reliability. For split-half reliability, numerically (but not significantly) better reliability was also observed for random slopes over difference scores and for studies with an interstimulus interval (250ms) when compared to those without (0ms). Counter to our predictions, we found no evidence of an effect of age and length of the test–retest interval on reliability. Yet, when considering the sample size, it is possible that the absence of a moderating effect of these factors may reflect lack of power since (West et al., 2018; West, Shanks, et al., 2021) reported a clear pattern of better test-retest and split-half reliability in adults than children. Furthermore, when a dichotomous approach (children vs adults) is taken instead, when examining the test-retest reliability, the overall reliability is only significant for adults (r = .36), but not children (.11), but there's still no significant difference between age groups. Additionally, the variability in the time scale between test and retest in this sample was quite limited. Finally, the absence of a moderating effect of the total number of trials may be due to the fact that all studies included in this meta-analysis used 500 or more trials per session. Even though we found no evidence for an effect of the number of trials, this should not be interpreted to suggest that the number of trials does not impact the reliability of the SRTT, primarily because experimental studies often adopt a considerably smaller number of trials (as low as 192 trials, (Schmalz et al., 2019)) than the ones reported here (Kidd, 2012; Kidd & Kirjavainen, 2011; Schmalz et al., 2019; Stoodley et al., 2006). Thus, this pattern does not necessarily reflect the number of trials adopted in previous experiments. It is possible that increasing the number of trials substantially more than this could lead to improvements in test-retest reliability by reducing trial noise (Rouder & Haaf, 2020).

Although the SRTT is well-known for producing a robust procedural learning effect at group-level, the findings from the present study raise questions about its suitability for individual differences research, since poor reliability contributes to attenuation of the association between measures (Rouder et al., 2019). Hierarchical modelling has been suggested as a way to disattenuate correlations (Matzke et al., 2017; Rouder et al., 2019), however, despite producing less biassed estimates than naive sample-effect correlations, the estimates are still highly variable (Rouder et al., 2019)**.** Thus, further investigation into the reasons for the lack of retest reliability is warranted, alongside efforts to develop tasks that are more suitable for eliciting adequate between subject variability. More reliable measures of procedural learning will help clarify whether the absence of correlations between procedural learning in the SRTT and language and literacy (Lammertink et al., 2020; West, Melby-Lervåg, et al., 2021) and between distinct measures of procedural learning (Arnon, 2020; Kalra et al., 2019; Siegelman & Frost, 2015; West, Melby-Lervåg, et al., 2021) reflect the lack of shared variance between these measures or whether individual differences fail to be captured due to poor reliability. Thus, the reliability issue is not only one issue of statistical importance, but also will help clarify theoretical issues pertaining to procedural learning as a construct and its role in language and literacy development and disorders.

Despite the common attribution of poor reliability to the use of difference scores (Castro-Schilo & Grimm, 2018; Trafimow, 2015), poor test-retest reliability cannot be solely explained by the measures used to index procedural learning, in light of similarly poor retest reliability when ratio scores or random slopes are used, and adequate split-half reliability regardless of the measure used. Yet, as previously recommended (Haines et al., 2020; von Bastian et al., 2020), the adoption of measures that more closely resemble the data generating process and that account for processing speed and trial noise (e.g., through Bayesian hierarchical modelling, see Haines et al., 2020; Rouder & Haaf, 2019), will potentially fare better at capturing the construct of interest. Unlike difference scores, which only provide point estimates of the individuals' performance, hierarchical models include information at the group and individual level, which has been found to better capture individuals' true ability (Haines et al., 2020). In the present meta-analysis, examined ratio and random slopes as these represent current practices in the field, but future research using more sophisticated models may be better able to separate measurement error from true individual differences.

In addition to resolving measurement and analytical challenges, it may be fruitful for future research to consider how performance in the SRTT may interact with other cognitive processes. For example, procedural learning effects have been shown to be positively associated with attention, with individuals with better sustained attention skills showing a larger procedural learning effect (Franklin et al., 2016; Oliveira et al., submitted; West, Shanks, et al., 2021). Thus, if individuals' alertness and motivation were to change between test and retest that would be likely to manifest in variations in their performance, consequently affecting the consistency of their ranking between test and retest. This may be expected to be more marked in children, given that their attentional skills are still developing (Levy, 1980) and attentional fluctuations have been previously found to decrease between childhood into young adulthood (Conners et al., 2003; Fortenbaugh et al., 2015). These changes between test and retest would be less influential for split-half reliability, as they would represent shorter-scaled differences in performance that would be captured in both odd and even trials; this is consistent with the finding of better split-half than retest reliability in the SRTT.

Similarly, practice effects, which are often observed in memory tasks (Beglinger et al., 2005; Calamia et al., 2012; Palmer et al., 2018; Scharfen et al., 2018), would also be expected to affect test-retest reliability more than split-half reliability. In the context of procedural memory, there is also the question of the extent to which task stability should be expected: individual performance is expected to change with practice, with an initial stage of procedural learning usually being marked by improvements in speed and accuracy, followed by consolidation and later automatization of the learnt probabilistic information (Dahms et al., 2020; Doyon & Benali, 2005). If these stages are captured by the SRTT, at least until automatisation has occurred, then performance should be expected to change across time. Further, the ranking between participants may also change as a result of individual differences in the rate at which participants make the transitions between stages of learning (as has been observed in other memory tasks (Dikmen et al., 1999; Temkin et al., 1999).

Taken together, this meta-analysis demonstrates that procedural learning in the SRTT exhibits suboptimal test-retest reliability, irrespective of the sampling and methodological manipulations explored here. Split-half reliability, on the other hand, is considerably better, indicating some degree of consistency within sessions. While some design features contributed to small improvements in reliability, none resulted in adequate reliability. It is possible that their cumulative impact could lead to significant increases in test-retest reliability, however due to our limited sample size that was not possible for this paper. Unfortunately, due to the lack of reporting of psychometric properties of the SRTT, further research is needed to adequately determine the impact of methodological factors by systematically investigating their influence on reliability. While it may not pose a major concern for group comparisons, individual differences research needs to be interpreted in light of the low measurement reliability of the SRTT (Parsons et al., 2019). The absence of correlations between measures thought to tap the same construct is often inferred as pointing towards domain specificity or lack of shared variance between measures, when it may simply reflect attenuation due to measurement error (Rouder et al., 2019). Until adequate reliability is established for existing procedural memory tasks, or new reliable measures are developed, the field of procedural memory will continue to be hampered by underspecification of its components and a poor understanding of its relationship with cognitive constructs of interest, such as language.

Chapter 4. Limited Evidence of An Association Between Language, Literacy, And Procedural Learning Across Typical and Atypical Development: A Meta-Analysis

Abstract

The ability to extract patterns from sensory input across time and space is thought to underlie the development and acquisition of language and literacy skills, particularly the subdomains marked by the learning of probabilistic knowledge. Thus, impairments to the procedural learning mechanisms are hypothesised by the procedural deficit hypothesis to underlie neurodevelopmental disorders such as dyslexia and developmental language disorder. In the present meta-analysis, comprising 2396 participants from 39 independent studies, the relationship between language, literacy, and procedural memory on the Serial Reaction Time task (SRTT) was assessed across children and adults with typical development, dyslexia and DLD. Based on the procedural/declarative model a positive relationship was expected between procedural memory and language and literacy measures for the typically developing group; however, no such relationship was observed. This was also the case for the disordered groups (*p*s > .05). Also counter to the procedural deficit hypothesis, the magnitude of the relationship between procedural learning and grammar and phonology did not differ between TD and DLD (*p*s > .05), nor between the TD and dyslexic group on reading, spelling, and phonology (*p*s > .05). Whilst lending little support to the procedural/declarative model and the procedural deficit hypothesis, we consider that these results may be the consequence of poor psychometric properties of the SRTT as a measure of procedural learning.

Introduction

Procedural learning refers to the ability to learn, consolidate and control motor and cognitive skills that require the integration of statistical, probabilistic and sequence knowledge (Batterink et al., 2019; Packard & Knowlton, 2002; Ullman, 2004; Ullman et al., 2020; Ullman & Pullman, 2015). This memory system has been proposed to support the development and acquisition of language (Ullman, 2004; Ullman et al., 2020), specifically for linguistic subdomains that require the extraction of patterns, such as phonology and grammar (Conway et al., 2008; Christiansen et al., 2012; Ullman, 2004; Ullman & Pierpont, 2005; Ullman et al., 2020). Deficits of procedural memory are also claimed to be a core causal factor in dyslexia and developmental language disorder (Ullman, 2004). However, recent meta-analyses examining the relationship between procedural learning and language and literacy provide weak support for this hypothesis (Hamrick et al., 2018; Lammertink et al., 2020; West, Melby-Lervåg, et al., 2021). Furthermore, extant correlational meta-analyses have been limited as they mainly focussed on grammar and vocabulary (Hamrick et al., 2018; Lammertink et al., 2020), with the exception of West, Melby-Lervåg, et al. (2021) which combined measures of language and literacy. Therefore, here, we extend existing meta-analyses by producing the largest and most comprehensive meta-analysis examining relationships between procedural learning and different language and literacy subdomains, and directly comparing the magnitude of the relationship between language and literacy and procedural learning in typical and atypical populations with language-related learning difficulties (dyslexia and developmental language disorder).

Declarative/Procedural Model and the Procedural Deficit Hypothesis

According to the Declarative/Procedural model (Ullman, 2004), the detection and acquisition of regularities in the input has been hypothesised to underlie the development of language and literacy skills, in addition to other non-linguistic skills (e.g., musical timbre sequences, Tillmann & McAdams, 2004). The role of procedural learning in language is suggested to be particularly relevant for the acquisition of probabilistic/regularity-based components such as grammar and phonology (Ullman, 2004). Conversely, the declarative memory system, which underpins the acquisition of arbitrary and idiosyncratic knowledge, is thought to be relevant to the accumulation of vocabulary knowledge fundamental to the mental lexicon. The procedural deficit hypothesis further claims that a neurological impairment in the procedural memory system could account for the language and literacy difficulties (i.e., with grammar and phonology) experienced by individuals with DLD and dyslexia, respectively (Ullman, 2004; Ullman et al., 2020; Ullman & Pullman, 2015). Thus, the procedural deficit hypothesis emerges as an alternative core-deficit account of dyslexia and DLD which aims to explain the range of profiles observed within these diagnostic categories (Ullman et al., 2020; Ullman & Pierpont, 2005), including non-linguistic difficulties with motor skills, attention and working memory (Baird et al., 2010; Brookman et al., 2013; Buchholz & McKone, 2004; Delage & Frauenfelder, 2020; Fostick & Revah, 2018; Hill, 2001; Nicolson & Fawcett, 1994; Romani et al., 2011). Despite the predicted difficulties in procedural memory, individuals with DLD and dyslexia are thought to have relatively intact declarative memory. Moreover, it is claimed that individuals with DLD and dyslexia compensate for their procedural deficits by relying on the declarative system. Crucially, it has been predicted that the association between language/literacy and procedural memory might be the same or weaker for disordered populations than TD children/adults (Lum et al., 2012). A smaller association between language and literacy and procedural memory may occur for domains more likely to be compensated by declarative memory (e.g., grammar, word reading; Ullman et al., 2020), thus leaving little variability for procedural memory to explain. On the other hand, if compensation does not occur, or occurs to a lesser extent, a similar association to the TD group will be expected, with those with better procedural learning abilities in the disordered groups being expected to show better language and literacy abilities (Lum et al., 2012**.**

A substantial body of research has examined the procedural learning abilities of typically developing children and adults and those with these developmental disorders using the SRTT (Nissen & Bullemer, 1987). In this task participants are usually presented with a stimulus that appears in one of four locations on screen with participants being asked to respond as quickly as possible to its position by pressing the corresponding key in the keyboard. Unbeknownst to the participants, some of the positions of the stimulus follow a pattern. As participants learn the sequence, they begin to anticipate sequenced trials, resulting in faster response times to these trials compared to random trials (Barker, 2012; Nissen & Bullemer, 1987; Schwarb & Schumacher, 2012). Meta-analyses comparing the performance of individuals with dyslexia and DLD to typically developing individuals have found a group deficit in DLD and dyslexia in the SRTT (Lum et al., 2013, 2014; West, Melby-Lervåg, et al., 2021), with the magnitude of the effect size being significantly predicted by the interaction between age of participants (Lum et al., 2013, 2014) and the number of exposures to the sequence (Lum et al., 2014) or the type of sequence (Lum et al., 2013). Yet, as noted by West, Melby-Lervåg, et al. (2021), the effect size for this group difference tends to be small (*g* = -.30).

Contrary to the predictions of Ullman and colleagues that follow from the procedural/declarative model of language learning (Lum et al., 2012; Ullman, 2004), only a handful of studies have found correlations between the procedural learning effect captured by the SRTT and standardised measures of language in samples of TD children (Clark & Lum, 2017; Desmottes et al., 2016, 2017; Lum et al., 2012). Other studies have found no association (Desmottes et al., 2017; Gabriel et al., 2015; Henderson & Warmington, 2017; Siegelman & Frost, 2015; Vakil et al., 2015; West et al., 2019). Similarly, some studies have found associations between procedural learning and language in children with DLD (e.g., Desmottes et al., 2016) and dyslexia (e.g., Vakil et al., 2015), whereas others have not (e.g., DLD: Clark & Lum, 2017; Desmottes, Maillart, et al., 2017; Gabriel et al., 2015; Lum et al., 2012; dyslexia: Deroost et al., 2010; Henderson & Warmington, 2017; West et al., 2019). One explanation for the inconsistencies across studies arises from recent findings that the SRTT (as a measure of procedural learning) has low reliability both in children and adults (Kalra et al., 2019; Siegelman & Frost, 2015; Stark-Inbar et al., 2017; West et al., 2018; West, Shanks, et al., 2021). Poor psychometric properties are particularly problematic for individual difference studies as they contribute to the attenuation of the association between measures (Fleiss, 1986; Rouder et al., 2019; Spearman, 1904).

Another explanation for the inconsistent results is that a more nuanced approach is needed to account for when these correlations will arise, taking age and language proficiency into account. In four meta-analyses that included 16 studies, Hamrick et al. (2018) examined the link between language and declarative and procedural memory in first (children) and second language learners (adults). As proposed by Ullman and colleagues (Ullman, 2004; Ullman & Pierpont, 2005), in the early stages of second language learning language acquisition should initially be reliant on the declarative memory system, with an increase in the involvement of the procedural memory system with increasing exposure and proficiency. This is thought to occur due to the gradual nature of procedural learning which requires more exposures to stimuli but fewer attentional resources than declarative memory (Ullman, 2013) and culminates in the automaticity of the learned skill, which allows the behaviour to be performed effortlessly. Hamrick et al's (2018) findings were in line with these predictions, as they observed that for children measures of declarative memory correlated with lexical abilities (*r* = .41) and to a lesser extent with grammar (*r* = .16); grammar also positively correlated with procedural memory (*r* = .27). However, for adult second language learners subdivided into low and high proficiency groups, only the group with high proficiency showed a positive association between procedural learning and grammar (high: *r* = .55; low: *r* = -.01), whilst for the lower L2 proficiency group grammar was solely associated with declarative memory (high: *r* = -.07; low: *r* = .46). These results lend support to the Declarative/Procedural model as grammatical development appears to be associated with the procedural memory system in first language learners and at later stages of proficiency in L2 learners, whilst in individuals with low proficiency grammatical skills are more strongly associated with declarative memory (Ullman, 2004; Ullman et al., 2020). Consequently, if populations with language disorders are expected to perform similarly to low proficiency groups, then children with DLD may also rely primarily on the declarative system as a means of compensating for the impairment in procedural memory (Ullman et al., 2020).

However, contrary to this, two recent meta-analyses found no evidence of a relationship between procedural learning in the SRTT and language/literacy measures (Lammertink et al., 2020; West, Melby-Lervåg, et al., 2021). West, Melby-Lervåg, et al.'s (2021) meta-analysis comprehensively examined the role of procedural learning on language and literacy development across a set of tasks all considered to tap into procedural learning (e.g., SRTT, Hebb serial order learning task, Artificial Grammar learning and statistical learning tasks, amongst others). The results pertaining to the SRTT are of particular interest to the current study: data from 441 participants (drawn from 5 studies) revealed a negligible association between procedural learning and measures of language and decoding in TD children and adults (*r* = .03; West, Melby-Lervåg, et al., 2021). Similarly, the large-scale meta-analysis conducted by Lammertink et al. (2020) (N = 139 children with DLD and N = 573 TD children; 19 studies), also found no evidence of an association between procedural learning in the SRTT and expressive (*r* = .07) and receptive grammar (*r* = .05) in children with or without DLD (Lammertink et al., 2020), with no statistical difference between groups. As discussed below, whilst the absence of an association between grammar and procedural learning in children with DLD is consistent with the predictions of the procedural deficit hypothesis, this pattern of results for school-aged TD children and adults does not.

In light of these findings, there seems to be weak (or, at best, mixed) evidence for a relationship between procedural learning and language/literacy in the SRTT. Yet, given the predictions of the procedural deficit hypothesis, which does not propose a unique contribution of procedural learning to the acquisition of grammar, but for all rule-based knowledge, further research on the role of procedural memory on language and literacy development more broadly is required. Thus, contrary to previous meta-analyses which have mostly focussed on the relationship between vocabulary and grammar and procedural memory (Hamrick et al., 2018; Lammertink et al., 2020) or analysed the relationship between language and literacy and procedural memory only in a small sample of studies (West, Melby-Lervåg, et al., 2021), the present meta-analysis, with the largest sample to date, aims to extend these findings by analysing the relationship between procedural learning and different subdomains, namely grammar, vocabulary, phonology, spelling, reading, in school-aged children and adults (5-27 years) with and without language and literacy impairments.

Based on the procedural deficit hypothesis (Ullman & Pullman, 2015), distinct patterns of association are hypothesised to occur for these populations as the declarative memory system has been proposed to compensate for the procedural learning impairments in populations with DLD and dyslexia. Specifically, a stronger association between procedural learning and grammar would be expected for typically developing children than those with DLD (as found by Lum et al., 2012). Similar findings would be expected for children with dyslexia, where the declarative memory system would be expected to compensate for the literacy deficits which are also proposed to emerge as a consequence of procedural learning impairments (Lum et al., 2013; Ullman, 2004; Ullman et al., 2020). Furthermore, since the ability to compensate for the language and literacy deficits would be expected to increase with age as the declarative memory system matures (Lum et al., 2013; Ullman, 2004; Ullman & Pullman, 2015), age is expected to be a significant moderator of the association between procedural learning and language and literacy measures, especially for individuals with DLD and dyslexia.

Finally, considering the methodological variability in how the SRTT is designed and implemented across different studies (Schwarb & Schumacher, 2012) and how these methodological decisions have been found to moderate the magnitude of the difference between dyslexic and control groups in the SRTT, factors such as the a) number of trials, b) type of sequence and c) type of SRTT were included as potential moderators. Previous meta-analyses have found that studies with a larger number of trials (Lum et al., 2013, 2014) showed a smaller difference between disordered and TD groups. Similarly, the size of the effect was also found to be moderated by the type of sequence, with second order conditional sequences being associated with smaller effect sizes when comparing children with and without dyslexia in the SRTT (Lum et al., 2013), and type of SRTT, with deterministic sequences showing larger effect sizes than alternating sequences (West, Melby-Lervåg, et al., 2021) although note that this was not the case in Lammertink et al. (2020), who did not find a moderating effect for the type of sequence. Thus, by taking these factors into account, we aim to determine whether methodological decisions impact the strength of the association between language and literacy and procedural learning tasks and thus have contributed to the inconsistent pattern of findings across studies.

Aim and research questions

The current meta-analysis aims to provide a more comprehensive analysis of the predictions of the procedural deficit hypothesis. By including language and literacy measures across subdomains, this meta-analysis will be able to assess the core predictions of the procedural deficit hypothesis in children and adults with and without language and literacy impairments. Furthermore, by investigating the effect of the moderating variables, this meta-analysis may also provide some possible explanations for the inconsistent pattern of results in the literature. Unlike previous meta-analyses, the present study includes effect sizes for all available language and literacy measures as, instead of aggregating the effect sizes derived from a single sample, we take advantage of multilevel models to deal with the non-independent effect sizes, thus preventing information loss.

We addressed the following pre-registered (https://osf.io/vtdg3) research questions and hypotheses:

Research Question 1: We examined the relationship between procedural learning and language/literacy abilities in TD adults and children. Specifically, following the declarative/procedural model, we predicted correlations between procedural learning and grammar, phonology, reading and spelling (Ullman et al., 2020), whilst vocabulary was expected to only be weakly correlated with procedural learning (Hamrick et al., 2018; Ullman, 2004; Ullman et al., 2020) (Hypothesis 1 (H1).

Research Question 2: In line with the procedural deficit hypothesis (Ullman, 2004; Ullman et al., 2020; Ullman & Pullman, 2015), we anticipated that group membership would moderate the relationship between procedural learning and language/literacy abilities, with a) stronger associations expected between grammar, phonology and procedural learning for TD groups than DLD groups, based on the proposal that individuals with DLD compensate for difficulties in grammatical and phonological acquisition with the declarative learning system (H2) (Ullman, 2004; Ullman & Pullman, 2015; Ullman et al., 2020); and similarly b) stronger correlations between phonology, reading, spelling and procedural learning for TD groups than for dyslexic groups, as individuals with developmental dyslexia would be predicted to compensate for phonology, reading and spelling difficulties with the declarative learning system (H3) (Ullman, 2004; Ullman & Pullman, 2015; Ullman et al., 2020);

In a set of exploratory analyses, we also examined whether age, the number of SRTT sessions, sequence complexity, and SRTT type (deterministic vs probabilistic sequences) moderate the relationship between procedural learning and language/literacy abilities.

Methods

All materials for this meta-analysis are available (https://osf.io/ev2xw/), including the dataset and scripts necessary to replicate all reported analyses and plotting.

Search strategy

To find eligible studies, literature searches were conducted up to November of 2020 on Pubmed and Google Scholar using the following search terms:

PUBMED: procedural learning OR procedural memory OR sequence learning OR implicit learning AND language OR reading OR dyslexia OR language impairments

procedural learning OR procedural memory OR sequence learning OR implicit learning AND SRT OR Serial Reaction Time task AND language OR reading OR dyslexia OR language impairments

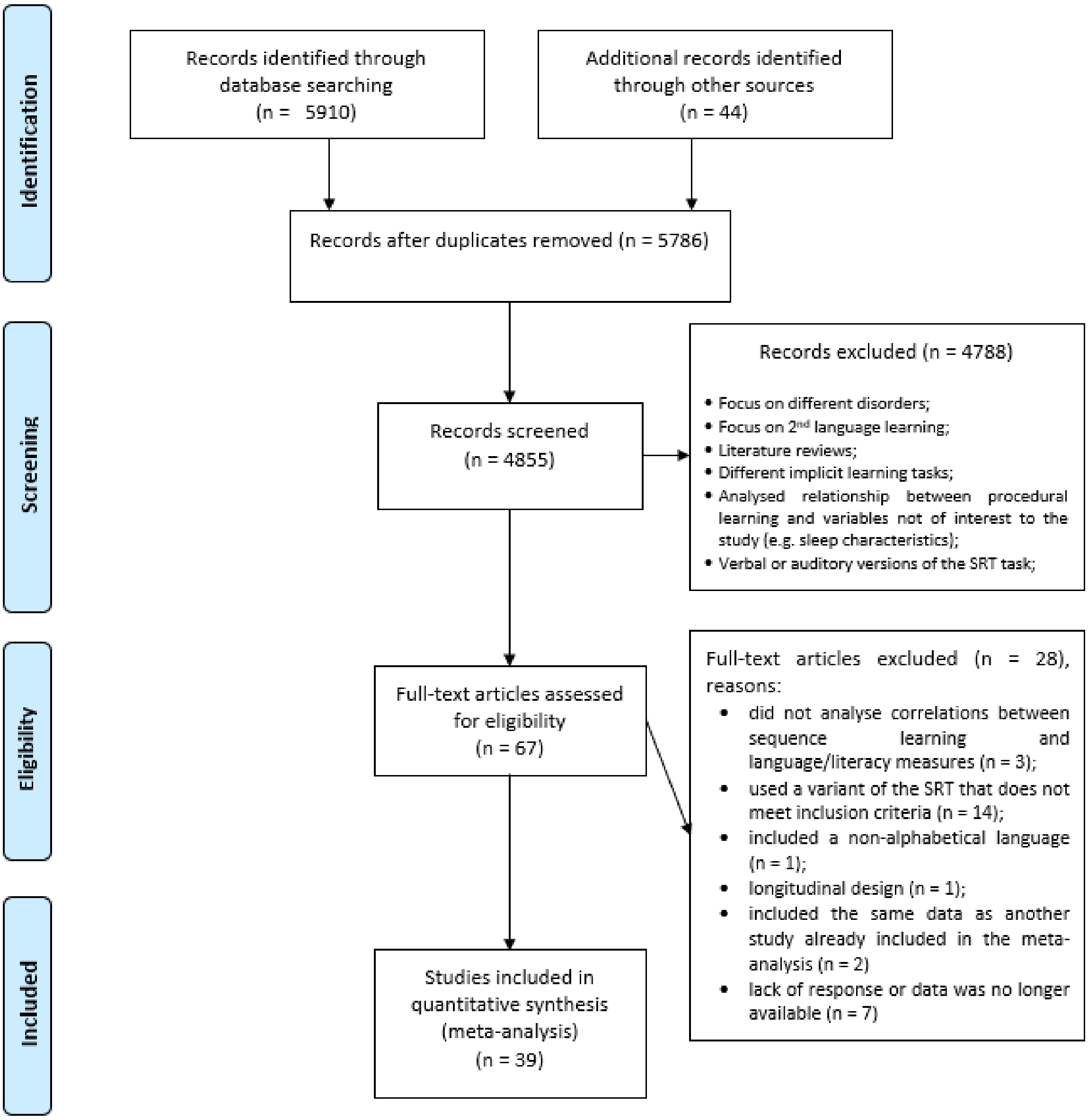
GOOGLE SCHOLAR: Relationship AND language AND “serial reaction time task” AND visual sequence learning

Once the search was completed, the first author screened all titles and abstracts (Figure 4.1) for records which analysed the procedural learning, language, or literacy skills of individuals with and without dyslexia and developmental language disorder. 67 records met these criteria and were then subjected to a full-text analysis against the inclusion and exclusion criteria to determine eligibility. Full-text eligibility was assessed using the following inclusion criteria: i) Population: TD controls and/or individuals with language/reading impairments of speakers of alphabetic languages; ii) Used a strictly visual deterministic, probabilistic or alternating SRTT with procedural learning computed as the difference between sequenced and random/improbable trials; Audio-visual SRTTs or tasks that alternated types of statistical dependencies were not included (e.g., Jackson et al., 2020); iii) Analysed the relationship between language/literacy and procedural learning and reported Pearson’s correlation (or equivalent) coefficients; iv) If correlations were missing but the required measures were used, we solicited the correlation coefficients directly from the authors; v) English language publication; vi) If the same results were published in multiple articles, these were only reported once in the meta-analysis; we selected the publication with the largest sample size and most comprehensive information; vii) Publication dates: between 2000 and 2020 to increase the chances of data availability; and exclusion criteria: i) Second language learning studies; dual task paradigms with the SRTT; longitudinal studies that measured procedural learning and language measures at different time points (e.g., language measures in infancy and procedural learning in adulthood); SRTTs which only involved 3 positions as typically 4 positions are used; studies that used adaptations that deviate substantially from the task described by Nissen and Bullemer (1987) (e.g., in the task used by Vicari et al. (2003) participants were expected to only provide a motor response to a subset of the stimuli).

Forward and backward searches were conducted on these records. Discrepancies regarding an article’s eligibility were resolved amongst the authors until a consensus was reached. Once the list of articles to be included in the meta-analysis was agreed upon it was sent to two independent experts in the field for feedback to ensure that relevant articles were not missed.

Figure 4.1.

PRISMA flowchart showing selection of studies for meta-analysis on the relationship between language and literacy and procedural learning


Identification
Screening
Eligibility
Included
Records identified through database searching (n = 5910)
Additional records identified through other sources (n = 44)
Records after duplicates removed (n = 5786)
Records screened (n = 4855)
Records excluded (n = 4788)
• Focus on different disorders;
• Focus on 2nd language learning;
• Literature reviews;
• Different implicit learning tasks;
• Analysed relationship between procedural learning and variables not of interest to the study (e.g. sleep characteristics);
• Verbal or auditory versions of the SRT task;
Full-text articles assessed for eligibility (n = 67)
Full-text articles excluded (n = 28), reasons:
• did not analyse correlations between sequence learning and language/literacy measures (n = 3);
• used a variant of the SRT that does not meet inclusion criteria (n = 14);
• included a non-alphabetical language (n = 1);
• longitudinal design (n = 1);
• included the same data as another study already included in the meta-analysis (n = 2)
• lack of response or data was no longer available (n = 7)
Studies included in quantitative synthesis (meta-analysis) (n = 39)

Data extraction

Articles included in the meta-analysis were coded by the first author and a second coder blind to the purpose of the study using the pre-registered data extraction form developed for the current review and available at https://osf.io/ev2xw/. Data extraction was compared until 100% agreement was reached between coders. The following moderators were included: age of participants, group (TD, DD/dyslexic or DLD groups), domain of interest (language or literacy), subdomains (phonology, vocabulary, grammar, reading or spelling[2]), type of SRTT (deterministic, probabilistic, or alternating), sequence complexity (first (FOC) or second (SOC) order conditional), number of trials the participants were exposed to, and number of sessions (including training and testing sessions).

Meta-analytic approach

The effect size metric of Pearson's r was used to represent the strength and direction of the relationship between procedural learning and language/literacy skills. Correlation coefficients were re-coded to reflect a positive relationship (higher procedural learning and higher language/literacy skills) when necessary. When missing data was identified, the corresponding authors were contacted by email requesting the relevant information; this had the aim of ensuring all correlations - rather than only significant ones reported in the published papers - were included, as reporting biases of this nature have been shown to inflate the overall effect (Kirkham et al., 2010). In the absence of a reply or data being unavailable, the articles were not included in the meta-analysis.

All correlation coefficients were converted from Pearson's r to Fisher's z scale as Pearson's r is not normally distributed (Hedges & Olkin, 1985). Since most studies contributed multiple correlation coefficients to the meta-analysis, to deal with the lack of independence across effect sizes and avoid reducing power by calculating the average for the effect sizes for these studies, robust variance estimation (RVE) was used for model estimation alongside small-sample corrections (Hedges et al., 2010; Tipton & Pustejovsky, 2015) via the *robumeta* package for R (Fisher & Tipton, 2015). RVE methods use a working model that approximates the dependence structure but does not require exact knowledge of the error distribution or covariance structure between effect sizes estimates. By using RVE methods, even when the working model is misspecified, the meta-regression coefficient estimates will be unbiased (Fisher & Tipton, 2015; J. E. Pustejovsky & Tipton, 2022).

Modelling

An intercept-only meta-regression was initially run to estimate the overall effect size between procedural learning and language/literacy across groups. Two separate random effects models for language and literacy with correlated effects dependencies were computed using the *robumeta*

[2] Other subdomains were coded (e.g., reading comprehension) yet due to their small sample size separate analyses for these subdomains were not conducted. Instead, they have only been included in the analyses for the overall effect of language or literacy.

package (Fisher & Tipton, 2015). Even though there is also evidence of hierarchical dependencies (e.g., some research groups contributed multiple studies to the meta-analysis), weights for correlational dependency were selected as recommended by Tanner-Smith et al. (2016) since this was the most common form of dependency in the present dataset. For all analyses, the in-study effect size correlation (ρ) was set at .8. In addition, sensitivity analyses were performed across varying values of rho (.0, .2, .4, .6, .8, 1.0) to assess whether results were robust to changes in rho values (Table 3).

To answer the research questions, separate RVE mixed-effects meta-regression models were performed with *group* as a moderating variable to determine whether there is a significant correlation between language and literacy for the TD group and disordered groups (represented in Table 2 as "*TD=DLD=DD=0*", with DD standing for the dyslexic group) and whether the magnitude of this effect is greater for this group than for the clinical groups (represented in Table 2 as "*TD vs DLD vs DD*" or a subset of the groups depending on the research questions). Further moderator variables (e.g., age, type of sequence) were examined to assess their effect on the association between procedural learning and language/literacy for the overall sample (Tables 1 and 3), with the categorical levels contrasted in a similar manner to group (e.g., "FOC vs SOC" - representing the contrast between first and second-order conditional sequences), using the *Wald_test()* function from the *clubSandwich* package (Pustejovsky, 2021). This function uses a method called approximate Hotelling's T^2 test which has been shown to perform adequately even with degrees of freedom close to 0 (Tipton & Pustejovsky, 2015).

After performing the meta-analytic calculations, Fisher's z overall estimates were converted back to Pearson's r for reporting the average correlation and 95% confidence interval (CI) for each model.

Bias and heterogeneity analyses

To test for the presence of publication/reporting bias, while taking account of the dependencies in the data analysis, the PET-PEESE estimates (Stanley & Doucouliagos, 2014) were computed using RVE via the *robumeta* package (Fisher & Tipton, 2015). PET (Precision-Effect Test) and PEESE (Precision-Effect Estimate with Standard Error) methods use the standard error and the sampling variance as predictors, respectively. The two-step method PET-PEESE is recommended because both PET and PEESE methods show bias, with PET being downwardly biased when the true effect is different from zero, while PEESE shows an upward bias when the true effect is zero (Stanley & Doucouliagos, 2014). The two-step process involves firstly assessing whether the PET estimate is significant, if the effect is significant then the PEESE estimate is used; if not, the PET estimate should be adopted (Stanley & Doucouliagos, 2014).

Since not all classical methods are available to analyse outliers, influential cases, and bias for models with dependency, effect sizes were aggregated via the *agg* function from the *MAc* package (Del Re & Hoyt, 2018). Following Borestein (2009), the default correlation between within-study effect sizes was set at .50, with complementary sensitivity analyses to ensure the robustness of the results (r = .10, .30, .50, .70, .90). The *influence* function from the *metafor* package was used to identify potential outliers and influential cases for the aggregated effect sizes for each model (Viechtbauer, 2010). To detect evidence of publication bias, funnel and contour-enhanced funnel plots were also produced (Galbraith 1988; Vandenbroucke 1988) using the *metafor* package (Viechtbauer, 2010). Rank correlation tests were performed to assess funnel plot asymmetry (Begg & Mazumdar, 1994).

Study heterogeneity was analysed using the Q statistics, I^2 and τ^2 (Higgins & Thompson, 2002). The Q test assesses heterogeneity by comparing the effect sizes across studies to determine whether all studies show the same effect (null hypothesis). However, this test has been shown to be poor at detecting true heterogeneity in smaller samples due to lack of power. Thus, the results of the Q test need to be considered in light of other measures of heterogeneity. I^2 represents the proportion of variability across studies due to heterogeneity rather than chance. This measure was introduced by Higgins and Thompson (2002) as being more interpretable and comparable across studies than Q statistics. Yet, given the reliance on the Q statistic for its calculation, I^2 is also often imprecise and/ or biased in small samples (von Hippel, 2015). For ease of interpretation, heterogeneity above 50% and 75% tends to be considered moderate and substantial, respectively (Higgins & Thompson, 2002)**.** Finally, τ^2 - the variance in the true effect sizes - was also interpreted since random effects models were adopted across analyses as we anticipated that studies did not represent an homogeneous population due to the methodological and sampling differences (Borestein, 2009)**.**

Correction for attenuation

In classical test theory observed scores are thought to reflect the true scores plus measurement error (Novick, 1966), with measurement error often limiting the size of the correlation between two variables (Spearman, 1904; Wiernik & Dahlke, 2019). Thus, as recommended by Wiernik and Dahlke (2019), the pooled correlation coefficient between language and literacy and procedural learning was corrected for attenuation using the Spearman's derivation (Spearman, 1904).

$$true\ correlation\ = \frac{observed\ correlation}{\sqrt{reliability(x).reliability(y)}}$$

This method adjusts the raw correlation by taking into consideration the reliability estimates for each measure (Spearman, 1904). Considering the scarce information about measurement reliability, an artefact distribution method was used, as measurement error correction was only performed after model estimation (Wiernik & Dahlke, 2019).

The overall reliability for the SRTT was estimated to be approximately .30, based on a recent meta-analysis conducted by Oliveira, Hayiou-Thomas, et al. (in prep). Unfortunately only a small number of studies reported the reliability of the tasks adopted (namely, Siegelman & Frost, 2015; West et al., 2018; West, Shanks, et al., 2021), thus the estimate may not be representative of the reliability of the SRTT used in the studies included in this meta-analysis. For the literacy and language measures, an overall reliability of .70 was selected based on the frequent test-retest reliability of standardised measures used to assess language and literacy. However, as discussed by Rouder et al. (2019), disattenuation methods are flawed and can produce highly variable estimates which can be both inflated and deflated, so these need to be interpreted with caution.

Results

In total, the meta-analysis comprised 39 independent studies, summarising 500 effect sizes and data from 2396 participants. Participants' age range: [5.2, 27.7], M = 12.69, SD = 5.64. See https://osf.io/ev2xw/ for the entire dataset.

Relationship between language and literacy and procedural learning

To directly test the predictions of the procedural/declarative model and the procedural deficit hypothesis separate analyses were required for the TD, DLD and dyslexic groups. Yet, before presenting those results, we started by examining the overall effects across populations for both literacy and language.

1.1 Overall effect

All studies (k = 39) were included in this analysis (citations marked with an asterisk). The estimated average correlation between procedural learning and overall language and literacy measures for all studies using RVE was Fisher's z = .06, 95% CI [.007, .12] SE = .03, t(32.7) = 2.3, p = .028, with a Pearson's r of .06, indicating a very modest - though statistically significant at this sample size - association between procedural learning and language/literacy. Sensitivity analyses indicated that the results are robust to different values of rho (Fisher's z varied between .0619 and .0620). There was no significant difference between language and literacy measures in terms of the strength of the relationship with procedural learning (F(1, 22.3) = .34, p = .563). Note that this result is not necessarily at odds with our predictions as the relationship between language and literacy and procedural memory may have been shifted downwardly by the inclusion of effect sizes from the disordered groups. A direct test of our predictions, based on the procedural/declarative model, requires separate analyses for TD and disordered groups.

To further explore the relationship between procedural learning and language/literacy separate RVE models were computed for language and literacy measures.

1.2 Language and procedural learning

Overall effect across participants

35 studies reported the relationship between procedural learning and language (marked with an asterisk). Fisher r-to-z transformed correlation coefficients ranged from −.97 to .91, with just over half of the estimates being positive (55%). However, the average correlation between procedural learning and spoken language measures was again very modest: Fisher's z = .06, 95% CI [-.002, .12] SE = .03, t(30.1) = 1.98, p = .057, with an equivalent Pearson's correlation of r = .06, 95% CI [-.002, .12]. Sensitivity analyses confirmed that this result was robust to different levels of rho. There was a moderate level of heterogeneity in the effect sizes τ^2 = .028, I^2 = 51.72, which was further explored using meta-regression analyses (presented in Tables 1-2). As before, this result offers no support for or against our predictions given that the overall estimate may be smaller due to the inclusion of the disordered groups.

Moderator analyses

Results of all separate RVE meta-regressions with moderator variables as predictors are shown in Table 1 and 2. F-tests were used to compare the estimates between levels of categorical predictors.

There was no evidence that group membership affected the magnitude of the pooled association between procedural learning and language (Table 1). Subgroup RVE meta-analyses (Table

2) for each linguistic domain (grammar, vocabulary, and phonology) were conducted separately to assess whether there were group differences in the relationship between language subdomains and procedural learning (H1/H2/H3). Since only one study included vocabulary measures for the dyslexic group, this group was removed from the vocabulary meta-regressions. The overall pattern was consistent across analyses, with no evidence of a significant relationship between procedural learning and language subdomains regardless of group (grammar: *F*(3, 3.02) = 1.40, *p* = .394; phonology: *F*(3, 67.83) = .78, *p* = .539; vocabulary: *F*(2, 11.5) = 1.3, *p* = .311) and there was no evidence that the strength of the association differed between the TD and DLD groups for grammar (*F*(1, 17.1) = 1.07, *p* = .315) or phonology specifically (F(1, 9.02) = 2.44, p = .153), nor between the TD and dyslexic groups for phonology (*F*(1, 7.95) = 0.628, *p* = .451). Thus, there was no evidence supporting H1 as there was no evidence of a relationship between language and procedural memory on the SRTT for the TD group. Furthermore, contrary to our predictions, the relationship between procedural learning and phonology did not differ between the TD and disordered groups (H2, H3), nor did the association between grammar and procedural memory differ between the TD and the DLD group (H2). Disattenuated correlations for the group comparisons (represented as R) are also presented in Table 4.2. These show small correlations for the TD group between procedural learning and grammar (R = .19) and vocabulary (R = .12). Small associations between procedural learning and phonology (R = .10) and grammar (R = .19) were also observed for the dyslexic group. For the DLD, on the other hand, there was a moderate association between phonology and procedural learning (R = .46). All the other correlations were negligible (R < .1).

The moderating effect of sampling and methodological differences (domain, age of participants, sequence complexity, type of SRTT, session or number of trials) was also tested on the entire sample, yet there was no evidence of a moderating effect of any of these factors on the relationship between language and procedural memory (Table 4.1).

Model diagnostics

Sensitivity analyses revealed consistent results across rho values for all meta-analytic analyses (see Tables 4.1 and 4.2). To assess the presence of influential studies, effect sizes were aggregated per study as these analyses are not available for RVE models. There was no evidence of influential studies.

Table 4.1

Results of all separate meta-regressions with moderator variables for language measures

Moderator (bolded) and level	***Study characteristics***			***Effect size***		***Test of significance***					***Heterogeneity***		***Sensitivity analysis***
	s	k	F	Fisher's z	r	SE	t	p	95% CI		τ^2	I^2	range
Group	35	323	__	__	__	__	__	__	__	__	.029	52.572	__
TD = DLD = DD = 0	__	__	1.23	__	__	__	__	.333	__	__	__	__	__
TD vs DLD vs DD	__	__	.124	__	__	__	__	.884	__	__	__	__	__
TD	__	__	__	.056	.056	.033	1.719	.100	-.012	.125	__	__	.0564; .0565
DLD	__	__	__	.083	.082	.090	.917	.376	-.112	.278	__	__	.0826; .0830
DD	__	__	__	.030	.030	.060	.500	.637	-.120	.179	__	__	.0298; .0299
Subdomain	35	313	__	__	__	__	__	_	__	__	.030	53.122	__
Grammar = Phonology = Vocabulary = 0	__	__	1.26	__	__	__	__	.318	__	__	__	__	__
Grammar vs Phonology vs Vocabulary	__	__	.213	__	__	__	__	.810	__	__	__	__	__
Grammar	__	__	__	.041	.041	.039	1.050	.307	-.041	.122	__	__	.0405; .0407

Phonology	__	__	__	.088	.088	.079	1.120	.291	-.089	.266	__	__	.0883; .0886
Vocabulary	__	__	__	.067	.067	.044	1.540	.160	-.032	.166	__	__	.0669; .0672
Age	35	323	__	.003	.003	.006	.575	.578	-.009	.015	.029	52.610	.0226; .0227
Sequence complexity	33	298	__	__	__	__	__	__	__	__	.026	50.631	__
FOC = SOC = 0	__	__	2.85	__	__	__	__	.082	__	__	__	__	__
FOC vs SOC	__	__	.130	__	__	__	__	.721	__	__	__	__	__
FOC	__	__	__	.081	.081	.045	1.810	.091	-.015	.177	__	__	.0813; .0815
SOC	__	__	__	.060	.060	.037	1.650	.124	-.019	.140	__	__	.0603; .0606
Session[a]	33	291	__	__	__	__	__	__	__	__	.028	52.786	__
T1 = T2 = T3 = 0	__	__	1.02	__	__	__	__	.462	__	__	__	__	__
T1 vs T2 vs T3	__	__	.930	__	__	__	__	.485	__	__	__	__	__
T1	__	__	__	.057	.057	.035	1.629	.115	-.015	.130	__	__	.0574; .0576
T2	__	__	__	.091	.090	.076	1.193	.282	-.099	.280	__	__	.0904; .0908
T3	__	__	__	-.020	-.02	.048	-.397	.732	-.241	.203	__	__	-.0189
# of Trials	35	323	__	__	__	.000	.062	.953	.000	.000	.029	52.591	__
Type of SRTT	35	323	__	__	__	__	__	__	__	__	.031	53.977	__
Det = Prob = Alt = 0	__	__	3.09	__	__	__	__	.213	__	__	__	__	__

Det vs Prob vs Alt	__	__	.494	__	__	__	__	.677	__	__	__	__	__
Det	__	__	__	.067	.067	.038	1.788	.086	-.010	.145	__	__	.0673; .0677
Prob	__	__	__	.035	.035	.053	.661	.549	-.120	.190	__	__	.0351; .0357
Alt	__	__	__	.021	.021	.006	3.596	.173	-.120	.190	__	__	.0206; .0207

Note. s = number of studies; k = number of effect size estimates; F values are from Approximate Hotelling-Zhang with small sample correction omnibus tests of the effects of moderators with more than two levels; r = Pearson's R correlation; standard errors (SE) and t values for individual levels of a moderator; *p* values correspond to F or t - values; 95 % CI corresponds to the Fisher's z; [a] Studies by Clark and Lum (2017b) and Desmottes et al. (2017) were removed from the meta-regression with session as a moderator variable as these studies did not compute correlations independently for each session.

Table 4.2

Results of all meta-regressions for language and literacy components with group as a moderator variable

Moderators (bolded) and levels	***Study Characteristics***			***Effect size***			***Test of significance***					***Heterogeneity***		***Sensitivity analysis***
	s	k	F	Fisher's z	r	R	SE	t	p	95% CI		τ^2	I^2	range
Grammar	25	137	__	__	__	__	__	__	__	__	__	0.023	44.432	__
TD = DLD = DD = 0	__	__	1.400	__	__	__	__	__	.394	__	__	__	__	__
TD vs DLD	__	__	0.210	__	__	__	__	__	.722	__	__	__	__	__
GroupDD	__	__	__	0.085	.085	.185	0.039	2.180	.274	-0.410	0.580	__	__	0.0849; 0.0851
GroupDLD	__	__	__	-0.017	-.017	-.037	0.060	-0.283	.782	-0.147	0.113	__	__	-0.0169; -0.0168
GroupTD	__	__	__	0.057	.057	.124	0.044	1.298	.214	-0.037	0.151	__	__	0.0572; 0.0574
Vocabulary[a]	18	63	__	__	__	__	__	__	__	__	__	0.017	35.482	__
TD = DLD = 0	__	__	1.300	__	__	__	__	__	.311	__	__	__	__	__
TD vs DLD	__	__	1.490	__	__	__	__	__	.243	__	__	__	__	__
GroupDLD	__	__	__	0.014	.014	.031	0.036	0.389	.706	-0.067	0.096	__	__	0.0141; 0.0142
GroupTD	__	__	__	0.089	.089	.194	0.054	1.665	.129	-0.031	0.210	__	__	0.0892; 0.0895

Phonology	16	111	__	__	__	__	__	__	__	__	__	0.028	42.235	__
TD = DLD = DD = 0	__	__	0.778	__	__	__	__	__	.539	__	__	__	__	__
TD vs DLD	__	__	2.440	__	__	__	__	__	.153	__	__	__	__	__
TD vs DD	__	__	0.628	__	__	__	__	__	.451	__	__	__	__	__
GroupDD	__	__	__	0.047	.047	.103	0.065	0.725	.498	-0.116	0.211	__	__	__
GroupDLD	__	__	__	0.212	.209	.456	0.138	1.536	.192	-0.155	0.579	__	__	__
GroupTD	__	__	__	-0.018	-.017	-.037	0.041	-0.427	.678	-0.109	0.074	__	__	__
Reading	15	130	__	__	__	__	__	__	__	__	__	0.004	12.981	__
TD = DLD = DD = 0	__	__	2.060	__	__	__	__	__	.213	__	__	__	__	__
TD vs DD	__	__	0.003	__	__	__	__	__	.956	__	__	__	__	__
GroupDD	__	__	__	0.040	.040	.087	0.077	0.518	.625	-0.154	0.233	__	__	0.0392; 0.0426
GroupDLD	__	__	__	0.100[c]	.100	.218	0.043	2.305	.114	-0.047	0.247	__	__	0.1002; 0.0995
GroupTD	__	__	__	0.045	.045	.098	0.039	1.153	.287	-0.048	0.139	__	__	0.0451; 0.0473
Spelling[b]	5	12	__	__	__	__	__	__	__	__	__	0.017	34.435	__
TD = DD =0	__	__	1.600	__	__	__	__	__	.388	__	__	__	__	__
TD vs DD	__	__	0.896	__	__	__	__	__	.421	__	__	__	__	__

GroupDD	__	__	__	0.252	.247	.539	0.159	1.588	.263	-0.489	0.993	__	__	0.2517; 0.2527
GroupTD	__	__	__	0.047[c]	.047	.103	0.088	0.537	.627	-0.226	0.320	__	__	0.0473; 0.0470

Note. s = number of studies; k = number of effect size estimates; F values are from Approximate Hotelling-Zhang with small sample correction omnibus tests of the effects of moderators with more than two levels; r = Pearson's R correlation; R = disattenuated Pearson's R correlation; standard errors (SE) and t values for individual levels of a moderator; p values correspond to F - or t - values; 95 % CI corresponds to the Fisher's z; [a] The DD (k = 2) and [b] DLD (k = 1) groups were not included in these analyses due to the small sample size.

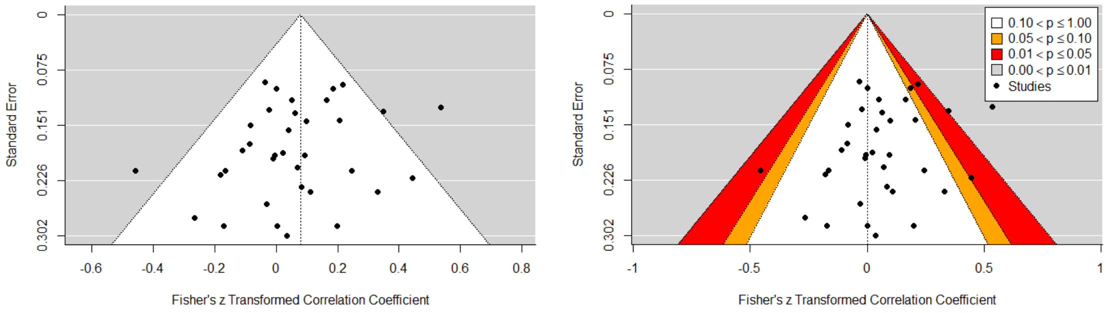
Publication bias

Several assessments of publication bias were conducted on the aggregated data, with the exception of PET and PEESE models which examined publication bias on all effect sizes. Funnel plot and contour plots for the estimates are shown in Figures 4.2A and 4.2B. Visual inspection shows no obvious evidence of plot asymmetry or overrepresentation of studies in the significance contours. These results are further supported by the non-significance of the rank correlation test (Kendall's τ = -.03, p = .880).

In line with previous results, both PET and PEESE models with RVE showed no evidence of publication bias (PETrve: *b*1 = -.739, *p* = .120; PEESErve: *b*1 = -1.560, *p* = .210).

Figure 4.2

Funnel plot showing study level effect sizes plotted against standard error. An asymmetric distribution is taken as evidence of publication bias; A: Funnel plot (left panel) and B: contour-enhanced funnel plot (right panel)



1.3 Literacy and procedural learning

Overall effect

A total of k = 18 studies were included in the analyses. Similar findings to those obtained for language were observed for literacy, both for the overall effect, and the meta-regressions examining moderators. Fisher's *z* for individual studies varied between -.48 to .91, with 57% positive effect sizes. The meta-analytical model revealed a significant, but again very modest, relationship between literacy and procedural learning, Fisher's *z* = .05 (Pearson's *r* = .05), 95% CI [.003, .10], SE = .02, *t*(13.5) = 2.29, *p* = .039, with evidence of a small amount of heterogeneity in the effect sizes (τ^2 = .011, I^2 = 29.48). Sensitivity analyses indicated that this result was consistent across rho values (values ranged from .0495 to .0496). Similarly to previous results, the weak relationship between literacy and procedural learning may reflect the inclusion of the disorder groups, thus it does not speak directly to our hypotheses.

Moderator analyses

All results from the meta-regressions are presented in Table 4.2-4.3.

Group membership was not a significant predictor of the magnitude of the relationship between literacy and procedural memory and there was no evidence that this relationship differed from zero for all groups (Table 4.3). To answer our research questions (H1/H2/H3), separate meta-regressions were conducted for reading and spelling to determine whether there were differences between groups for these subdomains. Given the low number of effect sizes for the DLD group for spelling (n = 1), the DLD group was omitted from these analyses, but was included in the remaining analyses. There was no evidence that the magnitude of the pooled effect size differed between the TD and dyslexic groups for spelling (*F*(1, 2.7) = .90, *p* = .421) and reading (*F*(1, 7.44) = .003, *p* = .956). Additionally, the relationship between procedural learning and spelling was not significant for any group (*F*(2, 1.95) = 1.60, *p* = .388). The same pattern was observed for reading (*F*(2, 7.64) = .99, *p* = .414). Again, there was no evidence supporting our hypothesis for a relationship between literacy and procedural memory in the TD group (H1), nor did the magnitude of the relationship between spelling and reading and procedural memory differed between the TD and dyslexic groups (H3). Yet, the disattenuated correlations (see Table 4.2) show a small relationship between spelling and procedural abilities for the TD group (*r* = .10), whilst for the DLD group there was a small association between reading and procedural abilities (R = .22). Of particular interest, the spelling abilities of the dyslexic group moderately correlated with procedural learning (R = .54).

For all sampling and methodological moderators, the magnitude of the relationship was not significantly different between categorical levels nor from zero, except for *session*. The relationship between literacy and procedural learning was moderated by *session* (*F*(2, 6.36) = 5.89, *p* = .036), with a higher correlation between these variables for the second session (Fisher's *z* = .164, 95% CI [.035, .294] SE = .045, *t*(3.67) = 3.65, *p* = .025) than for the first (Fisher's z = .023, 95% CI [-.029, .075], SE = .024, *t*(12.80) = .95, *p* = .361). The difference in the magnitude of the effect size estimate between sessions 1 and 2 was statistically significant (*F*(1, 5.13) = 8.40, *p* = .033), with a higher effect for session 2 than session 1.

Model diagnostics

Sensitivity analyses show that the findings did not differ depending on the value of rho (see Tables 4.2 and 4.3). The study by West, Shanks, et al. (2021) was identified as a potential influential effect size, with an aggregated effect size higher than expected. Given that outlier detection was conducted on the aggregated data, no further actions were taken.

Table 4.3

Results of all separate meta-regressions with moderator variables for literacy measures

Moderator (bolded) and level	***Study characteristics***			***Effect size***		***Test of significance***					***Heterogeneity***		***Sensitivity analysis***
	s	k	F	Fisher's z	r	SE	t	p	95 % CI		τ^2	I^2	range
Group	18	155	__	__	__	__	__	__	__	__	.013	32.056	__
TD = DLD = DD = 0	__	__	1.300	__	__	__	__	.347	__	__	__	__	__
TD vs DLD vs DD	__	__	0.127	__	__	__	__	.883	__	__	__	__	__
TD	__	__	__	.038	.037	.030	1.230	.249	-.031	.106	__	__	.0801; .0808
DLD	__	__	__	.054	.054	.069	0.786	.493	-.174	.281	__	__	.0532; .0539
DD	__	__	__	.080	.080	.065	1.237	.254	-.072	.232	__	__	.0374; .0377
Domain^1	16	143	__	__	__	__	__	__	__	__	.006	16.164	__
Reading = Spelling = 0	__	__	2.440	__	__	__	__	.165	__	__	__	__	__
Reading vs Spelling	__	__	0.854	__	__	__	__	.401	__	__	__	__	__
Reading	__	__	__	.048	.048	.027	1.800	.103	-.012	.108	__	__	.0481; .0497
Spelling	__	__	__	.120	.120	.074	1.620	.185	-.091	.331	__	__	.1196; .1203
Age	18	155	__	.001	.001	.004	0.154	.884	-.010	.011	.014	32.988	__
Sequence complexity	16	125	__	__	__	__	__	__	__	__	.018	37.434	__
FOC = SOC = 0	__	__	2.220	__	__	__	__	.177	__	__	__	__	__

FOC vs SOC	__	__	0.692	__	__	__	__	.427	__	__	__	__	__
FOC	__	__	__	.028	.028	.029	0.972	.383	-.050	.105	__	__	.0277; .0282
SOC	__	__	__	.069	.069	.036	1.886	.098	-.016	.153	__	__	.0686; .0688
Session[2]	18	152	__	__	__	__	__	__	__	__	.008	21.693	__
T1 = T2 = 0	__	__	5.890	__	__	__	__	**.036**	__	__	__	__	__
T1 vs T2	__	__	8.400	__	__	__	__	**.033**	__	__	__	__	__
T1	__	__	__	.023	.023	.024	0.947	.361	-.029	.075	__	__	__
T2	__	__	__	.164	.163	.045	3.646	**.025**	.035	.294	__	__	__
Number of trials	18	155	__	.000	.000	.000	1.118	.312	.000	.000	.011	28.483	4.43e-05; 4.23e-05
Type of SRTT	18	155	__	__	__	__	__	__	__	__	.016	35.618	__
Det = Prob= Alt = 0	__	__	0.967	__	__	__	__	.543	__	__	__	__	__
Det vs Prob vs Alt	__	__	0.291	__	__	__	__	.669	__	__	__	__	__
Deterministic	__	__	__	.046	.046	.023	2.041	.070	.005	.097	__	__	.0455; .0462
Probabilistic	__	__	__	.067	.067	.053	1.259	.308	-.117	.251	__	__	.0678; .0669
Alternating	__	__	__	.001	.001	.081	0.011	.993	-1.025	1.026	__	__	-.0004; .0012

Note. s = number of studies; k = number of effect size estimates; F values are from Approximate Hotelling-Zhang with small sample correction omnibus tests of the effects of moderators with more than two levels; r = Pearson's R correlation; standard errors (SE) and t values for individual levels of a moderator; p values correspond to F or Fisher's z values; 95 % CI corresponds to the Fisher's z; [1] The effect of *subdomain* was only analysed for spelling and reading; [2] Given the small number of studies for the third session, this level was omitted from the analyses as the parameters could not be estimated.

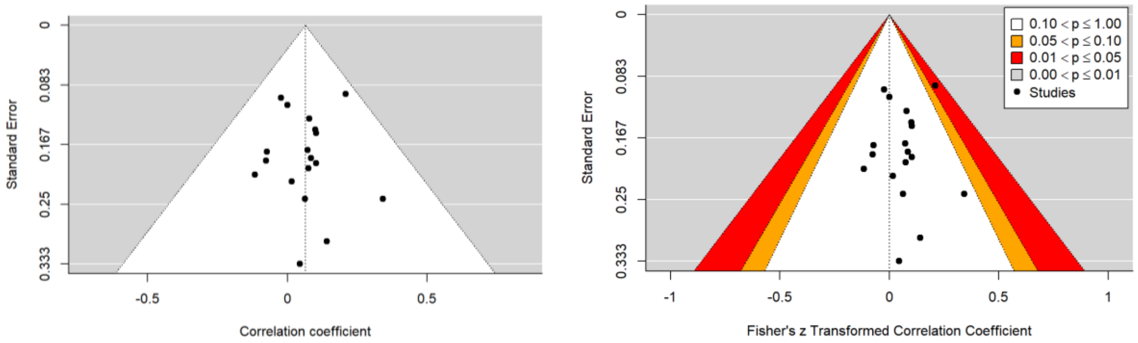
Publication bias

For publication bias, visual inspection of the funnel (shown in Figure 4.3A) and contour (is shown in Figure 4.3B) plots of the aggregated effect sizes revealed no evidence of asymmetry. This is consistent with the non-significant rank test for plot asymmetry (Kendall's τ = -.026, *p* = .880), thus suggesting low likelihood of publication bias.

PET and PEESE RVE models also showed no evidence of publication bias (PETrve: *b*1 = .02, *p* = .961; PEESErve: *b*1 = .17, *p* = .840).

Figure 4.3

Funnel plot showing study level effect sizes plotted against standard error. An asymmetric distribution is taken as evidence of publication bias; A: Funnel plot (left panel) and B: contour-enhanced funnel plot (right panel).



Discussion

Previous studies testing the procedural deficit hypothesis have primarily focussed on group-level testing, finding mixed results. Conversely, here we take a continuous approach to examining individual differences in literacy and language as predictors of procedural learning, in typical and literacy/language disordered populations, as predicted by the procedural/declarative model. This study comprised a large-scale meta-analysis which found no evidence of support for the procedural/declarative model. Counter to this hypothesis (and our predictions) but in keeping with recent smaller-scale meta-analyses, the results revealed only a negligible relationship between procedural learning and language and literacy for the overall sample. Neither association remained significant for the separate groups. Turning to the separate subdomains of language and literacy, as expected, vocabulary did not significantly correlate with procedural learning for any of the groups, nor were there differences in the pooled effect size between groups. However, procedural learning was also uncorrelated with grammar, phonology, reading and spelling in TD children, which is counter to our hypotheses and the predictions of the procedural/declarative model. Furthermore, the magnitude of the relationship between language and literacy and procedural learning did not differ between groups; specifically, the size of the effect did not differ between the TD and DLD groups for grammar and phonology, nor did it differ between TD and dyslexic groups for phonology, reading and spelling. Together, these results provide minimal to no support for the procedural/declarative model or the procedural deficit hypothesis.

Whilst the absence of correlations between language and literacy measures and procedural learning for the disordered groups may be taken as supportive evidence for the procedural deficit hypothesis (see Lum et al., 2012), this was observed alongside a similar pattern for the TD group. Thus, these results point to an overall lack of association between language, literacy and procedural learning as previously observed by Lammertink et al. (2020) and West, Melby-Lervåg, et al. (2021). Additionally, even though there was no evidence of differences between disordered groups on the magnitude of the relationship between language/literature and procedural learning, it is not clear from the procedural deficit hypothesis whether distinct patterns would be expected for the disordered groups when both groups are expected to show deficits in the same language/literacy domains, as is the case of phonology for dyslexia and DLD.

The findings from this meta-analysis replicate and extend the results of recent meta-analyses which found no association between grammar and procedural learning in children (Lammertink et al., 2017) and between language and decoding measures and procedural learning on the SRTT (West, Melby-Lervåg, et al., 2021). However, these findings are at odds with those obtained by Hamrick et al. (2018). The significant relationship between procedural learning and grammatical abilities observed by Hamrick et al. (2018) for first language learners may have been due to the small sample size, and inclusion of low powered studies which have been found to often upwardly bias the overall estimate (Loken & Gelman, 2017; Turner et al., 2013). Importantly, subsequent larger published studies reported a non-significant relationship between procedural learning and grammatical abilities (e.g., Llompart & Dąbrowska, 2020; West et al., 2018). There was no evidence of a moderating role of age in the present meta-analysis, thus suggesting that the strength of the relationship between procedural learning and language/literacy was not influenced by the age of the participants counter to Hamrick et al. (2018) but in line with the finding from Lammertink et al. (2020). However, only children older than 5 years old were included in the present meta-analysis. Thus, we cannot rule out that procedural learning may be more tightly associated with language and literacy acquisition at earlier stages of development. The association between procedural learning and grammar abilities in adult second language learners observed by Hamrick et al. (2018) may capture this early stage of language acquisition, when linguistic rule-based knowledge is accumulated and integrated into more abstract and complex grammatical structures. Thus, it will be important for future work to take a broader age perspective. Another possible explanation for the discrepancy between the present results and Hamrick et al. (2018) is that the latter focused on studies that used deterministic SRTTs, whereas the present review included both deterministic and probabilistic tasks; however, there were no differences in the effect sizes for these task variants here, ruling out this explanation.

Furthermore, exploratory meta-regressions on the whole sample were conducted to assess the impact of methodological differences on the magnitude of the relationship between language and literacy and procedural learning. There was no evidence that sequence complexity, number of trials and type of SRTT affected the association between language and literacy and procedural learning. However, the number of SRTT sessions was found to be a significant moderator of the relationship between procedural learning and literacy, such that the size of the effect was higher, even though still small, for procedural learning captured during a second session than for a first session. This is consistent with the suggestion by Conway et al. (2019), that correlations between literacy and procedural learning may emerge only for later sessions when procedural learning is more robust since knowledge and skill acquisition in this memory system tends to be gradual and require multiple exposures to the stimuli. This suggests that procedural learning that takes place after multiple training sessions may provide a more reliable predictor of individual differences than after a single session.

Importantly, the present findings suggest that while the SRTT can be adequately used for examining group differences (Hedge et al., 2018), this task may not provide a reliable measure of individual differences (Hedge et al., 2018). In line with this, in their meta-analysis, West, Melby-Lervåg, et al. (2021) found group differences between individuals with dyslexia and DLD and the TD group on procedural learning across measures (SRTT, Hebb learning task, artificial grammar and statistical learning tasks, weather prediction task). However, they found a non-significant relationship between continuous measures of language-and procedural learning. Whilst a reliable measure of individual differences requires considerable inter-individual variance allowing for the ranking of individuals, a sensitive measure of group differences does not. This interpretation concurs with recent evidence demonstrating the poor test-retest reliability of the procedural learning scores obtained with the SRTT (e.g., Kalra et al., 2019; Siegelman & Frost, 2015; Stark-Inbar et al., 2017; West et al., 2018; West, Shanks, et al., 2021). Given this, one must consider that the weak/absent correlations observed here may not necessarily refute the procedural/declarative model and the procedural deficit hypothesis, but rather, the SRTT may be insufficiently sensitive to individual differences to provide an adequate test of these hypotheses.

Test-retest reliability refers to the measure's consistency in ordering participants' performance at different time points (Kottner & Streiner, 2011). Measurement error and low variance between individuals have been suggested to decrease reliability (e.g., Fleiss, 1986; Hedge et al., 2018). Thus, the reliability of each measure will inform how much the raw correlations have been attenuated (e.g., Fleiss, 1986; Rouder et al., 2019; Spearman, 1904). The issue of attenuation has long been discussed (Spearman, 1904), yet, despite good progress, correlation recovery is still suboptimal as the methods available, whilst less biassed than raw correlations, still produce highly variable estimates (Rouder et al., 2019). One such method was proposed by Spearman (1904) and it proposes that disattenuation of a correlation between two measures can be accomplished by taking into account the reliability of each measure. We took this approach to estimate disattenuated correlations between language and literacy and procedural learning, but this did not change the pattern of results; correlations remained very low except for the correlations between procedural learning and phonology in DLD and spelling in dyslexia. Crucially, whilst disattenuated correlations may provide a better understanding of the true correlations, the Spearman (1904) method has been found to produce highly variable estimates thus these should be interpreted with caution. This is of special relevance given the pattern for low correlations for the TD group for which the procedural/declarative model makes clear predictions. Importantly, these results do not appear to be explained by publication bias. In summary, there appears to be some support for a procedural memory impairment in individuals with dyslexia and DLD in line with the procedural deficit hypothesis, as indicated by group-level studies. However, in the absence of evidence for a relationship between language and literacy and procedural memory measured continuously, it is still unclear whether procedural memory underlies the development of language and literacy and thus an impairment in procedural memory may represent a risk factor for dyslexia and DLD in line with the multiple deficit model (Pennington, 2006), with this relationship not being captured due to methodological limitations.

Whilst the SRTT has been shown to engage similar brain regions motor skill learning (Keele et al., 2003; Pascual-Leone et al., 1996; Robertson et al., 2001; Torriero et al., 2004), little is known about whether the abilities required to perform the SRTT translate into real-world procedural learning abilities (Mathews, 1997) and indeed whether quantitative differences in procedural learning on the SRTT are meaningful. A related issue is the lack of correlation between different tasks purporting to measure procedural learning, even when task demands have been carefully matched (Erickson et al., 2016). While this could again be explained by poor psychometric properties (Arnon, 2020; Siegelman & Frost, 2015; West et al., 2018; West, Shanks, et al., 2021), it is also possible that procedural learning is not a unified ability that can be similarly captured by all these tasks (Bogaerts et al., 2021). Thus, it may be that some measures of procedural learning are more relevant to the acquisition of language and literacy than others. This may explain the significant relationship between artificial grammar and statistical learning tasks and language-related abilities found by West, Melby-Lervåg, et al. (2021). This challenges the view of procedural learning as a general capacity that underlies the acquisition of all probabilistic knowledge irrespective of modality and domain (Conway et al., 2019). As suggested by Bogaerts et al. (2021) and Siegelman, Bogaerts, Christiansen, et al. (2017), future empirical work should focus on understanding the computations underlying procedural learning acquisition in each task so that these can be better mapped onto linguistic abilities.

Finally, the absence of evidence for a relationship between procedural learning on the SRTT and language/literacy also raises the possibility that group differences on the SRTT may be unrelated to language and literacy skills. The SRTT is not a pure measure of procedural learning and has been shown to rely on attention and working memory (Arciuli, 2017; Sengottuvel & Rao, 2013a; D. R. Shanks & St. John, 1994; West, Shanks, et al., 2021). Thus, in light of the evidence that individuals with dyslexia and DLD often have weaknesses in executive function (DLD: Marini et al., 2020; dyslexia: Romani et al., 2011; M. J. Snowling et al., 2020) and working memory (DLD: e.g., Baird et al., 2010; dyslexia: e.g., Fostick & Revah, 2018) group differences may actually reflect differences in other cognitive skills.

The present meta-analysis provides the most comprehensive overview to date of the relationship between procedural memory and language and literacy across children and adults with and without language and literacy disorders. The results provide little evidence for a relationship between continuous measures of language and literacy and procedural learning as indexed by the SRTT, thus calling into question the validity of the procedural/declarative model and procedural deficit hypothesis as a framework for understanding language acquisition. However, future research is needed to ascertain and improve the psychometric properties of the SRTT before this theoretical framework can be robustly tested. An important step will be for future research to adopt the practice of reporting test-retest reliability, allowing researchers to analyse the impact of reliability on their outcomes of interest (Parsons et al., 2019). Additionally, more sophisticated models such as meta-analytic structural equation modelling may be better suited for assessing the relationship between language and literacy and procedural learning as latent variables, whilst taking measurement error into account. Such a meta-analysis should ideally include measures of procedural learning from multiple tasks that tap into different abilities across subdomains; as well as potential confounding variables such as attention and memory. This model would have the potential to shed light on the moderating and mediating effects of procedural learning on language and literacy in children and adults with and without neurodevelopmental disorders. Such future research endeavours will be important in advancing our understanding of procedural memory, and its putative role in language and literacy acquisition, with potential for informing practice and intervention.

Chapter 5. Procedural Learning in The Serial Reaction Time Task in Individuals with and Without Dyslexia: Group-Level And Individual Differences

Abstract

An impairment in the procedural memory system, responsible for the extraction and assimilation of regularities from sensory input, has been suggested to play a causal role in neurodevelopmental disorders such as dyslexia. In this study, we compared the performance of adults with and without dyslexia on the Serial Reaction Time task across three sessions. Alongside it, this experiment presents the first examination of the stability of procedural learning in a dyslexic population. Across sessions, there was no evidence of an impairment in the dyslexic groups, thus lending no support to the procedural deficit hypothesis. There was limited evidence for a relationship between language/literacy abilities and procedural learning, with only a significant association between nonword repetition and procedural learning for the dyslexic group (*r* = .31). However, given that the reliability of the SRTT was well-below adequate psychometric standards in both groups, the lack of correlations between these measures may reflect attenuation. Nonetheless, a positive correlation between attention and procedural learning was observed in both groups, suggesting that individuals with better attentional abilities showed more evidence of procedural learning in the SRTT (*r*s between -.34 to -.47). In light of these findings, it is unclear whether a procedural learning impairment can account for the behavioural and cognitive profile observed in individuals with dyslexia.

Introduction

The ability to extract patterns from sensory input across time and space is thought to underlie the development and acquisition of language and literacy skills (Batterink et al., 2019; Saffran, 2018; Ullman et al., 2020). For instance, procedural learning is a core neurocognitive system implicated in the acquisition, consolidation and automatization of cognitive and motor skills and habits (L. R. Squire, 1984, 1994; L. R. Squire & Zola, 1996). This memory system has been argued to be important for language and literacy acquisition (Bitan & Karni, 2004; Bogaerts et al., 2021; Spencer et al., 2015), and it has been suggested that dyslexia, a neurodevelopmental disorder characterised by reading and spelling impairments in around 5% to 10% of the population (M. J. Snowling et al., 2020), is at least partly caused by a neural abnormality in this memory system (Ullman et al., 2020). However, despite meta-analyses lending some support to this theory (Lum et al., 2013; West, Melby-Lervåg, et al., 2021), prior studies comparing individuals with dyslexia and typically developing controls on procedural memory tasks have found mixed results. Inconsistent evidence has also been observed when examining the correlations between literacy and procedural memory (Hamrick et al., 2018; Lammertink, Boersma, Rispens, et al., 2020; West, Melby-Lervåg, et al., 2021). Whilst the absence of an association between procedural learning and literacy abilities may indicate that these constructs are unrelated, the poor psychometric properties of the tasks used to measure procedural learning may have contributed to the attenuation of the correlations between these variables (Kalra et al., 2019; Oliveira et al., submitted; West et al., 2018; West, Shanks, et al., 2021). Indeed, in order to adequately test the predictions of the procedural deficit hypothesis, reliable indices of procedural memory are required. The psychometric properties of these tasks have recently begun to be examined in typically developing populations, but to the best of our knowledge, no studies have yet investigated the stability of the procedural learning effect in disordered populations. In our previous work with a general population sample (Oliveira et al., submitted; Chapter 2 of this thesis), we found suggestive evidence that procedural learning effects may stabilise after multiple sessions, in line with the suggestion by Conway et al. (2019) and Palmer et al. (2018). Here, we extend existing evidence on the psychometric properties of the SRTT by focusing on the test-retest reliability of this task across multiple sessions in populations of adults with and without dyslexia. With the aim of testing the procedural deficit hypothesis, we examine both group differences in procedural learning and associations between SRTT performance and measures of language and literacy and how these results manifested over different sessions of procedural learning. Finally, we explore a putative link between procedural learning and attention in individuals with and without dyslexia as an effort to better understand the locus of procedural learning differences in dyslexia, and whether this is linked to the stability of the task in dyslexia and TD development.

The procedural/declarative model (Ullman, 2016a) is one of the models that aims to explain the role of procedural memory in language acquisition. This model claims that the declarative memory system is responsible for the fast acquisition of arbitrary information such as lexical knowledge, whilst the procedural memory system is involved in the more gradual acquisition of probabilistic information (including phonology, grammar, reading and spelling abilities). Focusing on literacy more specifically, children may implicitly learn and automatize the mapping between grapheme and phoneme in spelling (Gingras & Sénéchal, 2019; Treiman & Kessler, 2011) and reading (Steacy et al., 2019; Treiman, Kessler, & Bick, 2003; Treiman, Kessler, Zevin, Bick, & Davis, 2006) using statistical information in the input. This may be especially relevant in more opaque orthographies where the correspondence between grapheme and phoneme is less straightforward (Arciuli, 2018) and explicit instruction of all regularities may not be feasible (Apfelbaum et al., 2013), for example, children may draw on the frequency and position of double consonants (Pacton et al., 2001), and in consonantal and vowel context (Treiman & Kessler, 2006).

Evidence for the role of procedural memory in language learning is multifaceted. Firstly, procedural memory and language share brain systems such as basal ganglia and frontal cortex, such that abnormalities in the procedural memory system are expected to lead to linguistic impairments (Ullman & Pierpont, 2005). This association has been observed in clinical populations with damage to the basal ganglia, such as in Parkinson’s and Huntington’s disease, which often show simultaneous motor and linguistic impairments (Ullman & Pierpont, 2005). An additional, key line of evidence comes from behavioural experiments showing associations between tasks indexing procedural learning (including the SRTT, artificial grammar learning, and statistical learning tasks) and measures of language and literacy (Hamrick et al., 2018; West, Melby-Lervåg, et al., 2021). Based on this evidence, Ullman and colleagues (Ullman, 2014; Ullman et al., 2020) proposed the procedural deficit hypothesis, arguing that an impairment to the procedural memory system could underlie the language and/or literacy impairments that characterise dyslexia. dyslexia is a heterogeneous neurodevelopmental disorder, in which the defining difficulties with word-level reading and spelling are often accompanied by an array of secondary impairments in working memory (e.g., Fostick & Revah, 2018), motor functioning (Ramus, Pidgeon, & Frith, 2003), executive function and attentional control (e.g., Romani et al., 2011). The dominant theoretical account of dyslexia - the phonological deficit hypothesis - proposes a core deficit in the representation, storage and/or retrieval of speech sounds (Ramus, 2003). This hypothesis has had considerable success in accounting for the literacy difficulties in dyslexia as evidenced by the meta-analysis conducted by Melby-Lervåg et al. (2012) which demonstrated a large deficit in phonemic awareness in children with dyslexia when compared to TD children (pooled effect size estimate: -1.37) and reading-matched children (pooled effect size estimate: -0.57). Furthermore, intervention studies targeting phonological awareness have led to significant improvements in literacy abilities (Hulme et al., 2012), however this hypothesis cannot satisfactorily address the above-mentioned secondary impairments (M. J. Snowling et al., 2020)

The procedural deficit hypothesis attempts to address this gap, suggesting that the location, severity, and extent of the neuroanatomical abnormalities in the procedural memory system gives rise to multiple profiles, substantial heterogeneity across individuals with dyslexia and secondary difficulties that could emerge from dysfunctions in distinct areas of the procedural memory system and its circuitry. Specifically, working memory, and other executive functions, are thought to be supported by the posterior parietal-anterior dorsal striatum (anterior caudate/putamen) - prefrontal circuitry (Draganski et al., 2008; Ullman et al., 2020). According to this hypothesis (Ullman et al., 2020), reading difficulties may be explained by the procedural dysfunction, possibly leading to perceptual impairments and difficulties in learning predictive associations that underlie the learning of grapheme-phoneme (and phoneme-grapheme) mappings. Additionally, problems with speech-sound representations and learning of speech-sound categories may also occur since these skills are thought to be acquired implicitly, through gradual exposure to distributions of phonetic features. Finally, the procedural deficit hypothesis proposes that individuals with dyslexia may be able to compensate for some, if not all, of their procedural deficits by relying on the declarative memory system, using strategies such as chunking, as well as drawing on semantic and explicit knowledge (Ullman & Pullman, 2015).

Supporting the procedural deficit hypothesis, a meta-analysis conducted by Lum et al. (2013) found evidence for smaller procedural learning effects on the SRTT (Nissen & Bullemer, 1987) for individuals with dyslexia (N = 186, aged 7-15 years) compared to typically developing controls (N = 203, aged 7-15 years; average weighted standardised mean difference of .499, 95% CI [.204, .693]). In the SRTT, one the of the most commonly used measures of procedural memory, four rectangles are typically displayed on screen with a stimulus (e.g., a smiley face) appearing in one of the four positions on each trial. Participants are asked to respond to the position of the stimulus as quickly as possible by pressing the corresponding key on the keyboard. Unbeknownst to the participants, the presentation of some of the stimuli follows a pattern. Evidence of procedural learning in this task is taken as the difference between RTs for trials that follow the sequence and those which do not, on the assumption that if participants learn the sequence, they should be able to anticipate the position of the upcoming stimulus and respond faster. Additionally, Lum et al. (2013) observed that the size of the difference between groups diminished with age when participants had more practice on the SRTT or when they were tested on a SRTT with a second order conditional sequence (i.e., higher order sequences involve determining which stimulus (n) will follow is based on element n-2). The authors argued that the better performance of older participants with dyslexia could be due to the compensatory role of declarative memory, since as suggested by Lum et al. (2013) the age effect only emerged when using second order conditional sequences thought to place increased demands on the declarative memory system. This group-level difference was replicated by West, Melby-Lervåg, et al. (2021) in a more recent meta-analysis that included 610 participants with developmental language disorder and dyslexia and 698 typically developing participants. In this study, an effect size of *g* = .-28, 95% CI [-.46, -.09] revealed worse performance on the SRTT for participants with dyslexia. Counter to Lum et al. (2013), despite observing a larger numerical group difference for adults relative to children, the age effect was not statistically significant. It is worth noting, however, that although these meta-analyses show an overall deficit in SRTT performance in dyslexic groups, the results for individual studies have been variable and have not always replicated this finding (e.g., null group effects on the SRTT reported by Deroost et al. (2010), Gabay et al. (2012a), Henderson and Warmington (2017) and Rüsseler et al. (2006)).

Crucially, these meta-analyses, and the majority of studies, have only examined group differences in a single practice session. It is unclear whether performance in a single session is sufficient to probe the different stages of procedural learning. Procedural learning is thought to involve an initial phase of fast acquisition after exposure (Doyon et al., 2009), followed by a second stage marked by a slower learning rate, trending towards an asymptote which is thought to be required for consolidation to occur (Hauptmann & Karni, 2002; Hauptmann et al., 2005; Karni et al., 1998). As the initial memory becomes increasingly robust and resistant to interference through memory consolidation (Hauptmann et al., 2005; Nemeth et al., 2010; Song, 2009; Song, Howard, & Howard, 2007), automatisation of the new learned skill may allow for its effortless performance - often referred to as 'proceduralisation' (Doyon & Benali, 2005). Whilst these studies have often used the same sequences in the first and follow-up sessions, our previous findings (Oliveira et al., submitted) suggest that this may not be a requirement for consolidation and automatisation to occur as we observed a positive relationship between the degree of similarity between sequences used at test and retest and the procedural learning effect in the second session. Thus, suggesting that some knowledge is carried on from session to session, however, with this design, the distinct stages of procedural learning cannot be isolated. Despite this disadvantage, from a practical perspective, the adoption of distinct sequences with shared elements arguably more likely represents the input children would receive in a naturalistic environment. Given that the earlier phase of fast acquisition may lead to substantial changes in the ranking order between test and retest due to individual differences in the learning trajectory, it is likely that once participants' gains start showing less practice effects in the procedural learning effect in later sessions that this will translate into better reliability. It is then that the associations between measures of procedural learning and language or literacy should be most clearly apparent not only because attenuation would occur to a lesser degree but also because this stage would likely reflect more closely the individuals' true learning ability. From a theoretical perspective, as suggested by Hedenius et al. (2013, 2021), it is unclear whether individuals with dyslexia are slower learners and thus require more practice to reach saturation (e.g., Hauptmann et al., 2005) or whether irrespective of the amount of practice they continue to show impaired procedural learning when compared to TD individuals. A small number of studies have examined the issue of the time course of procedural learning in participants with and without dyslexia. Unlike the findings from the meta-analyses described above, Hedenius et al. (2013, 2021), who used the same sequences at test and retest, only observed differences in procedural learning between dyslexic and TD children in the alternating SRTT at their 24-hour follow-up sessions, but not in the initial sessions. On the other hand, the studies by Gabay et al. (2012) (day 1 and 24h later) and Henderson and Warmington (2017) (days 1, 2 and 8) which compared the procedural learning abilities of adults with and without dyslexia on a deterministic SRTT, did not observe differences between groups in any of the sessions. Due to the limited number of studies examining later sessions, we can only speculate that these differences may reflect developmental changes. In line with Lum et al. (2013) and West, Melby-Lervåg, et al. (2021), who reported larger group differences for children than adults, it is possible that this age difference would also be observed in later sessions, whereby adults are better able to compensate for their difficulties than children for whom the declarative memory system is still under maturation (Finn et al., 2016). However, it is possible that extraneous variables with similar maturational trajectories such as attention may account for these findings.

While the procedural deficit hypothesis makes clear predictions of a group-level deficit in measures of procedural learning in populations with dyslexia and other language-learning disorders, the model's predictions are more complex when taking an individual differences perspective. This is because of the proposed compensatory role of declarative memory. Specifically, the procedural/declarative model (Lum et al., 2012) expects procedural memory to correlate with language and literacy in typically developing populations. However, if the procedural memory is impaired in dyslexia (and not supporting language/literacy abilities to the same extent) then a null correlation between procedural learning and language/literacy may be predicted. Instead, stronger associations might be predicted between the compensatory declarative memory system and language/literacy abilities. However, the literature remains unclear on the expected pattern of correlations: Whilst there is some evidence of correlations between literacy (or literacy related abilities) and procedural learning and/or consolidation (TD group: ; dyslexic group: Jiménez-Fernández et al., 2011; Vakil et al., 2015), other studies have failed to find such correlations in children and adults with and without dyslexia (TD group: Desmottes, Meulemans, et al., 2017; Henderson & Warmington, 2017; Vakil et al., 2015; dyslexic group: Deroost et al., 2010; Henderson & Warmington, 2017; Vakil et al., 2015). In recent meta-analyses (Oliveira, Hayiou-Thomas et al., in prep; West, Melby-Lervåg, et al., 2021), there was no evidence for a relationship between literacy measures and procedural learning on the SRTT across groups with and without neurodevelopmental disorders. Nor was there any evidence that the magnitude of the association differed between TD and dyslexic groups. However, disattenuated correlations (i.e., adjusted for reliability) in Oliveira, Hayiou-Thomas et al. (in prep, Chapter 4 of this thesis) revealed a large correlation between procedural learning and spelling (r = .53) for the dyslexic group, with only negligible to small correlations between language/literacy and procedural learning for the TD group. Reconciling these contradictory findings under the procedural deficit hypothesis is difficult.

One factor that may help to account for the mixed findings in this field is the psychometric properties of the procedural learning effect that is measured by the SRTT, particularly its test-retest reliably. Similar to other well-established paradigms within cognitive psychology (Hedge et al., 2018), the SRTT yields robust and replicable results at the group level, but recent investigations demonstrate that it has suboptimal psychometric properties, which make it difficult to apply to individual differences research (Oliveira et al., submitted; West et al., 2018; West, Shanks, et al., 2021). Even though, split-half reliability, reflecting the consistency in individual's procedural learning scores within a testing session, has been shown to vary between moderate and adequate (.49 - .92; (Lammertink, Boersma, Rispens, et al., 2020; West et al., 2018; West, Shanks, et al., 2021), test-retest reliability has been consistently shown to fall below acceptable psychometric standards (i.e., r <.70; Burlingame et al., 1995; Nunnally & Bernstein, 1994) ranging from .21 to .47 (Kalra et al., 2019; Siegelman & Frost, 2015; Stark-Inbar et al., 2017; West et al., 2018; West, Shanks, et al., 2021; with the exception of .71 reported in West et al., 2021). Test-retest reliability, unlike split-half reliability, examines the consistency of the scores obtained by participants across sessions; thus, changes to the ranking order of participants between test and retest will be translated into poor reliability (Berchtold, 2016; Nunnally & Bernstein, 1994). Since observed scores are thought to reflect true scores and measurement error (Fleiss, 1986; Novick, 1966), high degrees of measurement error may lead to fluctuations in the observed scores across time, resulting in attenuation of the correlation between measures of interest (although note that in smaller samples the opposite may occur due to chance, Loken & Gelman, 2017), as well as greater uncertainty in parameter estimation (Loken & Gelman, 2017).

Additionally, measurement error may reflect the impure nature of the SRTT (and other procedural memory tasks, see Arnon, 2020; Kalra et al., 2019; Siegelman & Frost, 2015; West et al., 2018). For example, while procedural learning on the SRTT has been found to be to a certain extent independent of working memory (Janacsek & Nemeth, 2013; Unsworth & Engle, 2005) (cf. Bo et al., 2011; Medimorec et al., 2021) and intelligence (Danner et al., 2017; Feldman et al., 1995; A. S. Reber et al., 1991), accumulating evidence suggests a role for explicit awareness (Shanks et al., 2005; Stark-Inbar et al., 2017) and attention. Specifically, impaired procedural learning has been observed in dual-task paradigms (Röttger et al., 2019; Thomas et al., 2004), while studies incorporating independent measures of attention have found a positive relationship with procedural learning (Sengottuvel & Rao, 2013a; West, Shanks, et al., 2021). Conversely, mind-wandering measured by experience-sampling thought probes during the SRTT has been shown to have a negative association with procedural learning (Franklin et al., 2016). However, the role of attention in procedural learning is still under debate. There is evidence suggesting that attention is a mechanism which, through top-down (i.e., goal-driven) and bottom-up (i.e., stimulus-driven) processes (Awh et al., 2012; Failing & Theeuwes, 2018), reduces the available information by selecting the relevant stimuli and allowing for the detection and extraction of regularities from the sensory input (van Moorselaar & Slagter, 2019; J. Zhao & Luo, 2017). Thus, it is possible that attention may be required for establishing associations between temporally separated events, particularly when learning more complex sequences (Jimenez & Mendez, 1999; Cohen et al., 1990). Thus, whilst procedural learning is often being characterised as an automatic process independent of attentional resources, this assertion may be stage dependent: in a neuroimaging study (Thomas et al., 2004) lower activation in the parietal cortex, associated with visual-attentional processes, was observed for sequenced than random blocks in adult participants performing a deterministic SRTT. This finding was taken to indicate that as learning becomes automatised, it becomes less dependent on visual attention. Similar results were reported by Lum et al. (2019), where they examined behavioural and electrophysiological changes whilst performing a deterministic SRTT, focussing on three components, two of these likely to capture attentional processes - P1 and N1 and P3 which is typically observed in oddball versions of the SRTT for deviant trials. Even though P1 and N1 are both thought to capture attention processes, evidence suggests that P1 is modulated by attentional processes related to visuospatial processing whilst N1 is modulated by evaluation of a visual target (Vogel & Luck, 2000; Warbrick et al., 2014). When performing the SRTT, the amplitude of the P1 component decreased across sequenced trials followed by an increase in the random block, thus mirroring the procedural learning effect. No such pattern was observed in other components such as N1 and P3, thus suggesting a decrease in the reliance on visuospatial processing over time. To conclude, since attention may be involved in the early stages of procedural learning, it is possible that the weaknesses in attention often observed in populations with dyslexia (Ebert & Kohnert, 2011; Paulesu et al., 2014) might not only contribute to the group differences in procedural learning, but also to poorer stability of the SRTT in disordered populations as fluctuations in attention throughout the task and from day to day might contribute to changes in the procedural learning effect.

Rationale for the current study

Given the recognition that adequate psychometric properties are a requirement for reliable, and potentially replicable, results (Matheson, 2019), our previous efforts have been aimed towards establishing the test-retest reliability of the probabilistic SRTT in typically developing adult populations and to investigate which factors may affect its stability. In Oliveira et al. (submitted) we observed that increasing the number of sessions led to numerically higher stability on the SRTT (session 1-2: r = .43; session 2-3: r = .60), especially when using random slopes as opposed to simple difference scores. The reliability for later sessions may point to the superior stability of later stages of procedural memory, however given the use of different sequences at each time point, it is likely that even at follow-up sessions earlier stages were still being captured. Finally, when testing the assumptions of the procedural deficit hypothesis (Ullman, 2004), i.e., that language and literacy skills would correlate with procedural learning, no support was found. Procedural learning was only found to moderately correlate with vocabulary in the third, and last, session, but not with measures of literacy, nonword repetition or sentence repetition. Attention, on the other hand, showed a moderate relationship with procedural learning, with individuals with better offline attentional abilities showing more evidence of procedural learning on the SRTT, this is consistent with previous findings (Sengottuvel & Rao, 2013a; West, Shanks, et al., 2021). The relationship between attention and procedural learning may potentially indicate a reciprocal interaction between these variables, whereby attention allows for the reduction in perceptual noise (Gottlieb, 2012) and the extraction of the probabilistic by attending to relevant stimuli, but there’s also evidence that once regularities are detected, these may facilitate performance by reducing the amount of attention captured by distractors (Wang & Theeuwes, 2018). Thus, this leads to the hypothesis that weaknesses in attention often observed in dyslexia (Ebert & Kohnert, 2011; Paulesu et al., 2014) may place these populations at increased risk for procedural learning deficits, not only because individuals with poorer attentional abilities may be less able to attend to the relevant stimuli to support procedural learning (van Moorselaar & Slagter, 2019; J. Zhao & Luo, 2017), but also because due to the repetitive and long duration of the SRTT these individuals are more likely to show declines in attentional resources across the session which may lead to task disengagement (Fortenbaugh et al., 2017). It is also possible that poor attention may increase the variability in performance throughout the task, contributing to increased noise in the procedural learning effect, and thus reduced stability. Thus, the current study aims to determine whether attention (measured via a Psychomotor Vigilance task) is associated with both procedural learning performance and reliability.

In summary, this study examined the cognitive and procedural learning abilities of adults with dyslexia, with the aim of testing the predictions of the procedural deficit hypothesis by examining group and individual differences in procedural learning and its role in language and literacy. Importantly, these group and individual differences were considered in the context of also examining the stability of performance on the probabilistic SRTT over three sessions. Even though we have adopted different sequences at each time point, these were highly similar sequences, therefore at least some degree of procedural knowledge will be carried across sessions as suggested by our previous findings where similarity positively correlated with procedural learning in session 2 (Oliveira et al., submitted). Thus, it is likely that this experiment will capture both earlier and later stages of procedural learning in the follow-up sessions, however it will not be possible to disentangle learning from later stages as these will likely co-occur. We suggest that the greater stability expected in later sessions will result in more robust associations with measures of language and literacy and provide a more sensitive test of the predictions of the procedural deficit hypothesis. Furthermore, we aim to investigate the association between procedural learning and attention that has previously been reported in the general population. Given the substantial variability in sustained attention observed in dyslexia, we expect that the association with the SRTT will be particularly marked in the current study, yet if the poorer attentional abilities of this group lead to a less reliable procedural learning effect, then it could result in more attenuation of the associations between these measures.

PRE-REGISTERED HYPOTHESES

1. Are there group differences in procedural learning?
    a. Participants are expected to demonstrate evidence of procedural learning in all sessions. This will be indicated by faster and more accurate responses to probable versus improbable trials, particularly during later epochs;
    b. According to the procedural deficit hypothesis, adults with dyslexia are expected to show less evidence of procedural learning than TD adults;

   c. Participants with dyslexia will show slower RTs than the TD group but this will not account for the differences in procedural learning when participants' average speed is taken into account;

2. Is the procedural learning effect stable across sessions?
   a. Both groups will show higher split-half reliability than test-retest reliability, and for the control group these will be in a similar range to that found in previous studies;
   b. Reliability in the dyslexic group will be either equivalent to, or lower than, that of the control group (exploratory hypothesis).
   c. Higher test-retest reliability is expected for both groups for later sessions (2 and 3) than between session 1 and 2);

3. Does procedural learning correlate with languages and literacy measures?
   a. According to the procedural/declarative model, procedural learning should correlate with language and literacy measures. Later (more stable) sessions should in principle be more highly correlated with language/literacy, if there is a true correlation between these measures is present:
      i. Empirically, given our previous findings (Oliveira et al., submitted), these correlations are absent in TD groups;
      ii. However, these correlations may emerge in dyslexic participants if they exist only at the lower end;

4. Does procedural learning correlate with offline attention measures?
   a. Both groups will demonstrate a positive correlation between procedural learning and sustained attention, as measured by the psychomotor vigilance task;
   b. Adults with dyslexia are expected to show higher intra-individual variability in sustained attention than TD adults;
   c. Participants with higher intra-individual variability in sustained attention will show lower test-retest reliability on the SRTT;

5. Do participants with and without dyslexia differ in the amount of explicit awareness of the sequence?
   a. Based on previous studies (e.g., Du & Kelly, 2013; Russeler, Gerthe & Münte, 2006) we expect no significant differences in explicit awareness between groups;

Method

Participants

Sixty-two adults with dyslexia and fifty-six neurotypical adults aged between 18 and 35 years took part in the experiment. All participants were native English speakers, with normal or corrected-to-normal hearing and vision. Participants were recruited through various avenues, including Prolific, the participant panel at the Department of Psychology, University of York, and social media, and were compensated with £15 or a £15 Amazon gift voucher for their time. Adults with dyslexia confirmed having received a formal diagnosis of dyslexia and not having other documented neurodevelopmental disorders. At a group level, participants with dyslexia and TD adults were matched for age and non-verbal ability, but those with dyslexia performed significantly more poorly on standardised tests assessing literacy (spelling, word and nonword reading) (see Table 5.1), as well as self-reported significantly more literacy difficulties on the Adult Reading Questionnaire (Snowling et al., 2012). Participants with dyslexia who performed within the normal range in reading and spelling tasks (i.e., > 90) were excluded from the study (N = 7), as were TD participants performing below the cut-off on the literacy tasks (i.e., <90) (N = 10).

The intended sample size was 42 participants (acceptable range 42 – 50) per group who completed all sessions. As described in our pre-registration (https://osf.io/qf258), given our limited resources, power was optimised for the analyses of test-retest reliability using the test-retest reliability coefficient from our previous study (*r* = .42) as the smallest effect size of interest (Oliveira et al., submitted). Thus, we cannot ensure that the multilevel models are fully powered.

Ethical approval was provided by the Department of Psychology Ethics Committee at the University of York.

Table 5.1

Mean (and SD) scores for all background measures

Measure	***Test***	***TD group***				***Dyslexic group***				***F, p values***
		Raw scores		Standardised scores		Raw scores		Standardised scores		
		Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range	
Age		26.20 (5.70)	18 - 35	-	-	25.74 (4.80)	18 - 35	-	-	0.01, p = .9344
Matrix reasoning	WASI – II	21.37 (3.31)	13 - 28	52.19 (8.23)	34 - 73	20.95 (2.89)	12 - 26	50.64 (6.52)	32 - 65	0.59, p = .4434
Word reading	WIAT – III	71.00 (2.06)	65 - 74	105.39 (6.95)	93 - 121	60.91 (8.09)	26 - 71	86.79 (10.57)	50 - 103	**48.71, p < .001**
Nonword reading	WIAT – III	47.86 (2.36)	41 - 51	111.09 (6.67)	93 - 121	32.57 (8.65)	14 - 48	80.26 (12.71)	58 - 112	**159.17, p < .001**
Spelling	WIAT – III	57.00 (4.17)	48 - 63	113.05 (11.28)	91 - 132	38.83 (7.60)	21 - 55	78.67 (10.33)	60 - 106	**159.17, p < .001**
Sentence Recall	CELF 5	67.29 (7.39)	49 - 78	10.38 (3.26)	5 - 18	64.11 (5.65)	50 - 75	8.73 (2.16)	5 - 14	**6.69, p = .0117**
Nonword Repetition	CToPP-2	20.93 (3.32)	11 - 29	11.02 (3.10)	3 - 19	17.41 (3.24)	10 - 23	7.89 (2.63)	2 - 13	**19.47, p < .001**
Vocabulary	WASI – II	38.08 (6.95)	20 - 52	-	-	37.83 (4.62)	22 - 46	-	-	0.08, p = .7848
Conners	-	34.19 (13.74)	11 - 60	-	-	36.34 (12.28)	15 - 69	-	-	0.12, p = .7352
ARQ	ARQ	12.36 (4.26)	5 - 23	-	-	24.99 (5.90)	13 - 34	-	-	**105.37, p < .001**

Note. CELF - 5 UK, Clinical Evaluation of Language Fundamentals - Fifth Edition (Semel, Wiig, & Secord, 2017); CToPP-2, Comprehensive Test of Phonological Processing 2 (Wagner et al., 2013); WASI - II, Wechsler Abbreviated Scales of Intelligence – Second edition (Wechsler, 2011); WIAT – III, Wechsler Individual Achievement Test - Third UK Edition (Wechsler, 2009); ARQ, Adult Reading Questionnaire (Snowling et al., 2012)

Measures and procedure

The experiment used a mixed-subjects design with each participant performing the SRTT at three time points approximately one week apart. All sessions started with the administration of the SRTT followed by the standardised tests.

The nonverbal probabilistic Serial Reaction Time task used in the present study follows the protocol outlined in Oliveira et al. (submitted, Chapter 2 of this thesis). Participants were presented with an array of four rectangles displayed horizontally, with a stimulus (smiley face) appearing in one of the four positions on each trial. Participants were asked to press the corresponding key on the keyboard to the position of the stimulus on screen as soon as possible. Trials only proceeded once a response was made. Unknown to the participant, and given the probabilistic nature of this task, the position of the stimuli on screen follows either a probable (90% of the trials) or an improbable (10% of the trials) sequence. Procedural learning in this task is reflected in a difference in RTs between improbable and probable trials, since learning of the underlying sequence would allow participants to anticipate the position of the following stimulus.

The SRTT comprised two second order conditional 12 items sequences on each session, the pairs of sequences were randomised across participants, resulting in six groups, and were taken from Shanks, Wilkinson and Channon (2003): probable sequence A – 314324213412; improbable sequence B – 431241321423; Schwaneveldt and Gomez (1998): probable sequence C – 121342314324; improbable sequence D – 123413214243; and Kaufman et al. (2010): probable sequence E - 121432413423; improbable sequence F - 323412431421.

Following the SRTT, in the first session, literacy abilities were assessed using the spelling and word and nonword reading tasks from the Wechsler Individual Achievement Test, third edition UK (WIAT-III UK), as well as a self-report reading questionnaire - the Adult Reading Questionnaire (Snowling et al., 2012). Language abilities were assessed in session 2 using the expressive vocabulary task from the Wechsler Abbreviated Scale of Intelligence - Second Edition (WASI-II); the nonword repetition task from Comprehensive Test of Phonological Processing - Second Edition (CTOPP-2); and the sentence recall task from the Clinical Evaluation of Language Fundamentals - Fifth Edition (CELF-5 UK). Finally, in the third and last session, after completing the SRTT, participants' explicit knowledge of the sequence was measured through two generation tasks (Wilkinson & Shanks, 2004). The first generation task, under inclusion conditions, required each participant to generate 100 guesses of the sequence learned on the SRTT in the last session; whilst in the exclusion condition participants were asked to produce a sequence as different as possible from the training sequence. Each sequence generated by the participants was coded into the total amount of triplets recalled. Following the generation tasks, the attention abilities of the participants were assessed using a self-report questionnaire - Conners' Adult ADHD Rating Scales (CAARS) whilst nonverbal intelligence was assessed through the Reasoning Matrix task from the Wechsler Abbreviated Scale of Intelligence - II (WASI- II). All cognitive measures were delivered and scored in accordance with manual instructions.

At the end of each session, sustained attention was measured experimentally through a 5-min version of the Psychomotor Vigilance task (Reifman et al., 2018) to further explore the relationship between attention and procedural learning. This task has been shown to have adequate psychometric properties (Oliveira et al., submitted), with the median of the RTs providing the most stable estimate of individuals' attention abilities. An ex-Gaussian analysis was also performed on the Psychomotor Vigilance task as the tau parameter (representing the skewness or variability of the slow responses) has been suggested as an endophenotypic marker of attentional disorders which better captures moment-to-moment fluctuations in performance than other traditional measures of RTs (Gooch et al., 2012; Henríquez-Henríquez et al., 2015; Lin et al., 2015; van Belle et al., 2015).

Tasks were designed and hosted on the Gorilla (Anwyl-Irvine et al., 2020) and Pavlovia (Bridges et al., 2020; Peirce et al., 2019) platforms.

Statistical analyses

H1: Assessing group differences in procedural learning

RTs were grouped into epochs of 200 trials. Note that the first two trials of each epoch were removed for the purposes of analysis, as these are not predictable in the current task structure, which uses second-order conditionals (requiring participants to consider the positions of the two previous trials to be able to predict the following position). Outlier RTs were identified using a moving criterion based on sample size given the uneven number of probable and improbable trials, with the cutoff point depending on the number of correct trials per condition and block (Van Selst & Jolicoeur, 1994; Cousineau & Chartier, 2010). After removing outlier trials, participants whose mean accuracy and RTs were 2.5 standard deviations below the mean were removed from further analyses.

RTs were log transformed due to the skewed nature of the RTs (Brysbaert, & Stevens, 2018). The residuals were then inspected to determine whether they met the assumptions of normality for fitting the linear mixed effects models.

Linear mixed models were fitted to the RTs in order to compare the performance of adults with and without dyslexia. An initial model comprised *Probability* (Probable vs Improbable), *Epoch* (successive contrasts), *Session* and *Group* (Control vs dyslexic) as fixed effects and Participants as random effects. For the three-level factor of session two orthogonal contrasts were set: delay1 which contrasts session 1 with sessions 2 and 3 (S1 vs S2 & S3) and delay2 contrasts the performance in sessions 2 and 3. All dichotomous variables were contrasted using effect coding. All interactions between the predictors were analysed. Following theoretical evidence and Barr et al.'s (2013) recommendation, a backward selection from the maximal random-effects structure was fitted, with the random structure being simplified if necessary due to convergence issues according to the goodness of fit (AIC - Akaike Information Criterion; Akaike, 1974) with a smaller value indicating a better fit.

H1, H3 and H4: Assessing group differences in procedural learning and the role of language, literacy and attention on the procedural learning effect

Five additional models exploring the role of language and attention abilities on procedural memory were fitted. All models include *Probability*, *Session* and *Group* as fixed effects, with the addition of a predictor for language and attention, depending on the model. These models did not include *Epoch* as a fixed effect (unlike the initial model described above) and were estimated on the last 600 trials when learning is expected to be more robust, in order to avoid computational issues. The dependent variable and random effects remained the same. For the models including language and literacy abilities as a predictor, all measures were centred and standardised. For the literacy model, a composite measure for literacy was computed as variables (i.e., nonword reading, word reading and spelling) were highly correlated (r > .5). Separate models were fitted for each of the remaining variables: vocabulary, nonword repetition and sentence recall. For attention, given the low correlation between attention measures (conners and median PVT1: r = .07, conners and median PVT2: r = .03, conners and median PVT3: r = .10), only the PVT measures were used as the attention index. The median performance on the PVT was adopted for the modelling as this variable has been previously found to be highly reliable (Oliveira et al., submitted).

R software - version 4.1.0 (RStudio Team, 2015) and *lme4* package (Bates, Maechler & Bolker, 2012) were used to fit all linear mixed effects models. P-values for these models were obtained through the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017) and corrected for multiple comparisons using the Holm-Bonferroni method (Holm, 1979). Thus, all statements regarding significance reflect the analyses after correction for family-wise error rates. All figures were produced using the package *ggplot2* (Wickham, 2009). DFbetas were calculated and standardised after model selection via the *influence.ME* package (Nieuwenhuis, Pelzer, & te Grotenhuis, 2012). Participants with z-scores greater than +/- 3.29 were reported as influential cases.

H2: Assessing the reliability of the procedural learning effect

To establish the reliability of the SRTT, two measures of procedural learning were computed: difference scores and random slopes. Difference scores, the most commonly used procedural learning measure for the SRTT, were computed as the simple difference between improbable and probable RTs, with a positive value indicating procedural learning. Random slopes for each participant/session were obtained by running a linear mixed effects model with response time as a dependent variable and *Probability* and *Group* as predictors. For the random structure participants were introduced as a random intercept and probability as a slope.

To measure the split-half reliability for both sessions/groups, trials were separated into probable and improbable trials. Consecutive trials were labelled as odd or even. The split-half reliability was calculated by running Pearson's correlations for the procedural learning (difference scores/random slopes) effect for even and odd trials.

Pearson's correlations between the procedural learning scores (difference scores/random slopes) from session 1 and session 2 and from session 2 and 3 for each group were taken as measures of test-retest reliability. Bland-Altman plots (Bland & Altman, 1986) were plotted to assess the levels of agreement. Fisher's r-to-z transformations were computed to determine whether there were significant differences between reliability measures for each group (dyslexia vs TD) and contrasts (test-retest reliability: S1-S2 vs S2-S3).

H3 and H4: Assessing the relationship between procedural learning and cognitive abilities

The relationship between language, literacy, cognitive, explicit awareness measures and procedural learning was explored through Pearson's correlations for each session by group. In similarity with previous models, differences in the magnitude of the relationship between procedural learning and cognitive measures were explored through Fisher-z transformations, whilst correcting for multiple comparisons using the Holm-Bonferroni method (Holm, 1979). Bayesian Pearson's correlations were also computed alongside using the BayesFactor package (Morey & Rouder, 2022) as delineated in the second chapter, since non-significant findings do not provide evidence for the null hypothesis as these may occur due to lack of evidence for a correlation but also due to lack of power. Bayes factors above 3 or below ⅓ were interpreted as providing support for the alternative or null, respectively (Jeffreys, 1961).

H5: Assessing group differences in explicit awareness

Explicit awareness performance was analysed by comparing the performance of each group in the generation task for each condition (inclusion and exclusion conditions). A two-way ANOVA was conducted with *Group* and *Condition* as predictors, to determine whether there are significant differences between groups in the total number of triplets recalled of the sequence at end of session 3.

Results

RT data were available for 118 participants for Session 1 (dyslexia: N = 62; TD: N = 56), for 100 participants for Session 2 (dyslexia: N = 49; TD: N = 51) and 95 for the last session (dyslexia: N = 46; TD: N = 49). Data from 17 participants (dyslexia: N = 7; TD: N = 10) was removed from the analyses as these did not meet inclusion criteria (i.e., dyslexic participants performing above the 90th percentile on at least one of the literacy measures and controls performing below the 90th percentile on literacy measures). The remaining participants failed to return for follow-up sessions.

H1: Assessing group differences in procedural learning

As illustrated in Figure 5.1, RTs decreased with practice, both within and across sessions, with no significant difference between groups on overall RTs, yet there was a significant difference between groups prior to removal of influential participants[3]. This is reflected in the significant effect of *Epoch* for the later contrasts (Epoch4-3 and Epoch5-4), even though they did not survive correction for multiple comparisons, and *Session* for both contrasts (Delay1 and Delay2). The significant effect of *Probability* reveals that there was a difference in RTs between probable and improbable trials, with faster RTs for probable trials than improbable, indicating procedural learning. The procedural learning effect significantly increased with practice with a significant interaction between *Probability* x *Epoch* and from Session 1 to Session 2 (significant interaction *Probability* x *Session for* Delay 1) but did not differ between Sessions 2 and 3 (nonsignificant interaction *Probability* x *Session for* Delay 2), possibly indicating that participants showed less practice effects after the first session. The lack of a three-way interaction between *Probability*, *Epoch* and *Session* for both contrasts (Delay 1 and 2) suggests that the learning trajectory within each session was similar across sessions.

[3] All results with and without influential points are available at Open Science Framework (OSF) project pages

Importantly, group comparisons revealed no overall significant differences in RTs, as evidenced by the non-significant *Group* effect. There was also no difference in the procedural learning effect between TD adults and adults with dyslexia, as evidenced by the non-significant interactions of *Probability* x *Group, Probability* x *Session* (Delay1 and 2) x *Group* and *Probability* x *Epoch* x *Session* x *Group* (Delay1 and 2).

Figure 5.1

Mean and 95% CI RTs for probable and improbable trials per Epoch and Session for TD and dyslexic groups (Session 1 on the left, Session 2 in the centre and Session 3 on the right)

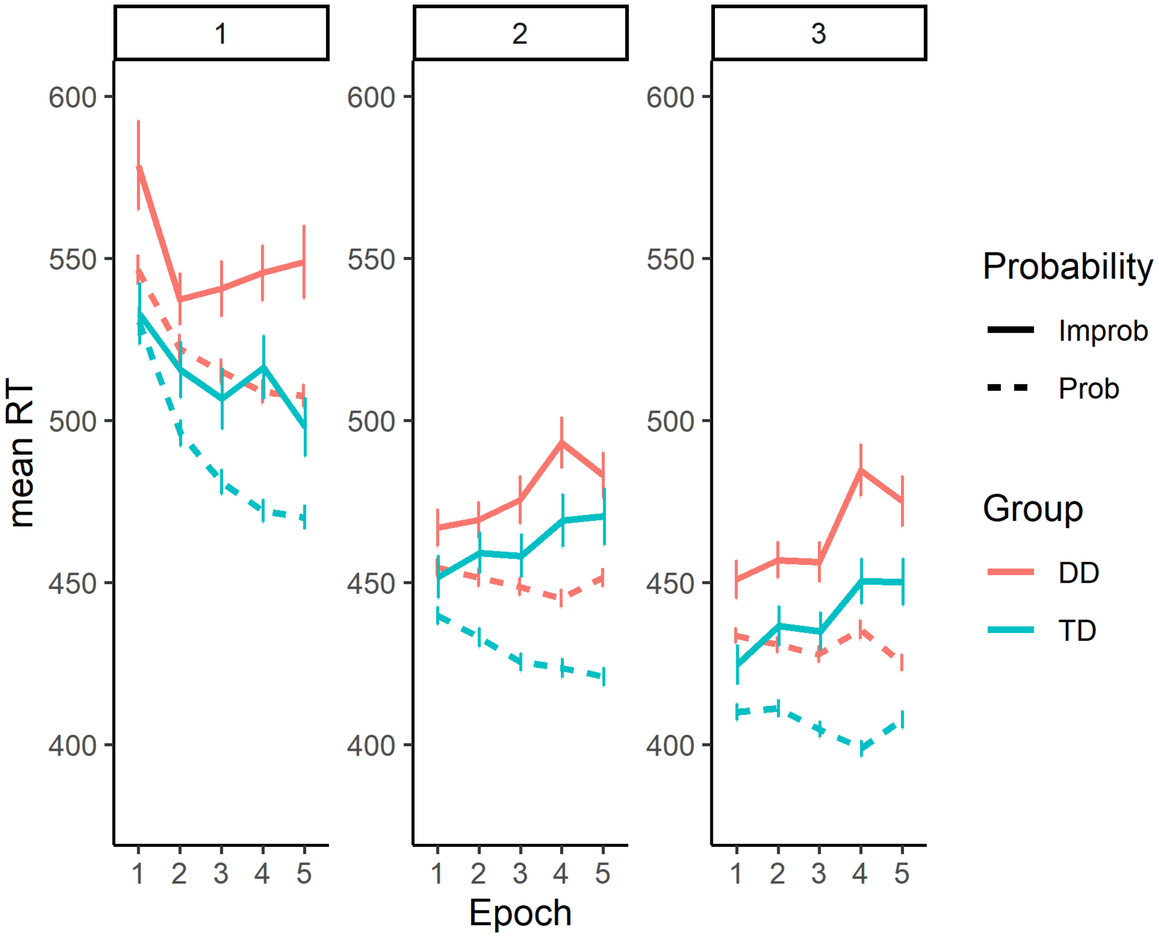


Table 5.2

Predictors of the group effect on the magnitude of procedural learning

Fixed effects	*b*	*SE*	*t*	*p*	*CI*	
(Intercept)	**6.105**	**0.014**	**449.019**	**<.001**	**6.078**	**6.132**
Probability	**0.031**	**0.001**	**22.683**	**<.001**	**0.028**	**0.034**
Epoch2-1	-0.010	0.005	-1.970	.052	-0.020	0.000
Epoch3-2	-0.008	0.004	-2.054	.043	-0.015	0.000
Epoch4-3	0.008	0.004	1.923	.058	0.000	0.016
Epoch5-4	-0.009	0.004	-2.460	.015	-0.017	-0.002
Delay1 (S1 vs S2 and S3)	**-0.044**	**0.003**	**-14.467**	**<.001**	**-0.050**	**-0.038**
Delay2 (S2 vs S3)	**-0.021**	**0.004**	**-5.848**	**<.001**	**-0.028**	**-0.014**
Group	0.021	0.014	1.553	.124	-0.006	0.048
Probability x Epoch2-1	**0.012**	**0.002**	**6.132**	**<.001**	**0.008**	**0.015**
Probability1 x Epoch3-2	0.004	0.002	1.887	.059	0.000	0.007
Probability1 x Epoch4-3	**0.018**	**0.002**	**9.071**	**<.001**	**0.014**	**0.022**
Probability x Epoch5-4	**-0.008**	**0.002**	**-3.896**	**<.001**	**-0.012**	**-0.004**
Probability x Delay1	**0.003**	**0.000**	**6.971**	**<.001**	**0.002**	**0.004**
Probability x Delay2	0.002	0.001	2.490	.013	0.000	0.003
Epoch2-1 x Delay1	**0.015**	**0.001**	**10.937**	**<.001**	**0.012**	**0.017**
Epoch3-2 x Delay1	0.001	0.001	0.620	.535	-0.002	0.003
Epoch4-3 x Delay1	0.004	0.001	2.798	.005	0.001	0.007
Epoch5-4 x Delay1	0.003	0.001	2.019	.043	0.000	0.006
Epoch2-1 x Delay2	0.008	0.002	3.246	.001	0.003	0.012
Epoch3-2 x Delay2	-0.003	0.002	-1.252	.211	-0.008	0.002
Epoch4-3 x Delay2	0.003	0.003	1.046	.295	-0.002	0.008
Epoch5-4 x Delay2	-0.003	0.003	-1.225	.220	-0.008	0.002
Probability x Group	-0.002	0.001	-1.821	.072	-0.005	0.000
Epoch2-1 x Group	-0.005	0.005	-0.949	.345	-0.014	0.005
Epoch3-2 x Group	0.005	0.004	1.337	.184	-0.002	0.012

Epoch4-3 x Group	0.007	0.004	1.659	.100	-0.001	0.015
Epoch5-4 x Group	-0.005	0.004	-1.430	.155	-0.013	0.002
Delay1 x Group	0.000	0.003	0.136	.892	-0.006	0.006
Delay2 x Group	0.001	0.004	0.265	.792	-0.006	0.008
Probability x Epoch2-1 x Delay1	0.002	0.001	1.228	.219	-0.001	0.004
Probability x Epoch3-2 x Delay1	-0.001	0.001	-0.514	.607	-0.003	0.002
Probability x Epoch4-3 x Delay1	0.001	0.001	0.457	.648	-0.002	0.003
Probability x Epoch5-4 x Delay1	0.002	0.001	1.466	.143	-0.001	0.005
Probability x Epoch2-1 x Delay2	-0.001	0.002	-0.488	.625	-0.006	0.003
Probability x Epoch3-2 x Delay2	0.000	0.002	0.197	.844	-0.004	0.005
Probability x Epoch4-3 x Delay2	0.001	0.003	0.250	.803	-0.004	0.006
Probability x Epoch5-4 x Delay2	0.000	0.003	-0.170	.865	-0.005	0.005
Probability x Epoch2-1 x Group	-0.003	0.002	-1.519	.129	-0.007	0.001
Probability x Epoch3-2 x Group	0.000	0.002	-0.084	.933	-0.004	0.004
Probability x Epoch4-3 x Group	0.000	0.002	-0.208	.835	-0.004	0.004
Probability x Epoch5-4 x Group	-0.002	0.002	-0.949	.343	-0.006	0.002
Probability x Delay1 x Group	0.000	0.000	-1.037	.300	-0.001	0.000
Probability x Delay2 x Group	0.003	0.001	3.233	.001	0.001	0.004
Epoch2-1 x Delay1 x Group	-0.002	0.001	-1.397	.162	-0.004	0.001
Epoch3-2 x Delay1 x Group	**-0.006**	**0.001**	**-4.738**	**<.001**	**-0.009**	**-0.004**
Epoch4-3 x Delay1 x Group	0.003	0.001	2.374	.018	0.001	0.006
Epoch5-4 x Delay1 x Group	0.001	0.001	0.515	.607	-0.002	0.004
Epoch2-1 x Delay2 x Group	-0.003	0.002	-1.370	.171	-0.008	0.001
Epoch3-2 x Delay2 x Group	0.001	0.002	0.298	.765	-0.004	0.005
Epoch4-3 x Delay2 x Group	0.004	0.003	1.445	.148	-0.001	0.009
Epoch5-4 x Delay2 x Group	**-0.010**	**0.003**	**-3.804**	**<.001**	**-0.015**	**-0.005**
Probability x Epoch2-1 x Delay1 x Group	-0.002	0.001	-1.767	.077	-0.005	0.000
Probability x Epoch3-2 x Delay1 x Group	-0.001	0.001	-0.442	.659	-0.003	0.002

Probability x Epoch4-3 x Delay1 x Group	0.002	0.001	1.357	.175	-0.001	0.005
Probability x Epoch5-4 x Delay1 x Group	-0.001	0.001	-0.413	.680	-0.003	0.002
Probability x Epoch2-1 x Delay2 x Group	0.002	0.002	0.739	.460	-0.003	0.006
Probability x Epoch3-2 x Delay2 x Group	-0.001	0.002	-0.548	.584	-0.006	0.003
Probability x Epoch4-3 x Delay2 x Group	-0.002	0.003	-0.777	.437	-0.007	0.003
Probability x Epoch5-4 x Delay2 x Group	0.004	0.003	1.438	.150	-0.001	0.009

Random effects	Variance	SD
Participant: (Intercept)	0.015	0.121
Participant: Epoch2-1 (Slope)	0.002	0.041
Participant: Epoch3-2 (Slope)	0.001	0.029
Participant: Epoch4-3 (Slope)	0.001	0.032
Participant: Epoch5-4 (Slope)	0.001	0.027
Participant: Delay1 (Slope)	0.001	0.026
Participant: Delay2 (Slope)	0.001	0.029
Participant: Probability (Slope)	0.000	0.011

H1 and H3: Assessing group differences in procedural learning and the role of language and literacy on the procedural learning effect

The same pattern of results emerged for each model incorporating literacy, nonword repetition, sentence recall and vocabulary (see Tables 5.3-5.7). These all showed clear evidence of procedural learning, with a significant effect of *Probability*. There was suggestive evidence that the difference between probable and improbable trials increased with practice, indicated by the interaction between *Probability* and *Session*, but this effect did not survive correction for multiple comparisons, except for the vocabulary model. Furthermore, and similar to the initial model described in the previous section, these models indicate that overall RTs decreased with practice across sessions (significant effect of *Session* for one or both contrasts).

Whilst there was a trend indicative of slower RTs for individuals with lower nonword repetition scores (with the reverse pattern for literacy), notably only for sessions 1 and 2, the relevant interactions were not significant following correction for multiple comparisons (i.e., *Literacy* and *Session*, and *Nonword repetition* and *Session*). Most relevant for the aims of the study, and contrary to our hypothesis, there was no evidence that *Literacy*, *Nonword repetition* or *Sentence recall* predicted the procedural learning effect, as evidenced by the non-significant interactions between *Probability* and these variables. Instead, there was a significant interaction between *Vocabulary* and *Probability*, indicating that participants with better vocabulary abilities showed more evidence of procedural learning. However, this interaction did not survive correction for multiple comparisons. Finally, there was no indication that dyslexia status affected the relationship between procedural learning and language or literacy measures, as there were neither main nor interaction effects for *Group*.

Table 5.3

Literacy as a predictor of the group effect on the magnitude of procedural learning

Fixed effects	*b*	*SE*	*t*	*p*	*CI*	
(Intercept)	**6.175**	**0.040**	**156.240**	**<.001**	**6.096**	**6.254**
Probability	**0.039**	**0.005**	**8.256**	**<.001**	**0.029**	**0.048**
Delay1 (S1 vs S2 and S3)	**-0.041**	**0.008**	**-4.974**	**<.001**	**-0.058**	**-0.025**
Delay2 (S2 vs S3)	-0.016	0.009	-1.750	.084	-0.033	0.002
Group	-0.015	0.040	-0.379	.706	-0.094	0.064
Literacy	-0.054	0.045	-1.204	.232	-0.143	0.035
Probability x Delay1	0.008	0.003	2.930	.005	0.002	0.013
Probability x Delay2	0.003	0.004	0.779	.439	-0.005	0.012
Probability x Group	-0.002	0.005	-0.332	.741	-0.011	0.008
Delay1 x Group	-0.006	0.008	-0.741	.461	-0.023	0.010
Delay2 x Group	-0.018	0.009	-2.060	.043	-0.036	-0.001
Probability x Literacy	0.003	0.005	0.533	.596	-0.008	0.013

Delay1 x Literacy	-0.007	0.009	-0.787	.434	-0.026	0.011
Delay2 x Literacy	-0.025	0.010	-2.524	.014	-0.045	-0.005
Group x Literacy	0.087	0.045	1.946	.055	-0.002	0.176
Probability x Delay1 x Group	-0.005	0.003	-2.059	.043	-0.011	0.000
Probability x Delay2 x Group	0.001	0.004	0.134	.894	-0.008	0.009
Probability x Delay1 x Literacy	-0.005	0.003	-1.755	.084	-0.011	0.001
Probability x Delay2 x Literacy	-0.003	0.005	-0.531	.597	-0.012	0.007
Probability x Group x Literacy	0.001	0.005	0.109	.913	-0.010	0.011
Delay1 x Group x Literacy	-0.002	0.009	-0.177	.860	-0.020	0.017
Delay2 x Group x Literacy	0.006	0.010	0.581	.563	-0.014	0.026
Probability x Delay1 x Group x Literacy	0.004	0.003	1.524	.132	-0.001	0.010
Probability x Delay2 x Group x Literacy	0.001	0.005	0.205	.838	-0.009	0.011

Random effects	*Variance*	*SD*
Participant: (Intercept)	0.017	0.129
Participant: Probability (Slope)	0.000	0.013
Participant: Delay1 (Slope)	0.001	0.025
Participant: Delay2 (Slope)	0.001	0.026
Participant: Probability x Delay1 (Slope)	0.000	0.006
Participant: Probability x Delay2 (Slope)	0.000	0.009

Table 5.4

Nonword repetition as a predictor of the group effect on the magnitude of procedural learning

Fixed effects	*b*	*SE*	*t*	*p*	*CI*	
(Intercept)	**6.114**	**0.015**	**405.090**	**< .001**	**6.084**	**6.144**
Probability	**0.040**	**0.002**	**21.885**	**<.001**	**0.036**	**0.044**
Delay1 (S1 vs S2 and S3)	**-0.042**	**0.003**	**-13.137**	**<.001**	**-0.048**	**-0.035**
Delay2 (S2 & S3)	**-0.024**	**0.004**	**-6.062**	**<.001**	**-0.032**	**-0.016**

Group	0.015	0.015	0.998	.321	-0.015	0.045
Nonword repetition	-0.023	0.017	-1.291	.200	-0.057	0.012
Probability x Delay1	0.003	0.001	2.909	.005	0.001	0.005
Probability x Delay2	0.004	0.002	1.929	.057	0.000	0.007
Probability x Group	-0.001	0.002	-0.706	.482	-0.005	0.002
Delay1 x Group	-0.002	0.003	-0.708	.481	-0.009	0.004
Delay2 x Group	0.005	0.004	1.266	.210	-0.003	0.013
Probability x Nonword repetition	0.001	0.002	0.397	.692	-0.003	0.005
Delay1 x Nonword repetition	-0.008	0.004	-2.094	.040	-0.016	0.000
Delay2 x Nonword repetition	0.005	0.005	0.993	.324	-0.005	0.014
Group x Nonword repetition	0.002	0.017	0.141	.888	-0.032	0.037
Probability x Delay1 x Group	0.000	0.001	-0.388	.700	-0.003	0.002
Probability x Delay2 x Group	0.002	0.002	1.180	.242	-0.001	0.006
Probability x Delay1 x Nonword repetition	0.001	0.001	0.710	.480	-0.002	0.004
Probability x Delay2 x Nonword repetition	0.000	0.002	0.040	.968	-0.004	0.004
Probability x Group x Nonword repetition	0.001	0.002	0.386	.701	-0.003	0.005
Delay1 x Group x Nonword repetition	-0.005	0.004	-1.304	.196	-0.013	0.003
Delay2 x Group x Nonword repetition	-0.004	0.005	-0.751	.455	-0.013	0.006
Probability x Delay1 x Group x Nonword repetition	0.000	0.001	0.027	.978	-0.003	0.003
Probability x Delay2 x Group x Nonword repetition	0.003	0.002	1.303	.196	-0.001	0.007

Random effects	*Variance*	*SD*
Participant: (Intercept)	0.044	0.210
Participant: Probability (Slope)	0.000	0.015
Participant: Delay1 (Slope)	0.001	0.023
Participant: Delay2 (Slope)	0.001	0.027
Participant: Block (Slope)	0.003	0.054
Participant: Probability x Delay1 (Slope)	0.000	0.006

Participant: Probability x Delay2 (Slope)	0.000	0.011

Table 5.5

Sentence recall as a predictor of the group effect on the magnitude of procedural learning

Fixed effects	*b*	*SE*	*t*	*p*	*CI*	
(Intercept)	**6.111**	**0.014**	**441.872**	**<.001**	**6.083**	**6.138**
Probability	**0.041**	**0.002**	**23.873**	**<.001**	**0.038**	**0.045**
Delay1 (S1 vs S2 and S3)	**-0.039**	**0.003**	**-13.083**	**<.001**	**-0.044**	**-0.033**
Delay2 (S2 & S3)	**-0.019**	**0.004**	**-5.435**	**<.001**	**-0.026**	**-0.012**
Group	0.033	0.014	2.364	.021	0.005	0.060
Sentence recall	-0.012	0.015	-0.779	.438	-0.043	0.019
Probability x Delay1	0.003	0.001	2.947	.004	0.001	0.005
Probability x Delay2	0.002	0.002	1.249	.215	-0.001	0.005
Probability x Group	-0.002	0.002	-1.250	.215	-0.006	0.001
Delay1 x Group	0.000	0.003	0.140	.889	-0.005	0.006
Delay2 x Group	0.003	0.004	0.949	.346	-0.004	0.010
Probability x Sentence recall	0.002	0.002	1.122	.265	-0.002	0.006
Delay1 x Sentence recall	-0.001	0.003	-0.268	.790	-0.007	0.006
Delay2 x Sentence recall	0.002	0.004	0.632	.530	-0.005	0.010
Group x Sentence recall	-0.009	0.015	-0.561	.577	-0.039	0.022
Probability x Delay1 x Group	0.000	0.001	-0.343	.732	-0.002	0.002
Probability x Delay2 x Group	0.001	0.002	0.582	.562	-0.002	0.004
Probability x Delay1 x Sentence recall	0.001	0.001	0.643	.522	-0.001	0.003
Probability x Delay2 x Sentence recall	-0.002	0.002	-1.254	.214	-0.006	0.001
Probability x Group1 x Sentence recall	0.003	0.002	1.377	.172	-0.001	0.006
Delay1 x Group x Sentence recall	0.003	0.003	0.927	.357	-0.003	0.009
Delay2 x Group x Sentence recall	0.004	0.004	1.059	.293	-0.004	0.012
Probability x Delay1 x Group x Sentence recall	-0.001	0.001	-0.807	.422	-0.003	0.001

Probability x Delay2 x Group x Sentence recall	0.001	0.002	0.342	.733	-0.003	0.004

Random effects	*Variance*	*SD*
Participant: (Intercept)	0.039	0.198
Participant: Probability (Slope)	0.000	0.014
Participant: Delay1 (Slope)	0.001	0.023
Participant: Delay2 (Slope)	0.001	0.027
Participant: Block (Slope)	0.003	0.053
Participant: Probability x Delay1 (Slope)	0.000	0.006
Participant: Probability x Delay2 (Slope)	0.000	0.010

Table 5.6

Vocabulary as a predictor of the group effect on the magnitude of procedural learning

Fixed effects	*b*	*SE*	*t*	*p*	*CI*	
(Intercept)	**6.104**	**0.013**	**481.185**	**<.001**	**6.079**	**6.130**
Probability	**0.040**	**0.002**	**23.950**	**<.001**	**0.037**	**0.043**
Delay1 (S1 vs S2 and S3)	**-0.039**	**0.003**	**-13.970**	**<.001**	**-0.044**	**-0.033**
Delay2 (S2 & S3)	**-0.020**	**0.003**	**-6.175**	**<.001**	**-0.026**	**-0.013**
Group	0.042	0.013	3.286	.002	0.016	0.067
Vocabulary	-0.032	0.014	-2.301	.024	-0.059	-0.004
Probability x Delay1	**0.004**	**0.001**	**4.213**	**<.001**	**0.002**	**0.006**
Probability x Delay2	0.002	0.002	1.472	.145	-0.001	0.006
Probability x Group	-0.003	0.002	-1.696	.095	-0.006	0.001
Delay1 x Group	0.001	0.003	0.309	.758	-0.005	0.006
Delay2 x Group	0.002	0.003	0.640	.524	-0.004	0.008
Probability x Vocabulary	0.005	0.002	2.765	.007	0.001	0.009
Delay1 x Vocabulary	-0.002	0.003	-0.787	.434	-0.008	0.004
Delay2 x Vocabulary	0.001	0.004	0.206	.837	-0.007	0.008

Group x Vocabulary	-0.016	0.014	-1.126	.264	-0.043	0.012
Probability x Delay1 x Group	-0.001	0.001	-1.395	.168	-0.003	0.001
Probability x Delay2 x Group	0.002	0.002	1.082	.283	-0.001	0.005
Probability x Delay1 x Vocabulary	0.000	0.001	-0.185	.854	-0.002	0.002
Probability x Delay2 x Vocabulary	0.001	0.002	0.653	.516	-0.002	0.005
Probability x Group x Vocabulary	0.000	0.002	0.178	.859	-0.003	0.004
Delay1 x Group x Vocabulary	-0.005	0.003	-1.662	.101	-0.011	0.001
Delay2 x Group x Vocabulary	0.008	0.004	2.188	.032	0.001	0.016
Probability x Delay1 x Group x Vocabulary	-0.001	0.001	-0.673	.503	-0.003	0.001
Probability x Delay2 x Group x Vocabulary	0.002	0.002	0.907	.367	-0.002	0.005

Random effects	*Variance*	*SD*
Participant: (Intercept)	0.012	0.110
Participant: Probability (Slope)	0.000	0.012
Participant: Delay1 (Slope)	0.001	0.023
Participant: Delay2 (Slope)	0.001	0.025
Participant: Probability x Delay1 (Slope)	0.000	0.006
Participant: Probability x Delay2 (Slope)	0.000	0.010

H1 and H4: Assessing group differences in procedural learning and the role of attention on the procedural learning effect

A final model was fitted to the RTs to address whether individuals' attentional skills as measured by the Psychomotor Vigilance Task are associated with performance on the SRTT. RTs decreased with practice as evidenced by a significant effect of *Session* for both contrasts (Delay 1 and Delay 2). As in the previous models, there was no main effect of *Group* or a *Session* x *Group* interaction.

Similar to previous models in this study, there was clear evidence of procedural learning as evidenced by a *Probability* effect, which again benefitted from practice as evidenced by the two-way interaction between *Probability* and *Session* for Delay1. The absence of a significant interaction between *Probability* and *Session* for Delay2 suggests that the procedural learning effect plateaued after session 2. Furthermore, both groups showed a similar procedural learning effect given the non-significant interaction between *Probability* x *Group* and *Probability* x *Group* x *Session* for both Delay2. The trend for an interaction between *Probability* x *Group* x *Session* for Delay1 indicates that the TD group showed a larger procedural learning effect than the dyslexic group from session 1 to 2, but this contrast did not survive correction for multiple comparisons.

Participants' attentional skills were not associated with overall RT, or with the procedural learning effect, evidenced by non-significant effects of *Attention* and no interactions between *Probability* x *Attention* and *Probability* x *Session* x *Attention*. The effect of attention on procedural learning did not significantly differ between groups as there was a non-significant interaction between *Probability* x *Group* x *Attention* and *Probability* x *Session* x *Group* x *Attention* for both contrasts (Delay 1 and Delay 2).

Table 5.7

Predictors of the attention effect on the magnitude of procedural learning

Fixed effects	*b*	*SE*	*t*	*p*	*CI*	
(Intercept)	**6.100**	**0.013**	**480.043**	**<.001**	**6.075**	**6.126**
Probability	**0.039**	**0.001**	**26.761**	**<.001**	**0.036**	**0.042**
Delay1 (S1 vs S2 and S3)	**-0.038**	**0.003**	**-14.038**	**<.001**	**-0.043**	**-0.033**
Delay2 (S2 vs S3)	**-0.020**	**0.003**	**-6.148**	**<.001**	**-0.027**	**-0.014**
Group	0.029	0.013	2.262	0.027	0.003	0.054
Attention	0.014	0.008	1.756	0.081	-0.002	0.031
Probability x Delay1	**0.004**	**0.001**	**4.224**	**<.001**	**0.002**	**0.006**
Probability x Delay2	0.002	0.002	1.404	0.164	-0.001	0.005
Probability x Group	-0.002	0.001	-1.578	0.119	-0.005	0.001
Delay1 x Group	-0.001	0.003	-0.322	0.749	-0.006	0.005
Delay2 x Group	0.001	0.003	0.190	0.850	-0.006	0.007
Probability x Attention	-0.001	0.002	-0.726	0.469	-0.005	0.002
Delay1 x Attention	0.006	0.003	1.855	0.068	0.000	0.013
Delay2 x Attention	0.001	0.004	0.340	0.735	-0.007	0.010

Group x Attention	0.000	0.008	0.012	0.991	-0.016	0.016
Probability x Delay1 x Group	-0.002	0.001	-2.257	0.027	-0.004	0.000
Probability x Delay2 x Group	0.002	0.002	1.414	0.162	-0.001	0.005
Probability x Delay1 x Attention	-0.001	0.001	-0.996	0.322	-0.004	0.001
Probability x Delay2 x Attention	0.000	0.002	0.049	0.961	-0.004	0.004
Probability x Group x Attention	0.003	0.002	1.435	0.154	-0.001	0.006
Delay1 x Group x Attention	0.001	0.003	0.147	0.883	-0.006	0.007
Delay2 x Group x Attention	-0.006	0.004	-1.435	0.155	-0.014	0.002
Probability x Delay1 x Group x Attention	0.001	0.001	0.640	0.524	-0.002	0.003
Probability x Delay2 x Group x Attention	0.002	0.002	0.823	0.413	-0.002	0.005

Random effects	*Variance*	*SD*
Participant: (Intercept)	0.013	0.112
Participant: Probability (Slope)	0.000	0.010
Participant: Delay1 (Slope)	0.000	0.022
Participant: Delay2 (Slope)	0.001	0.026
Participant: Probability x Delay1 (Slope)	0.000	0.006
Participant: Probability x Delay2 (Slope)	0.000	0.009

H2: Assessing the reliability of the procedural learning effect

Low to adequate (dyslexia: *r*s = .16 - .69, TD: *r*s = .06 -.78) split-half reliability was observed for both groups, with better split-half reliability observed in session 3 (Table 5.8). The split-half reliability in the first session was significantly higher for the dyslexic group for the last 600 trials when adopting random slopes, however none of the remaining contrasts were significant. Test-retest reliability was below psychometric standards (i.e., <.70) across contrasts and groups (TD: *r*s = -.02 - .33; DD: *r*s = .00 - .38). Whilst for the TD group there was a numerical improvement in test-retest reliability for later sessions, the reverse pattern was observed for the group with dyslexia (*p*s > .05). Similar patterns were observed for both difference scores and random slopes, and there was no difference between computing procedural learning based on all 1000 trials versus the last 600 trials (*p*s > .05).

Table 5.8

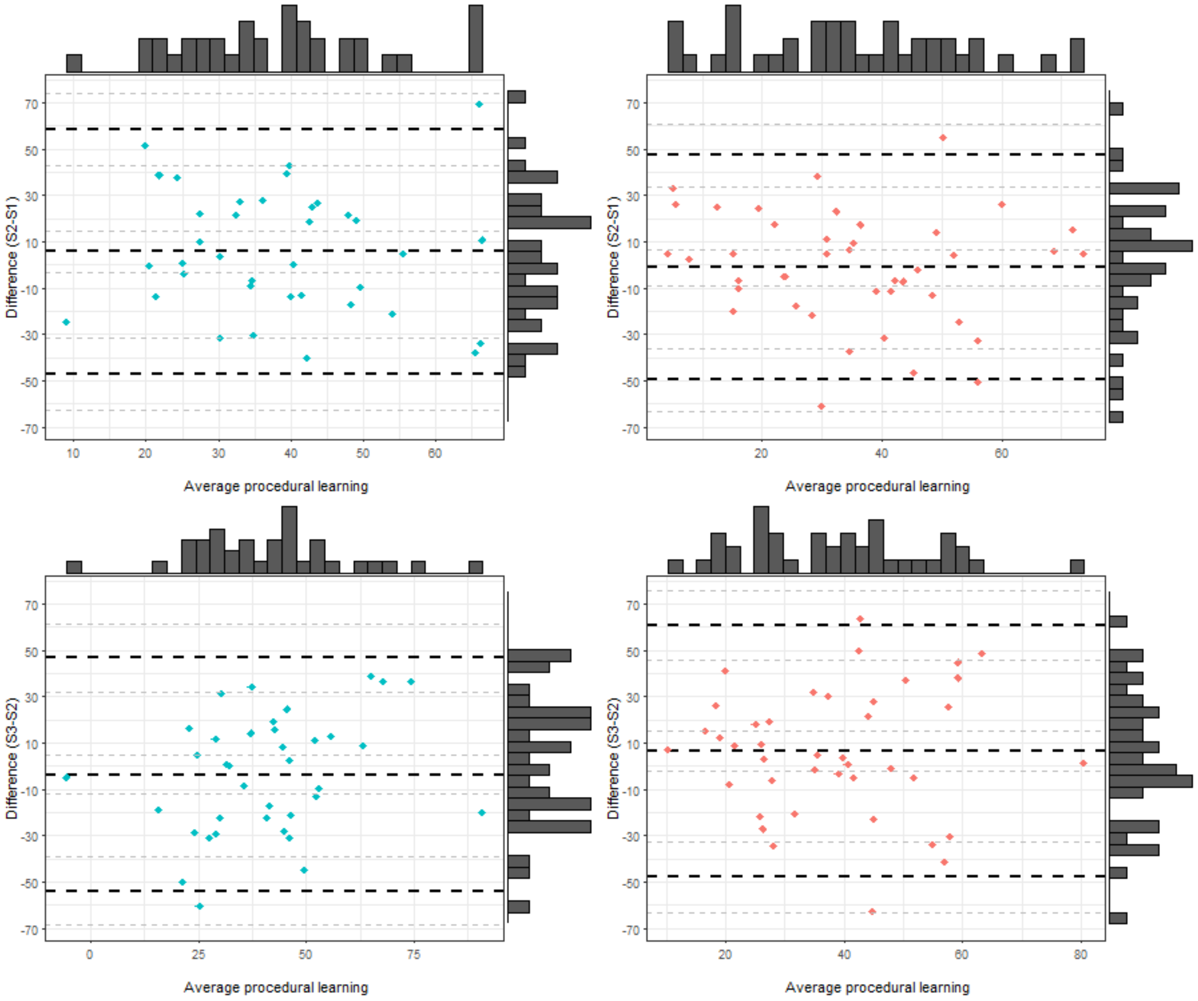
Split-half and test-retest reliability of the procedural learning measures for overall and last 600 trials of the SRTT for session 1 (SRTT1), session 2 (SRTT2) and session (SRTT3)

Measure		***Split-half reliability***						***Test-retest reliability***			
		TD (N 35-40)			DD (N 40 – 50)			TD (N 35 - 37)		DD (N 40 - 45)	
		SRTT 1	SRTT 2	SRTT 3	SRTT 1	SRTT 2	SRTT 3	Session 1-2	Session 2-3	Session 1-2	Session 2-3
Difference Scores	Overall	.46	.39	.71	.55	.19	.65	.07	.34	.14	.11
	Last 600 trials	.35	.24	.78	.51	.33	.69	.05	.33	.38	.10
Random Slopes	Overall	.25	.47	.72	.60	.35	.69	-.01	.33	.11	.07
	Last 600 trials	.06	.25	.76	.59	.16	.69	-.02	.33	.30	.00

Agreement in performance across sessions was examined using Bland-Altman plots (Figure 5.2). The 95% limits of agreement range between -46.82 to 58.51 for d=sessions 1 and 2 and between -53.67 to 46.70 for Sessions 2 and 3 for the TD group and between -49.48 to 47.49 for sessions 1 and 2 and between -47.81 to 60.92 for sessions 2 and 3 for the dyslexic group. Irrespective of groups, whilst only a small number of participants fell outside the limits of agreement, the limits are very wide, pointing to poor agreement between measures (i.e., a high level of variability in the magnitude of the procedural learning effect between sessions). Whilst for the TD group the precision of the limits of agreement improved slightly between earlier and later sessions, the opposite pattern was observed for the dyslexic group.

Figure 5.2

Plot of the mean of the two measurements against the differences between procedural learning in session 1 and session 2 (top) and session 2 and 3 (bottom) for the TD (left) and dyslexic (right) groups



In keeping with previous findings (Oliveira et al., submitted), the Psychomotor Vigilance task showed adequate split-half and test-retest reliability (i.e., *r*s >.70); with slightly better psychometric properties for the median RTs and lower reliability for the reciprocal (i.e., 1/RT). A two-way MANOVA revealed no significant differences between groups on the performance on the Psychomotor Vigilance task, for the tau parameter and median for all sessions (*p*s > .05), thus suggesting that these groups showed comparable attentional abilities.

Table 5.9

Descriptive statistics, split-half and test-retest reliability of the Psychomotor Vigilance task

Measure	***Descriptive statistics***									***Split-half reliability***			***Test-retest reliability***	
	Session 1			Session 2			Session 3			S1	S2	S3	S 1-2	S 2-3
	N	M	SD	N	M	SD	N	M	SD					
Lapses	89	3.69	4.53	85	3.78	4.46	85	5.18	6.11	.82	.73	.90	.37	.70
Mean RT	88	385.83	120.98	85	372.09	67.75	84	388.12	97.44	.74	.86	.86	.34	.62
Median	91	346.30	88.53	85	346.13	60.22	85	364.90	72.21	.93	.90	.94	.74	.79
Reciprocal	92	3.01	1.05	85	3.19	2.746	85	2.86	0.58	.78	.85	.89	-.003	.08

H3 and H4: Assessing the relationship between procedural learning and cognitive abilities

The relationship between cognitive measures and procedural learning on the SRTT was further explored using the random slopes for the last 600 trials, however none of the correlations survived corrections for multiple comparisons. In Appendix G Bayes factors and credible intervals for the bivariate correlations are presented.

Language and literacy. Against the predictions of the procedural/declarative model, there were no significant correlations between language and literacy and procedural learning for the TD group. Whilst there was a weak positive association between nonword repetition and procedural learning in session 3 for the dyslexia group (*r* = .31), Bayesian correlations did not support this (BF_{10} = 1.93)

Attention. Individual differences analyses for the TD group revealed significant correlations between procedural learning and attention suggesting that individuals with better attentional skills showed more evidence of procedural learning in session 3 (median for sessions 2, BF_{10} = 2.20, and 3, BF_{10} = 15.66). This association remained significant after accounting for explicit awareness (exclusion condition: *r* = -.37, *p* = .028). For the dyslexic group, attention abilities in the first session correlated positively with procedural learning in session 2 (median for session 1, BF_{10} = 3.38), indicating that, unlike the TD group, those with worse performance on the PVT showed more evidence of procedural learning. There was no evidence of an association between procedural learning and attention abilities as measured by the Conners scale (p < .05) for both groups.

Explicit awareness. For the TD group, procedural learning in session 1 was negatively correlated with explicit awareness in the inclusion condition (BF_{10} = 5.28) whilst procedural learning in session 2 (BF_{10} = 2.60) and 3 (BF_{10} = 10.87) was negatively correlated with explicit awareness in the exclusion condition. This suggests that participants with more explicit awareness of the sequence showed *less* evidence of procedural learning. There were no significant correlations between explicit awareness and procedural learning for the dyslexic group (*p*s > .05). Given these group differences in associations between explicit awareness and procedural learning, we also explored whether another meta-cognitive variable (i.e., enjoyment) would differentially associate with procedural learning. Indeed, for the TD group in sessions 1 (BF_{10} = 3.18) and 2 (BF_{10} = 2.51) there were positive correlations, suggesting that participants who enjoyed the SRTT more showed more evidence of procedural learning. There were no such correlations for the dyslexic group (*p*s > .05).

Table 5.10

Correlation matrix between procedural learning and cognitive measures

Measures		***TD group***			***Dyslexic group***		
		Procedural learning Session 1	Procedural learning Session 2	Procedural learning Session 3	Procedural learning Session 1	Procedural learning Session 2	Procedural learning Session 3
Age		-.34*	.27†	.06	-.11	-.05	-.24
Literacy	Word Reading	0.25	-.18	-.20	.05	.18	.04
	Nonword Reading	.20	.10	.002	-.06	.03	.07
	Spelling	.02	.12	.12	.10	.08	.18
Language	Vocabulary	.21	.26	.30	.13	.02	.10
	Nonword Repetition	.10	.16	.15	.03	-.07	.31*[a]
	Sentence Recall	.08	.20	-.13	.10	.29†	-.04
Nonverbal IQ	Matrix Reasoning	.01	.09	.24	.07	.01	.07
Attention	PVT median Session 1	-.08	.10	-.28†	.10	.35*[a]	.22
	PVT median Session 2	.01	-.12	-.34*	-.06	.09	.07
	PVT median Session 3	-.06	-.15	-.47**[a]	-.06	.07	.05
	PVT Reciprocal Session 1	.20	-.10	.29†	.21	-.08	-.03
	PVT Reciprocal Session 2	.06	.16	.33†	-.16	-.28†	-.14

	PVT Reciprocal Session 3	.26	.03	.46**[a]	-.08	-.15	-.18
	PVT tau Session 1	-.14	.02	-.39*[a]	.07	.08	-.04
	PVT tau Session 2	-.26	-.41*[a]	-.36*	-.10	-.25	.19
	PVT tau Session 3	.06	.001	-.17	.25	.18	-.05
	Conners scale	-.13	-.13	-.09	-.14	.13	.13
ARQ		.16	-.20	.24	-.04	.15	-.09
Explicit Awareness	Inclusion	-.41*[a]	-.08	.01	-.14	.01	-.01
	Exclusion	-.13	-.35*	-.45**[a]	.04	-.12	-.23
Enjoyment		.38*[a]	.35*	.22	-.06	-.02	-.03

Note. †p < .10; *p < .05; **p < .01; ***p < .001; [a] Correlations with Bayes factor equal or bigger than 3

H5: Assessing group differences in explicit awareness

Each generated sequence was coded for the number of triplets in common with the learnt sequence and compared against chance level, which was estimated to be 29.98 triplets, out of a maximum of 98.

Table 5.11

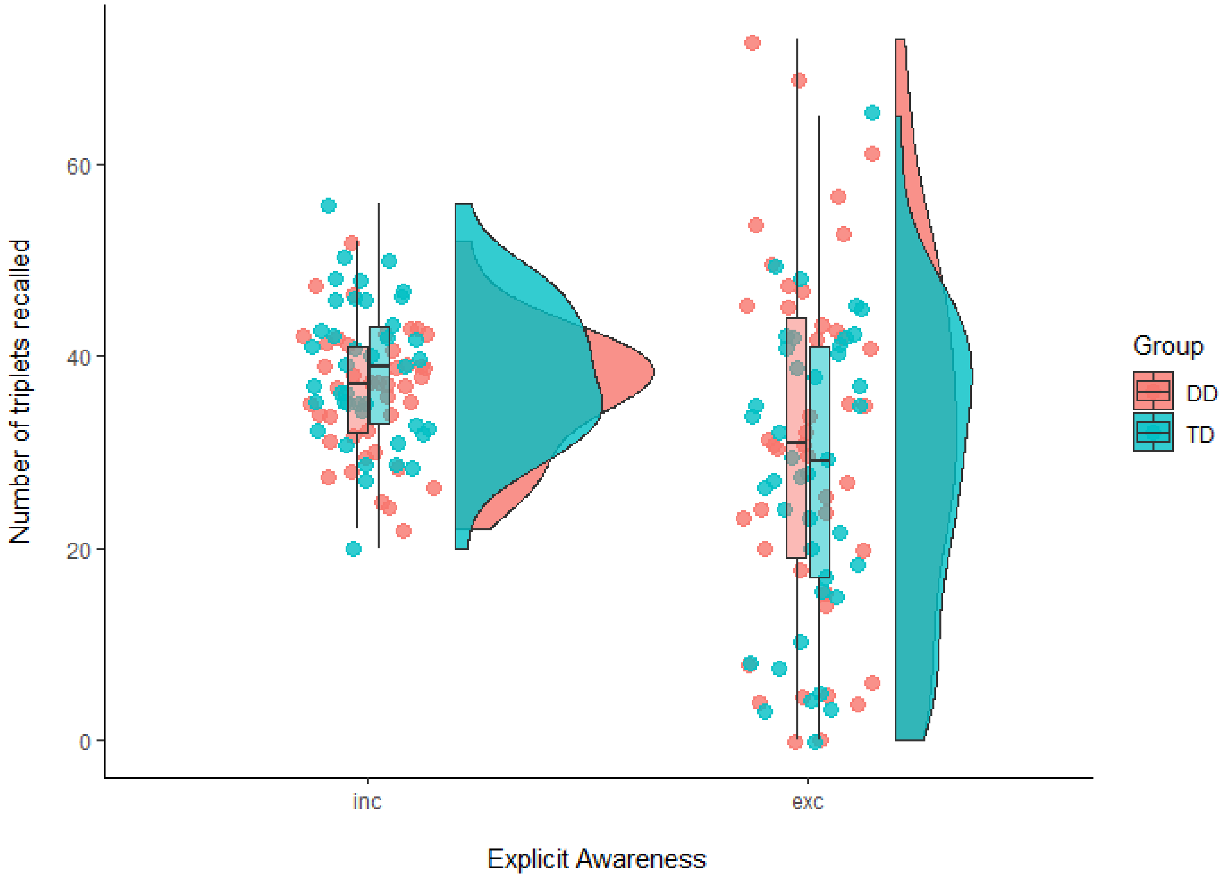
Means and standard deviations for the explicit tasks for each group

Triplets	***TD***			***dyslexic***		
	N	Mean	SD	N	Mean	SD
Inclusion	41	38.5	7.54	42	36.2	6.58
Exclusion	41	28.1	15.4	43	31.2	18.8

For both groups, recall of triplets was above chance in the inclusion condition (TD: *t*(40) = 7.21, *p* <.001; DD: *t*(41) = 6.12, *p* < .001), but not in the exclusion condition (TD: *t*(40) = -.78, *p* < .440; DD: *t*(42) = .43, *p* = .669). Since in the exclusion condition, unlike the inclusion condition, participants are asked to not generate the underlying sequence, this suggests some evidence of control over the explicit knowledge. A two-way ANOVA, with *Group* and *Condition* as predictors, revealed a significant effect of *Condition*, with the number of triplets generated in the inclusion condition significantly higher than in the exclusion condition (*F*(1, 163) = 13.93, *p* < .001). There were no significant differences between *Groups* (*F*(1, 163) = .04, *p* = .849), nor was there a *Group* x *Condition* interaction (*F*(1, 163) = 1.74, *p* = .190). This suggests that both groups showed similar levels of explicit awareness across conditions.

Figure 5.3

Plot of the distribution and probability density for the explicit awareness tasks under inclusion (right) and exclusion (left) conditions for dyslexic and TD participants



Discussion

This experiment is the first to examine the procedural learning abilities of adults with and without dyslexia in the context of task reliability over 3 learning sessions. Alongside taking a group level approach to examining differences in procedural learning, we also adopted an individual differences approach, comprehensively examining relationships between procedural learning and language, literacy, and attention. Both groups showed robust procedural learning effects and we found no support for the procedural deficit hypothesis at the group-level: There were no differences between groups in the magnitude or trajectory of the procedural learning effect within or across sessions. Furthermore, and compatible with our previous findings, there was only very limited evidence for a relationship between procedural learning and language/literacy, with only a significant and moderate correlation between procedural learning and nonword repetition in the dyslexic group. This could be at least partially explained by the poor psychometric properties of the SRTT, as again the procedural learning effect on the SRTT failed to meet psychometric standards in both typical and atypical populations (*r*s < .70). Numerically higher reliability in later sessions was observed in the TD group than in earlier sessions, replicating the pattern observed in our previous study. Unexpectedly, the opposite pattern was observed for the dyslexic group. Irrespective of the psychometric issues, we replicated the correlation between attention and procedural learning in both groups, suggesting that the relationship between attention and procedural learning is stronger, and more robust to the limitations of weak psychometric properties, than the relationship between procedural learning and language and literacy abilities.

As predicted, and replicating our previous research (Oliveira et al., submitted), we found overall higher split-half than test-retest reliability for the procedural learning effect in both groups, particularly in session 3. Crucially, the test-retest reliability was well below psychometric standards for both groups, indicating that the SRTT may not be suitable to be used to assess procedural learning in typical or atypical populations. Even though the test-retest reliability improved from earlier (sessions 1-2) to later sessions (sessions 2-3) in TD participants (akin to Oliveira et al., submitted), the opposite pattern was observed for the dyslexic group. Nonetheless, it is important to emphasise that these patterns were only numerical and did not represent a statistically significant difference across sessions.

It is notable that the retest reliability coefficients were lower here (TD group session 1-2: *r*s = -.02 - .07; TD group session 2-3: *r*s = .33 - .34; dyslexic group session 1-2: *r*s = .14 - .38; dyslexic group session 2-3: *r*s = .00 -.11) than in our previous study (session 1-2: *r* = .43; session 2-3: *r* = .60). Whilst it is unclear what factors account for this, these findings only further highlight the poor stability of the procedural learning effect on the SRTT, whereby designs which have yielded better reliability in the previous literature (e.g., West, Shanks, et al., 2021) will likely fail to produce the same result. Instead, it is possible that, irrespective of the experimental design, individual differences in confounding factors may primarily contribute to the reliability of the SRTT. Specifically, unlike in our previous study examining the psychometric properties of the SRTT in TD adults across 3 sessions, in this study individual differences in the age of the participants, their enjoyment in performing the SRTT and the degree of explicit awareness influenced the magnitude of the procedural learning effect. Whilst enjoyment showed a positive relationship with procedural learning, explicit awareness and age showed the opposite pattern. Thus, it is likely that the effect of these variables on procedural learning may have contributed to changes in the ranking order between sessions. The influence of these variables is not inherently problematic, as they could influence participants' performance in the same manner across sessions, thus preserving the ranking order. However, given the poor reliability and the changes in the correlations between sessions, it is unlikely that these variables exert the same degree of influence on each session and/or that their influence on each participant is consistent across sessions.

Whilst attention was not a predictor of procedural learning in the mixed model, the correlation between attention and procedural learning was replicated. The findings from the mixed model likely reflect low power, as there was only one score per session for the attentional abilities of the participants, thus a four-way interaction (*Probability* x *Session* x *Group* x *Attention*) was unlikely to emerge even if there was evidence for an effect of attention on procedural learning. Future research should overcome these limitations by recruiting more participants or using online measures of attention. Importantly, unlike our previous study with TD adults, attention correlated with procedural learning in later rather than earlier sessions. Thus, even though attention has been suggested to be required for procedural learning to occur by reducing the sensory input to the relevant information, this correlation emerges later than the procedural learning effect. Thus, potentially highlighting the importance of top-down mechanisms whilst performing the same repetitive and long task for the third time. This is especially relevant when participants are being tested in a naturalistic setting where distractors are more likely to occur. Furthermore, the relationship between attention and procedural learning in the third session cannot be explained by explicit awareness, as this association remained significant after accounting for explicit awareness in the last session. Finally, however, we cannot rule out the possibility that the timeline of the correlations between attentional abilities and procedural learning may simply reflect the poor psychometric properties of the SRTT, whereby attenuation would be more accentuated in earlier than later sessions in the TD group, and the opposite pattern for the dyslexic group.

Interestingly, despite comparable attentional abilities, the dyslexic group did not show correlations between procedural learning and attention in the last session, instead the only correlation emerged in session 2. It is unclear whether these differences are associated with the stages of procedural learning. Since we adopted different sequences on each session, despite their high degree of similarity, it is possible that later sessions may not represent later stages of procedural learning, per se. Furthermore, if later sessions were to reflect later stages of procedural learning, then performance on the third session would likely be more automatised and thus less reliant on attentional resources (Seger & Spiering, 2011). Yet, testing across sessions may not be required for later stages to be captured as evidenced by Lum et al. (2019), who observed within-session changes in the activation of attentional components, modulated by the procedural learning effect, which decreased throughout the session whilst performing patterned trials, which was interpreted as indicating some level of automatisation. Unfortunately, our offline measure of attention was not sensitive enough to capture these time sensitive changes. However, as evidenced in the Bland-Altman plots, there was a distinct pattern in the overall procedural learning effect across sessions for each group. Whilst the TD group showed more evidence of practice effects (i.e., changes in the participants' score due to practice on the test) from session 1 to 2 as evidenced by the difference in procedural learning effect in session 2 and 1 (difference = 5.84) than 2 to 3 (difference = -3.49), the dyslexic group showed the opposite effect, with more gains in later sessions (difference = 6.56) than earlier (difference = -1.00). Even though this agrees with the pattern of more stability for later sessions in the TD group and in earlier sessions for the dyslexic group, these reflect only numerical differences which were not captured in the linear mixed model. Thus, replications of the present study are required to fully understand the trajectory of procedural learning in these groups. One possibility to test in future research is whether the dyslexia group may require more sessions until performance starts to stabilise, potentially as a consequence of group differences in consolidation and/or interference. However, speculating on the mechanisms by which procedural learning in one session affects the other is beyond the scope of the present experiment.

Having established the poor psychometric properties of the SRTT, it is thus not surprising that there were no significant correlations between language and literacy and procedural learning for any of the groups, with the exception of nonword repetition in the dyslexic group. Whilst attenuation of the correlations between these measures may have occurred due to the poor reliability of the SRTT, it is somewhat surprising however that we have replicated the correlation between attention and procedural learning across studies, irrespective of task manipulations and sampling characteristics (e.g., in the presence/absence of an interstimulus interval, in lab and online testing, and in populations with and without dyslexia). Thus, the poor reliability of the SRTT cannot be the only reason why we have failed to find correlations with measures of language and literacy.

The absence of group differences on the SRTT is also at odds with the predictions of the procedural deficit hypothesis (Ullman et al., 2020; Ullman & Pierpont, 2005). One possibility is that the age group tested in this study (i.e., young adults) may have been able to compensate for procedural learning difficulties by relying on their declarative system. According to the procedural deficit hypothesis, the declarative memory system, if preserved, should compensate for the procedural deficits (Ullman & Pullman, 2015) and, given that the declarative system has a prolonged maturational trajectory compared to the procedural memory system, it is likely that its compensatory role becomes more efficient from childhood into young adulthood (Cycowicz et al., 2001; Mandler & Robinson, 1978; Ofen, 2012; Ofen et al., 2007). This is in line with the findings from previous meta-analyses comparing individuals with and without dyslexia on the SRTT (though this pattern was also observed for the visual artificial grammar learning task (van Witteloostuijn et al., 2017)), which have reported smaller differences between groups for older participants, yet this pattern was only significant for Lum et al. (2013). However, two main arguments oppose this view. Firstly, the dyslexic group showed comparable performance to the TD group on the SRTT which arguably would not have been achievable by relying solely on the declarative system. Specifically, populations with preserved declarative memory but damage to the neural structures that support procedural memory (e.g., individuals with Parkinson’s and Huntington’s disease) still show impaired performance on the SRTT, even when using the verbal version of the task (Westwater et al., 1998; Sommer et al., 1999; Smith and McDowall, 2006; Clark, Lum & Ullman, 2014; Doyon et al., 1997, 1998; Molinari et al., 1997; Shin and Ivry, 2003). Secondly, if dyslexic participants are indeed able to fully compensate for their procedural learning deficits on the SRTT, it is unclear why the dyslexic group included in this study has not been able to compensate as well with the declarative system for their literacy and phonological abilities. As previously reported (Lohvansuu et al., 2021), the dyslexic participants tend to show pervasive difficulties that persist into adulthood. In this study, this was evident as participants struggled with the reading and spelling accuracy tasks, even though these have been found to more effectively respond to remediation interventions than fluency and comprehension issues (S. E. Shaywitz et al., 2008). Thus, it is unlikely that dyslexic participants' comparable performance to TD participants is fully explained by the declarative memory compensatory mechanisms.

Despite the lack of support for the procedural deficit hypothesis here, it is possible that this result is task specific, especially given the poor psychometric properties. Whilst there is evidence suggesting that the SRTT is more closely tied to the neural structures underlying procedural learning than other procedural learning tasks (hence our reason for selecting this task here), it is still possible that the SRTT is not sensitive enough for detecting procedural deficits in adults with dyslexia, at least when only examining behavioural data. Instead, neuroimaging or electrophysiological studies may prove useful to detect whether despite similar behavioural performance, this is accomplished by relying on different neural mechanisms. This would also allow for a better understanding of whether more activation is observed in the declarative memory system (i.e., hippocampus) when dyslexic participants perform the SRTT compared to the TD group. It is also possible that the procedural learning impairment is modality or computation specific. Even though the procedural deficit hypothesis assumes that procedural memory is a domain-general ability, previous studies which have examined the procedural learning abilities of individuals with dyslexia across paradigms have found selective impairments in one or more tasks, but not all (Henderson & Warmington, 2017; J. H. Howard et al., 2006; Jiménez-Fernández et al., 2011; Rüsseler et al., 2006; Vakil et al., 2015; Vicari, 2005). This is further supported by the null to small correlations between procedural tasks which are thought to tap the same construct and use similar statistical patterns (e.g., visual and auditory nonverbal adjacent tasks: Siegelman & Frost, 2015), though we recognise that attenuation may have occurred due to the poor psychometric properties of these tasks. Thus suggesting, as proposed by Bogaerts et al. (2021), that procedural learning tasks should be selected based on their relevance to a specific impairment instead of being viewed as interchangeable since they likely tap different computations and dimensions of procedural learning.

Finally, whilst the present findings lend only very limited support to the role of procedural learning in language and literacy development and impairments, it would be premature to conclude that the procedural memory system does not influence language and literacy. Instead, these findings raise questions about the timing of this relationship. If procedural memory exerts more influence in earlier stages of language and literacy acquisition, its distal influence may be harder to capture at later stages when adequate performance on complex language and literacy tasks may rely on a broader array of abilities. Similarly, and according to the multiple deficit model (Pennington, 2006) which proposes that the cause of dyslexia is multifactorial resulting from the interaction between genetic and environmental risk and protective factors, it is possible that a procedural learning deficit represents a risk factor for dyslexia, which is only observed in some individuals with dyslexia. Thus, explaining the mixed findings in studies examining the procedural learning abilities of individuals with dyslexia.

To conclude, our findings from an adult sample do not provide support for the procedural deficit hypothesis as there was no evidence of a procedural learning deficit in this population across sessions. Similarly, we also observed only very limited evidence for a role of procedural learning in language and literacy in adults. Whilst this pattern may be partially explained by the poor psychometric properties of the SRTT, in light of the more consistent correlations for attention, it does raise the possibility that if a relationship is present, it may be less robust than predicted by the procedural/declarative model. Future longitudinal studies are needed to systematically determine the role of procedural learning across development. Critically, however, before embarking on such an endeavour more reliable tasks are required.

Chapter 6. General Discussion

The procedural deficit hypothesis, a neurobiological account put forward as an alternative to cognitive theories of DLD and dyslexia, proposes that an impairment in the procedural memory system may give rise to language and literacy impairments (Ullman, 2004; Ullman et al., 2020; Ullman & Pierpont, 2005). Whilst this hypothesis has gathered considerable interest and been continually examined over the last two decades, the results have been mixed. As indicated in the literature review in **Chapter 1**, meta-analytical evidence supports group-level differences between individuals with TD and those with DLD and dyslexia (Lum et al., 2013, 2014; West, Melby-Lervåg, et al., 2021), yet it is still unclear whether these group-level differences reflect differences in procedural learning or extraneous factors. Despite the results of these meta-analyses, there are also cases where the group differences have not been replicated, perhaps owing to methodological differences in the procedural learning tasks. Given the limited evidence from individual differences research for an association between language/literacy and procedural learning (Hamrick et al., 2018; Lammertink et al., 2020; West, Melby-Lervåg, et al., 2021), if indeed procedural learning does play a role in language/literacy difficulties, it is unclear whether it represents a continuous risk factor or whether performance only has to reach a threshold for adequate development of language and literacy abilities. In light of the poor psychometric properties of the SRTT (as confirmed consistently across five experiments in this thesis) which may lead to attenuation of the correlations between these measures, it is possible that the discrepancy between group-level meta-analytic findings and individual differences studies reflects the reliability issues. In this thesis, we sought to address these empirical questions directly, thereby further examining the claims of the procedural deficit hypothesis.

To achieve this endeavour, we conducted four behavioural experiments with adults with and without literacy difficulties, using both lab-based and online administration of a probabilistic SRTT, and complemented the empirical studies by conducting two meta-analyses (drawing from studies using various versions of SRTTs). All, but the first experiment, were pre-registered with all data and analyses scripts added to the Open Science Framework. By utilising this methodological approach, we examined the psychometric properties of the SRTT in adults with and without dyslexia, as a reliable measure of procedural learning is a necessary condition to guarantee the validity of the findings. As well as examining group-level differences in procedural learning, we also explored the concurrent associations between procedural learning and language/literacy proficiency in adults with and without dyslexia and DLD. Additionally, the experimental studies reported additional cognitive measures previously shown to relate to procedural memory, such as sustained attention (Arciuli, 2017; West, Shanks, et al., 2021). This final chapter provides a summary and synthesis of each of the key findings, before considering the broader implications.

Summary of experimental studies

Chapter 2

In Experiments 1-2 and the supplementary experiment we investigated factors that may affect the stability of the SRTT in adults aged 17 - 60 years. Experiment 1 examined whether the similarity of sequences learned at test and retest would impact the magnitude and stability of the procedural learning effect. Test-retest correlations were low regardless of sequence similarity (r < .25). Thus, in experiment 2 we adopted different, but highly similar sequences, in line with West, Shanks, et al. (2021). Furthermore, experiment 2 added a third session to examine whether individual differences in procedural learning would stabilise with further training, whereby later sessions would be expected to lead to less procedural learning gains than earlier sessions. There was a small (but nonsignificant) improvement in stability for later sessions (session 1-2: *r* = .43; session 2-3: *r* = .60), such that test-retest reliability between sessions 2 and 3 approached acceptable psychometric standards (*r* > .70). In the supplementary experiment, we aimed to more closely replicate the study design employed by West, Shanks, et al. (2021), which remains the only study in the literature to report a test-retest coefficient for the SRTT that was >.70, by adopting an interstimulus interval (ISI) of 250 ms and recruiting participants from a wider age range (18-60 years). Counter to our predictions, the test-retest reliability continued to be poor for both groups with and without an ISI (*r*s < .50), suggesting that other design features may have contributed to the superior stability observed by West, Shanks, et al. (2021). Crucially, the stability of procedural learning on the SRTT remained suboptimal in all conditions (*r*s < .70). We argue here that this poses a serious obstacle to the use of this task as a sensitive indicator of individual differences and, ultimately, theoretical advance. Unlike the consistent pattern of poor stability across sessions, within-session reliability was substantially and consistently better, ranging from poor to excellent depending on task design and the analytical methods adopted (Experiment 1: *r*s = .50-.71, Experiment 2: *r*s = .23 - .91, Supplementary experiment: *r*s = .52 - .93).

In addition to examining stability across three (as opposed to two) sessions each separated by one week, Experiment 2 also examined concurrent associations between procedural learning and language, literacy, and attentional abilities. Stronger correlations between these measures were expected for later sessions if the stability of the procedural learning effect was better at later sessions, based on the assumption that procedural learning at later sessions may provide a better reflection of an individual’s true procedural learning ability. Despite the improved reliability at later sessions, and in contrast to the predictions of the procedural/declarative model, there was only a significant and positive correlation between procedural learning in session 3 and vocabulary abilities (*r* = .39). There was also a significant association between attention abilities (as measured by the Psychomotor Vigilance task which was administered in the first session) and procedural learning in sessions 1 and 2. This relationship was further replicated in the supplementary experiment in the ISI group using a shortened version of the PVT and in a follow-up experiment for the noISI group.

It was concluded from this set of experiments that, although the procedural learning effect was robust across samples, settings and task characteristics, its stability across sessions was consistently below psychometric standards. Similarly, the relationship between attention and procedural learning was replicated across experiments, even though poor reliability often leads to underestimation of the correlation between constructs due to measurement error.

Chapter 3

The experiments in Chapter 3 complemented those in Chapter 2 by examining the test-retest reliability of the SRTT across existing studies published in the literature, in order to determine which manipulations may result in higher reliability. In this meta-analysis (N datasets = 7), comprising 719 participants, we confirm the "reliability paradox" (Hedge et al., 2018). The overall retest reliability (*r* = .30, 95% CI = [.18, .42]) of the robust procedural learning effect elicited by the SRTT was found to be well below acceptable psychometric standards. However, split-half reliability within a session was higher, with an overall estimate of *r* = .68, 95% CI [.57, .77]. There were no significant effects of sampling (participant age), methodology (number of trials, sequence type, inclusion of an interstimulus interval, version of the SRTT task) or analytical decisions (whether all trials were included when computing the procedural learning scores; using difference scores, ratio scores, or random slopes as an index of learning). Thus, suggesting that the SRTT, irrespective of manipulations, does not meet psychometric standards for adequate test-retest reliability.

Chapter 4

The ability to extract patterns from sensory input across time and space is thought to underlie the development and acquisition of language and literacy skills, particularly in subdomains marked by the learning of probabilistic knowledge. Thus, impairments to procedural learning mechanisms are hypothesised by the procedural deficit hypothesis to underlie neurodevelopmental disorders such as dyslexia and DLD In a meta-analysis comprising 2396 participants from 39 independent studies, the relationship between language, literacy and procedural memory as indexed by the SRTT was assessed across children and adults with TD dyslexia, and DLD. Based on the procedural/declarative model a positive relationship was expected between procedural memory and language and literacy measures for the typically developing group; however, no such relationship was observed. We also did not observe a relationship between procedural memory and language and literacy for the disordered groups. Also counter to the procedural deficit hypothesis, the magnitude of the relationship between procedural learning and grammar and phonology did not differ between TD and DLD groups, nor between the TD and dyslexic groups on reading, spelling, and phonology. Whilst lending little support to the procedural/declarative model and the procedural deficit hypothesis, we consider that these results may be the consequence of poor psychometric properties of the SRTT as a measure of procedural learning.

Chapter 5

Experiment 4 assessed the predictions of the procedural deficit hypothesis by comparing the performance of individuals with and without dyslexia on this task on three separate occasions. This study also represents the first examination of the stability of procedural learning in a dyslexic population. Even though a large number of studies have examined procedural learning in dyslexia, the results have been inconsistent. Importantly, almost all the existing studies in the literature only assess procedural learning at a single time point, and thus may not be capturing the most stable stage of learning. Therefore, this study, using a similar design to Experiment 2 which resulted in the best observed level of reliability in our studies, examined group level differences and individual differences between procedural learning and literacy/literacy and attention abilities over three sessions. Contrary to the predictions of the procedural deficit hypothesis, there were no significant differences between dyslexic and TD groups on the magnitude of the procedural learning effect in any of the three sessions. Furthermore, both groups recalled more triplets of the underlying sequence than chance level in the inclusion condition (where they were asked to generate the sequence) but not in the exclusion condition (where instead they were asked to not generate the sequence), thus revealing some control over their explicit knowledge. Both groups showed comparable evidence of explicit awareness of the sequence being learned, thus any differences observed between groups are less likely to be accounted for explicit awareness.

Individual differences analyses with standardised measures of language and literacy showed that procedural learning only correlated significantly with nonword repetition (r = .31) for the group with dyslexia. On the other hand, attention was positively associated with procedural learning for both groups, replicating the findings from the previous experiments reported in this thesis (rs between -.34

to -.47). Finally, for the TD group, enjoyment and procedural learning were significantly and positively correlated (rs between .35 and .38). There was also a significant and negative relationship between procedural learning and explicit awareness under inclusion and exclusion conditions (*r*s between -.35 and -.45).

Taken together these findings do not lend support to the view that individuals with dyslexia show an impairment in the procedural memory system, and there was also only very limited evidence for a relationship between procedural learning and language/literacy in the typical adult population. However, there does seem to be a robust and replicable association between sustained attention and the size of the procedural learning effect, in both typical and atypical populations, which is considered in more detail below.

We now go on to consider a number of key themes that arise from these findings and highlight important directions for future research. We start by reflecting on the measurement of procedural learning, examining factors which may be associated with its poor stability, and then review the evidence from our experimental work for the procedural/declarative model and procedural deficit hypothesis. Finally, we conclude by considering the clinical implications of these findings and future research avenues.

Measuring procedural learning

A central issue here pertains to the measurement of procedural learning. Despite producing robust procedural learning effects, the SRTT showed consistently poor psychometric properties, particularly with respect to its stability across time. In Chapter 2, we investigated whether manipulating task and design characteristics such as the levels of similarity between sequences at test and retest, adding a third session, and the presence/absence of an ISI improved the psychometric properties of the SRTT. Whilst split-half reliability was often observed to meet the threshold for adequate reliability (i.e. r >.70; Burlingame et al., 1995; Nunnally & Bernstein, 1994), test-retest reliability was consistently below acceptable psychometric standards. These results were confirmed in the meta-analysis, irrespective of sampling, methodology and analytical decisions. Thus, the issue of reliability calls into question the suitability of the SRTT as an index of procedural learning for individual differences research.

Importantly, the issue of reliability of procedural learning tasks is not specific to the SRTT, as other measures of procedural memory have also been found to show poor reliability (e.g., artificial grammar learning: Kalra et al., 2019; probabilistic classification task: Kalra et al., 2019; Hebb task: West et al., 2018; auditory and visual statistical learning tasks: Arnon, 2020). Beyond this, the issues with reliability are not specific to procedural memory, with similar findings reported for other classic, widely-used experimental paradigms in cognitive psychology (e.g., Stroop task, Flanker task: Haines et al., 2020; Hedge et al., 2018; von Bastian et al., 2020). This phenomenon is referred to as the "reliability paradox" (Hedge et al., 2018), where experimental paradigms known for eliciting robust effects fail to capture stable individual differences. The reliability paradox is thought to be a consequence of the use of experimental tasks in individual differences research which have been designed to reduce variability between individuals to ensure that the phenomenon of interest is captured. Unfortunately, this reduction in between-subject variability has consequences for individual differences as it limits the ability of a test to differentiate between individuals (Hedge et al., 2018).

Relatedly, the use of difference scores as measures of procedural learning has thus been considered problematic given that subtracting the improbable and probable trials further reduces the variance between subjects (Enkavi et al., 2019). However, the use of difference scores is insufficient to explain the superiority of within-session stability, compared to stability across sessions. To attempt to address this, alternative methods for computing procedural learning have been explored (e.g., ratio scores which account for the individuals' overall speed and random slopes which integrate information at the individual and group level); however, none of the methods used thus far have consistently produced reliable results. This should not be taken as an indication that procedural learning on the SRTT is inherently unreliable; instead, it may call for more complex statistical models that are better able to isolate the effect of interest from measurement error. Whilst the hierarchical models adopted in our experiments and in Lammertink et al. (2020) and van Witteloostuijn et al. (2021) led to show improvement in reliability as these are more suitable for accounting for trial noise, which has been found to attenuate between measures (Rouder & Haaf, 2019, 2021), they have modelled each session independently. Instead, future research should aim to explore hierarchical modelling as adopted by Haines et al. (2020) for assessing the reliability of the SRTT, as this has led to considerable improvements in reliability in similar tasks (e.g., Stroop task, Flanker task). Yet, as highlighted by von Bastian et al. (2020) in the context of attention, when examining the magnitude and reliability of procedural learning, we often ignore the speed-accuracy trade-offs. Whilst it is possible that individual differences in this trade-off between speed and accuracy are unrelated to procedural learning, they likely add noise to the measurement of the construct of interest. Thus, it may prove useful to extend previous work by examining both RTs and accuracy, particularly if the speed-accuracy trade-off differs between participants and is expected to change from session to session.

As previously mentioned, the issue of reliability is vital for individual differences research, as poor psychometric properties have been found to attenuate correlations with other measures (Enkavi et al., 2019; Hedge et al., 2018; Rouder et al., 2019). In small samples, however, measurement error may exert the opposite effect (Loken & Gelman, 2017), which can be especially problematic in light of the small samples in the field of neurodevelopmental disorders (as evidenced in our meta-analysis) and the bias for significant findings to be published. First, poor reliability limits the validity of individual differences studies examining the predictions of the procedural/declarative model as, even if there is a true association between procedural memory and language and literacy abilities, it will likely be attenuated due to measurement error., This may potentially explain the mixed findings both in TD populations and disordered populations. Second, weak correlations among different tasks thought to index procedural memory (Arnon, 2020; Kalra et al., 2019; Siegelman & Frost, 2015; West et al., 2018) have led researchers to question unitary accounts of procedural memory, in support of more componential views (Arciuli, 2017). Yet, it is unlikely that correlations between these measures would emerge, even if they capture the same underlying construct given that the degree of attenuation is impacted by the poor reliability of both measures (Rouder et al., 2019). Thus, even though our efforts to understand the impact of some experimental decisions have led to limited improvements in reliability, it is crucial for the psychometric properties of experimental paradigms to be adequately inspected and, as suggested by Parsons et al. (2019), reporting of the reliability estimates should be standard practice. Not only would this practice of reporting reliability estimates allow the scientific community to more adequately assess the validity of findings, but it would also allow for meta-analytical work to examine which factors contribute to better reliability. Potentially due to the limited number of current studies which have analysed the psychometric properties of the SRTT, our findings were, unfortunately, inconclusive in identifying which, if any, factors could improve reliability. However, based on the simulation work presented in chapter 2, further consideration should be given to the ratio between probable and improbable trials, as a small number of improbable trials will result in a noisier estimate. Furthermore, whilst the work presented in this thesis has focussed on the SRTT, it may prove useful to also comprehensively examine the psychometric properties of other procedural memory tasks in parallel as this may highlight shared characteristics, which may increase our understanding of procedural learning as construct.

One such pattern that would be usefully explored across tasks is the consistently superior split-half reliability (relative to retest reliability) which has also been observed in previous studies using the SRTT (e.g., West et al., 2018, 2021), and also in measures of procedural learning (Hebb task: e.g., Bogaerts et al., 2018; West et al., 2018; visual statistical learning task: e.g., Arnon, 2019). It is still unclear which factors may contribute to this disparity between split-half and test-retest reliability. As previously mentioned, it is likely to be influenced by the temporal distance between trials, since split-half reliability looks at the correlations between odd and even trials, which are temporally more proximal than correlations across sessions (Wagenmakers et al., 2004). This could reflect practice effects (i.e., changes in procedural learning as learning progresses), or the impact of extraneous/interfering variables, whereby temporally close trials are more likely to reflect similar cognitive and emotional states.

The role of practice effects across sessions is an important issue both in the consideration of split half versus retest reliability, and also in our interpretation of the trajectory of procedural learning that we measure here. Even though we aimed to reduce the impact of practice effects on the stability of the SRTT by adding a third session, practice effects were still observed in later sessions, despite the use of alternate sequences. Given the adoption of distinct sequences at each time point, it is unclear whether these changes in performance across sessions reflect practice effects or later stages of procedural learning. Based on the findings from the first experiment, where we observed a positive relationship between similarity and procedural learning in session 2, it appears that the knowledge acquired in one session impacts the learning of subsequent sessions. However, the mechanisms which lead to this impact are beyond the scope of this thesis. Having adopted highly similar sequences, it is possible that shared patterns of the sequence between sessions may have been consolidated and automatised across sessions, whilst new inconsistent patterns were learnt alongside. This may suggest that more sessions are necessary for learning to become saturated, as has been shown in working memory tasks where practice effects are still observed until the 7th session, despite no longer being significant after the 4th administration (Scharfen et al., 2018). On the other hand, if automatisation in the procedural memory system can occur without offline consolidation, it would be possible for automatisation of the procedural learning to occur within a single session if given enough practice, thus requiring only a follow-up session. However, it is likely that due to the repetitive nature of the SRTT, participants would start to disengage from the task. Stark-Inbar et al. (2017) analysed the performance on an alternating SRTT which lasted 60 minutes and included 3825 trials per session. The test-retest reliability for this task was .46. Yet, this encompassed all trials, so it is unclear whether the stability of the procedural learning increased in later blocks. Future research may aim to explore the reliability of the SRTT across multiple administrations to determine whether once practice effects fail to occur, the stability of the SRTT increases. Yet, we recognise that, even if the reliability of the SRTT becomes adequate after 3+ administrations, its suitability for research and clinical purposes would still be questioned, as it would not be feasible for a test to have to be administered so many times to achieve a reliable score.

Finally, individual differences research assumes that there are stable differences between individuals in the construct of interest which may influence individuals' accumulated experience/learning over the long term, which, if adequately captured, would likely result in adequate stability. However, it is possible that the poor reliability of the procedural learning effect does not reflect a problem with the paradigm. Instead, this may indicate that there is insufficient variability in the procedural learning effect, as it may be sufficient for a minimum level of procedural learning ability to facilitate acquisition of cognitive and motor skills and habits. Therefore, the magnitude of the difference scores may carry only limited meaning, instead it may be more important whether the individual is able to extract any knowledge from the task, irrespective of its magnitude. This is in line with A. S. Reber (1989) proposal that procedural learning due to being evolutionarily old differs substantially from declarative memory as it is expected to show little between subject-variability. Following from this, if individuals do not differ enough from one another then measurement fluctuations will lead to substantial changes in ranking order. We return to this issue when examining the role of explicit awareness on the SRTT and the relationship between procedural learning and language/literacy.

The role of attention on the SRTT

The role of attention in the SRTT is still under debate. Even though procedural learning is thought to be independent from attention (Frensch et al., 1998; Heuer & Schmidtke, 1996; Schmidtke & Heuer, 1997), it is unclear whether this independence is stage dependent. Namely, there is some evidence suggesting that earlier stages of procedural learning may require attention for learning to occur, whilst once the performance becomes automatised it becomes less reliant on attention (Seger & Spiering, 2011). This is line with evidence showing impaired procedural learning when performing the SRTT in a dual-task condition (Röttger et al., 2019; Thomas et al., 2004) and the relationship between attention and procedural memory has also been previously observed (Sengottuvel & Rao, 2013; West, Shanks, et al., 2021; Franklin et al., 2016). Our experiments confirm the positive association between procedural memory and attention abilities, as measured by the Psychomotor Vigilance task, as this relationship was consistently observed across studies (Chapters 2 and 5) both in the typically developing and dyslexic populations. The role of attention in the task has been shown to be stage dependent, whereby a decrease in the attentional demands is thought to occur once the learning on the SRTT becomes more automatised (Lum et al., 2019; Thomas et al., 2004). In our experiments, we observed a significant relationship between attention and procedural learning in sessions 1 and 2 in experiment 2 in Chapter 2 and in sessions 2 and 3 in Chapter 5. Thus, it appears that performance on the SRTT was still dependent on attention in later sessions. However, given that participants are required to learn a new sequence, despite the high levels of similarity, in each session, it is possible that participants failed to demonstrate evidence of automatisation on the SRTT.

Alternatively, given our design (i.e., involving a large number of trials within each session), it is possible that later blocks within a single session may be capturing later stages of procedural learning as observed by Lum et al. (2019). However, given the reliability issues of the SRTT, analysing the correlations between attention and procedural memory for each block would likely lead to substantial attenuation of the correlation, as measurement error tends to decrease with the number of trials (Rouder et al., 2019; Rouder & Haaf, 2019). Given these limitations, future research should aim to use online measures of attention, such as pupillometry measures (Unsworth & Robison, 2016, 2018; S. Zhao et al., 2019), to examine the changes in tonic pupil size as learning progresses in the SRTT. Unfortunately, due to the pandemic restrictions, we were unable to conduct this planned study.

Given the possible impact of participants' attention abilities on the magnitude of the procedural learning effect, we also examined its influence on the reliability of the SRTT. To achieve this, ex-Gaussian analyses were performed on the Psychomotor Vigilance task as this method has been found to better capture intraindividual variability of response times, which have been considered a potential endophenotype for attentional difficulties. These exploratory ex-Gaussian analysis in Experiment 2 (Chapter 2) revealed that the test-retest reliability of individuals in the lower intra-individual variability group showed significantly better reliability for later sessions than that of individuals in the higher intra-individual variability group. Numerically higher reliability for the procedural learning effect for those who showed lower variability (lower tau parameter) on the PVT was also observed in the supplementary experiment (Appendix E). This suggests that individuals with more variability in their response times on the attention task show less stability in the procedural learning effect, possibly due to fluctuations in attention throughout the task, which may lead to disparities in the procedural learning effect between sessions. As previously mentioned, fluctuations in attentional states would be expected to also impact the within-session stability of the procedural learning effect, yet this should occur to a lesser degree since it would likely be captured by both halves of the task.

To conclude, the role of attention in the SRTT is still poorly understood, yet there is still considerable evidence suggesting a positive association between attention and procedural learning. Future research is required to further understand the trajectory of attention during the SRTT as most evidence has been gathered from offline measures of attention (cf. Lum et al., 2019; Thomas et al., 2004). A more continuous measure of attention which allows for the detection of fluctuations in attention during the SRTT will also help elucidate its effect on reliability. Finally, given the role of attention in language and literacy development and disorders it is crucial to determine whether the deficits in procedural learning cannot be accounted for by attentional difficulties. This is especially relevant given the larger group differences in procedural learning in children than adults, which may at least be partially explained by the fact that children's attentional abilities are still under development (McAvinue et al., 2012). Thus, longitudinal studies may prove useful in revealing the direction of the relationship between attention and procedural memory, as well as elucidating the construct of procedural memory, and to what extent attention and procedural memory are overlapping constructs.

The role of explicit awareness on the SRTT

The issue of explicit awareness has not been a primary focus in this thesis, yet it is worth reflecting on, as there have been some suggestions that explicit awareness may impact the reliability of the SRTT. Whilst we observed evidence of explicit knowledge of the sequence across experiments, as expected after extended training (Shanks et al., 2005; Wilkinson & Shanks, 2004), we only observed an association between explicit awareness (in both inclusion and exclusion conditions) and procedural learning in Chapter 5, which was moderate and in a negative direction. This is contrary to the remaining experiments which showed no evidence of a relationship between these measures. The negative correlation seems to be at odds with Cleeremans and Jiménez (2002) suggestion that the emergence of explicit awareness in the SRTT occurs as the representations of the underlying grow stronger, which would likely indicate that those with better procedural learning would potentially show more evidence of explicit awareness. Furthermore, this evidence is insufficient to determine whether the emergence of explicit knowledge had a negative effect on the stability of procedural knowledge. That said, it is possible that as individuals develop explicit knowledge of the sequence, the strategies they adopt might change. Explicit awareness of the sequence is not inherently problematic for the stability of the procedural learning effect, as long as its impact is similar across participants, thus preserving the rank order between test and retest. However, this is unlikely to be the case given the high degree of between-subject variability in explicit awareness observed in our experiments. Furthermore, the emergence of explicit knowledge has been shown to vary depending on participants' characteristics such as age (Verneau et al., 2014) and sleep architecture (Kirov et al., 2015).

Finally, given previous findings that declarative tasks show consistently higher levels of reliability than procedural measures (Kalra et al., 2019; West et al., 2018), it could prove useful to examine the test-retest reliability of measures of explicit awareness. As previously mentioned, the better reliability of declarative tasks may be related to higher inter-individual variability than procedural learning (Hedge et al., 2018; Reber, 1989), making implicit measures of learning on the SRTT more sensitive to measurement error. Furthermore, unlike procedural learning in the SRTT which is measured as a difference between conditions, explicit awareness is instead computed as the number of recalled items which match the underlying sequence. Given considerable evidence demonstrating the unreliability of difference scores (Cronbach, 1990; Hedge et al., 2018; May & Hittner, 2003), even though they are insufficient to explain the pattern of adequate split-half reliability observed in this thesis, it is likely that this would also contribute to better reliability. Current explanatory theories on the emergence of explicit knowledge in the SRTT suggest that either explicit awareness develops as the result of the stronger representations of the underlying sequence (Cleeremans & Jiménez, 2002) or as a result of the process of searching for a cause for changes that occur during task performance (e.g., variations in accuracy when switching from sequenced to random trials) (Frensch et al., 2003; Rünger & Frensch, 2008). Thus, considering that both accounts suggest that procedural memory is a requirement for explicit awareness, and that clinical studies with individuals with Parkinson's and Huntington's disease show impaired performance on the SRTT despite preserved declarative memory (Westwater et al., 1998; Sommer et al., 1999; Smith and McDowall, 2006; Clark, Lum & Ullman, 2014; Doyon et al., 1997, 1998; Molinari et al., 1997; Shin and Ivry, 2003), this may be useful to determine the validity of the procedural deficit hypothesis. Alternatively, it is likely that the SRTT is not a process-pure measure of procedural learning, but that both procedural and declarative memory systems are involved when performing the SRTT (Sun et al., 2005). The declarative and procedural memory systems have been found to work cooperatively, independently and in competition, yet the mechanisms underlying their engagement are still poorly understood, with some evidence suggesting that task demands play a role in the relative weight of each system when performing a task (Foerde et al., 2006). In the context of the SRTT however, there is some evidence demonstrating that declarative memory may either compete for resources or have an inhibitory effect on procedural learning (Galea et al., 2010). This is suggested by evidence showing that suppression of the declarative memory system (e.g., repetitive transcranial magnetic stimulation: Ambrus et al., 2020; Galea et al., 2010; long-term effects of alcohol: Virag et al., 2015; cognitive fatigue: Borragán et al., 2016; and hypnosis: Nemeth et al., 2013) may enhance procedural learning on the SRTT. If, indeed, both systems interact whilst performing the SRTT, researchers should consider the impact of declarative learning abilities when comparing individuals with and without language and literacy impairments, as these populations, contrary to the predictions of the procedural deficit hypothesis, have often been found to show impairments in the declarative memory system (Lum & Conti-Ramsden, 2013; Reda et al., 2021).

Procedural learning and language and literacy proficiency

The procedural/declarative model postulates that procedural memory, that is, the sensitivity to regularities from sensory input, is implicated in the development of phonology, syntax, and morphology, because these aspects of language are characterised by sequential and statistical information (Ullman, 2001a, 2001b, 2004, 2016b, 2016c; Ullman et al., 2020). Consequently, given the underlying role of procedural learning in language and literacy, impairments in this long-term memory system have been hypothesised to cause dyslexia and developmental language disorder, characterised by particular difficulties in phonology and morphosyntax respectively (procedural deficit hypothesis) (Ullman, 2001a, 2001b, 2004, 2016b, 2016c; Ullman et al., 2020). Despite considerable research, mixed findings have been obtained both at the group-level when comparing individuals with and without dyslexia and DLD in the SRTT (e.g., Lum et al., 2013, 2014; West, Melby-Lervåg, et al., 2021), and in individual differences research (e.g., Lammertink et al., 2020; West, Melby-Lervåg, et al., 2021). Even though individual studies show a mixed pattern, meta-analyses have thus far consistently reported group differences in procedural memory between these groups, thus supporting the procedural deficit hypothesis (Lum et al., 2013; West, Melby-Lervåg, et al., 2021). On the other hand, meta-analyses examining the association between procedural learning in the SRTT and language and literacy have failed to demonstrate evidence for the procedural/declarative model (Lammertink et al., 2020; West, Melby-Lervåg, et al., 2021). One notable exception is a meta-analysis by Hamrick et al. (2018), which reported an association between grammar and procedural learning in the SRTT in L1 children (first language) and in L2 adults with high linguistic proficiency.

The findings by Lammertink et al. (2020) and West, Melby-Lervåg, et al. (2021) are in agreement with the pattern of results from our experiments, as we have also found little evidence for a relationship between language and literacy and procedural memory across individual studies (chapters 2 and 5). Specifically, we observed a correlation between vocabulary and procedural learning in experiment 2 for the TD group and an association between nonword repetition and procedural learning in experiment 6 for the DD group. According to the procedural/declarative model (Ullman, 2001a, 2004, 2016a), arbitrary information such as the mental lexicon is thought to be acquired primarily through the declarative memory system. Thus, the association between vocabulary and procedural learning is not aligned with the predictions of the procedural/declarative model. Speculatively, it is possible that this correlation reflects the contamination with explicit awareness, as almost half of the participants had become aware of the sequence by the third session. If that were the case, then this finding is not necessarily inconsistent with the procedural/declarative model. However, the negligible correlations between procedural learning and explicit awareness cast doubt on this hypothesis.

The association between nonword repetition abilities and procedural learning for the dyslexic group, on the other hand, support the procedural deficit hypothesis. However, they emerge in the absence of correlational evidence from the TD group, casting some doubt as to whether this evidence is sufficient given the unclear predictions from the procedural/declarative model for disordered populations. As previously mentioned, the procedural/declarative model postulates that the same pattern of association between language/literacy and procedural memory is expected to occur in disordered groups, unless the declarative memory system has been able to compensate for the language and literacy deficits experienced by these populations. In such cases, it is possible that null correlations between procedural learning and language/literacy would be observed as procedural learning would fail to explain the variability in the language/literacy abilities. This same issue is encountered in our meta-analysis, which examined the role of language/literacy and procedural memory in the SRTT. Here, we observed no significant correlations between the SRTT and language and literacy in any of the groups, yet moderate to high disattenuated correlations for the DLD and dyslexic groups were observed between procedural learning and phonology (r = .45) and spelling (r = .53), respectively. In light of the small to negligible correlations between procedural learning and language/literacy abilities, including those disattenuated for poor reliability, for the TD group, it is unclear whether this evidence lends support to the procedural/declarative model.

However, it is also unclear from the procedural/declarative model, despite the assertion that individuals with better procedural learning abilities would be expected to show better language and literacy skills, whether a linear relationship between these variables is to be expected. Specifically, given (Reber, 1989) characterisation of procedural learning as showing lower inter-individual variability, this may suggest that variability may only be found at the extremes of procedural learning abilities. Given that between subject variability is necessary for discriminating between individuals (Hedge et al., 2018), correlations between these variables would more likely occur in disordered than TD populations. Therefore, despite some evidence supporting that individual differences in procedural learning may be related to language and literacy abilities, this evidence calls for more testable hypotheses from the procedural/declarative model.

Group-level comparison: Procedural learning in dyslexia

We further examined the predictions of the procedural deficit hypothesis by comparing the performance of individuals with and without dyslexia in a probabilistic SRTT. Similar to Bennett et al., (2008), Gabay et al. (2012a, 2012b) and Henderson & Warmington (2017), we found no evidence of poorer procedural learning abilities in the dyslexic groups across sessions. However, this finding is at odds with the findings from the meta-analyses conducted by Lum et al. (2013) and West, Melby-Lervåg, et al. (2021), which report procedural learning impairments in dyslexia. Crucially, both meta-analyses observed smaller differences between groups for adults than children, even though this pattern was only significant for Lum et al. (2013). Thus, it is possible that, as proposed by Lum et al., (2013), in older participants the declarative memory system may be able to compensate for their procedural deficits. The adoption of a probabilistic SRTT may also have contributed to the null findings observed in the present study as West, Melby-Lervåg et al. (2021) reported smaller group differences for alternating SRTTs (which are similar to probabilistic tasks), compared to deterministic SRTTs. Although our use of a probabilistic task means we cannot make a direct comparison between our results and the previous findings reported in West, Melby-Lervåg et al. (2021), we would have expected to find a similar pattern of diminished group differences due to the shared probabilistic nature of these versions. This may suggest that individuals with dyslexia have more marked difficulties meeting the computational demands of the deterministic SRTT, as this version mostly captures sequential learning, whilst probabilistic and alternating tasks tap into both sequential and statistical learning. That is, if the underlying computational deficit in dyslexia is in sequential learning, but not statistical learning more generally, then the group differences should indeed appear for deterministic but not alternating or probabilistic SRTTs, consistent with the pattern reported by West, Melby-Lervåg, et al. (2021)**.** Whilst it is unclear whether the group-level findings would have been different if a deterministic SRTT had been adopted, this does not seem to be the case in adult populations given previous null findings in this age range (e.g., Gabay et al., 2012a, 2012b; Henderson & Warmington, 2017). Furthermore, previous evidence (Kalra et al., 2019; Stark-Inbar et al., 2017) suggests that this task shows poorer psychometric properties than probabilistic versions, and thus would be less well-equipped to examine the relationship between procedural learning and cognitive abilities. Nonetheless, it would prove beneficial for future research to directly contrast children's performance on sequential and statistical learning tasks, to determine whether the locus of the difficulty lies in sequential learning.

Beyond distinctions in computational demands, it is possible that differences in other factors, such as the severity of the neurodevelopmental disorders and presence or absence of comorbidities, may impact group differences. In the context of our experiment, contrary to previous studies which have often failed to control for comorbidity, individuals with dyslexia with comorbid disorders were excluded from the research study. Ullman et al. (2020) proposed that abnormalities of procedural memory and its underlying circuitry may contribute not only to language and literacy difficulties, but also to speech disorders. Furthermore, the authors further did not exclude the possibility that other neurodevelopmental disorders such as attention deficit hyperactivity disorder and autism spectrum disorder may also be at least partly explained by an impairment in the procedural memory system. Thus, whilst evidence for procedural learning impairments across populations may indicate that these neurodevelopmental disorders share risk factors, it is also possible that unreported comorbidities between disorders account for some of the mixed findings, whereby only one or some of these disorders is characterised by procedural memory impairments.

Relatedly, it is possible that variability in inclusion criteria may also account for the mixed findings, as West, Melby-Lervåg, et al. (2021) reported that the severity of decoding difficulties did predict the magnitude of group-level differences in procedural learning between the TD and disordered groups with dyslexia and DLD. Furthermore, in the context of DLD, a recent study by Krishnan et al. (2021) reported differences in the basal ganglia only in a subgroup of children who showed the poorest performance on a grammatical task. Whilst we recognise that there is not a direct correspondence between neural and behavioural deficits, only future research will be able to speak to whether deficits in the cortico-basal ganglia-thalamocortical circuitry have a directly causal role in language and reading deficits or whether, in line with the multifactorial view of developmental disorders (Pennington, 2006), they represent an additional risk factor for these disorders, which combined with other risk factors might result in more severe manifestations of dyslexia and DLD. Thus, procedural memory deficits may not be necessary or sufficient to cause the language and literacy deficits observed in dyslexia and DLD and may only be present in a subset of cases (Ramus & Ahissar, 2012).

Individual differences: The relationship between procedural learning and language/literacy

The absence of correlations between procedural learning and language/literacy may reflect attenuation due to the poor psychometric properties of the SRTT. On the other hand, it is also possible that, given the magnitude of the disattenuated coefficients (which take account of the poor reliability of individual measures), there is no, or only a small, true correlation between these measures. Yet, as observed by Rouder et al. (2019), disattenuation of correlations often results in under- or overestimations of the true correlation, thus it is possible that even disattenuated correlations fail to capture the true association between language/literacy and procedural learning. Assuming that the disattenuated associations are a better representation of the true association between language/literacy and procedural learning, and thus that there is only limited evidence for the procedural/declarative model in adults, this should not be prematurely interpret as indicating that there is no role for procedural memory in language/literacy across development. Instead, it may still be possible for procedural memory to be associated with earlier stages of language and literacy development. Support for this hypothesis comes from the meta-analysis by Hamrick et al. (2018), which observed that in children (first language) grammatical abilities positively correlated with measures of both declarative and procedural memory, whilst in adults (second language learners) in the higher proficiency group, grammar positively correlated with procedural, but not declarative memory. In the low proficiency group, declarative memory correlated with grammar, but not with procedural learning. Thus, if the language disorders in dyslexia and DLD are thought to result from immature language systems (e.g., Bishop and Snowling, 2004), more equivalent to younger learners, it is possible that the pattern of findings for these populations would more closely resemble those observed by Hamrick et al. (2018) in children or adults with low proficiency. Unfortunately, due to the absence of declarative memory tasks in our experiment, we cannot speak to the performance in those measures. Furthermore, whilst the findings from Hamrick et al. (2018) do not explain the absence of correlations between procedural learning and language/literacy measures in adults, they do suggest that there are temporal dynamics involved in the relationship between these measures which would benefit from being investigated in a longitudinal design.

A correlation between procedural memory and language/literacy proficiency is also expected only if a valid index of the individual's underlying ability is available. In the case of procedural memory, despite some evidence supporting a linear relationship between language/literacy and learning on the SRTT (e.g., Hamrick et al., 2018), it is unclear whether a numerical difference in the procedural memory effect is meaningful and should therefore translate into a difference in language/literacy abilities. As argued by Schmalz et al. (2021), given the involvement of the procedural memory system in a wide range of cognitive functions such as object and scene perception (e.g., Coppola et al., 1998; Lauer et al., 2018) and social cognition (e.g., Monroy et al., 2017; Norman & Price, 2012; Ruffman et al., 2012), it is unlikely that impairments in procedural memory have a deterministic effect on behaviour. After all, if this were the case, children with DLD and dyslexia would likely show broader impairments in cognitive abilities than those predicted by the procedural deficit hypothesis, such as in visual object and scene perception (Bogaerts et al., 2021) and in social cognition (e.g., Monroy et al., 2017; Norman & Price, 2012; Ruffman et al., 2012). If, instead, procedural memory is viewed from a componential perspective, it is possible that the computations required to adequately learn on the SRTT are not involved in language and literacy abilities, at least not in adulthood. However, it is also possible that procedural memory plays a distal role in language and literacy acquisition; for example, in the context of reading acquisition, a minimum level of graphotactic knowledge, acquired via procedural learning, may be sufficient to facilitate reading acquisition (Schmalz et al., 2021). Consequently, the magnitude of the procedural learning effect may not have practical implications and more likely should be conceptualised/considered as a binary indicator of whether the individual is able to demonstrate evidence of learning irrespective of the size of the effect.

Finally, we return to the issue of task impurity as it does not only pose concerns for the reliability of the SRTT, but it may also explain some of the variability in individual differences across studies. More specifically, as previously mentioned, procedural learning in the SRTT has been found to positively correlate with attention (Sengottuvel & Rao, 2013; West, Shanks, et al., 2021; Franklin et al., 2016), thus suggesting that individuals with better attentional abilities show more evidence of procedural learning. Given this link, and the role of attention in language and literacy development (de Diego-Balaguer et al., 2016; Gavril et al., 2021), it is possible that previous findings for a relationship between procedural learning and language/literacy may be due to the shared correlation between these variables and attention. In line with this hypothesis, West, Shanks, et al. (2021) observed significant correlations between procedural learning and children's attainment on measures of reading, grammar, and arithmetic. However, in a latent variable path model, when procedural learning in the SRTT, measures of declarative learning and attention were entered as predictors of children's attainment, only attention and declarative memory explained unique variance, procedural learning did not. The third variable problem may also prove useful in explaining the higher disattenuated correlations between procedural learning and language/literacy observed in our meta-analysis, as (Smolak et al., 2020) observed significant intercorrelations between sustained attention, working memory and language abilities in the DLD, but not in the TD, group. Thus, if in TD participants there is a smaller relationship between attention and language, this should be reflected in the smaller relationship between language and procedural learning. Hence, it is important that future research accounts for the role of attention, and potentially other confounding variables (e.g., working memory), when examining the role of procedural learning in language and literacy abilities.

Crucially, these findings do not exclude the possibility of a general ability for learning probabilistic regularities that may contribute to language and literacy acquisition. Instead, they highlight the current theoretical and methodological issues that hamper our ability to draw conclusions about the relationship between these measures.

Clinical implications

Language and literacy impairments have been found to persist into adulthood (Clegg et al., 2005; Lohvansuu et al., 2021) and have significant impact on social interactions, self-esteem, and educational outcomes (Alexander-Passe, 2006; Conti-Ramsden et al., 2018; Mugnaini et al., 2009), potentially contributing to the increased prevalence of psychiatric disorders in these populations (Eadie et al., 2018; Mugnaini et al., 2009). Furthermore, individuals with DLD and dyslexia have been found to be employed in lower-skilled positions (Conti-Ramsden et al., 2018; Maughan et al., 2020).

Whilst interventions for dyslexia that focus on improving word decoding and single-word identification levels have been shown to be effective, those targeting fluency and comprehension have shown more variable outcomes (S. E. Shaywitz et al., 2008). Similarly, a meta-analysis conducted by Law et al. (2003) demonstrated the positive impact of speech and language therapy for children with expressive phonological and expressive vocabulary difficulties, yet there was less evidence for its effectiveness for those with receptive difficulties. Mixed findings were observed for the effectiveness of expressive syntax interventions. Therefore, given the long-lasting impact of language and literacy difficulties, it is crucial to better understand these neurodevelopmental disorders, particularly their causal mechanisms, so that more effective interventions can be devised. However, whilst the possibility that an impairment of the procedural memory system is promising for advancing the identification and remediation of these neurodevelopmental disorders, current evidence is insufficient to determine whether evaluating the value of interventions targeting the procedural memory system would be a worthy endeavour.

Implicit learning training which builds from the knowledge obtained from procedural learning research, whereby children are given support for implicitly learning of language and literacy abilities, has been suggested to be potentially valuable for language and literacy intervention (Apfelbaum et al., 2013; Arfé et al., 2018; Plante & Gómez, 2018). This can be achieved by increasing the salience of the probabilistic, sequential and statistical regularities, for example by considering the frequency and consistency with which a target would be presented (Plante & Gómez, 2018). Arfé et al. (2018) examined the effects of an implicit intervention on the phonological-orthographic mappings against a phonological treatment in elementary school-aged children with dyslexia and observed that, whilst both groups showed improvements in spelling the trained words after the 6 sessions of intervention, only the group in the implicit intervention condition generalised their acquired spelling knowledge to untrained items. In typical populations, however, explicit instruction of spelling rules has been found to result in more progress in spelling than implicit conditions (Burton et al., 2021; Cordewener et al., 2015). At this point, given the scarcity of evidence, it is still unclear whether individuals with language and literacy impairments would benefit from implicit training. However, as suggested by Plante & Gómez (2018), the incorporation of at least some of the principles derived from statistical learning research (e.g., variability principle, i.e., high input variability of the control elements promotes learning of the target) in the educational and treatment setting will increase the salience of these regularities. As indicated by the authors, this does not require the creation of new treatment methods; instead, modifications to existing treatments should enhance learning and generalisation. According to the authors (Plante & Gómez, 2018), this training is especially beneficial for populations with language impairments which may not benefit from explicit instruction due to their linguistic difficulties (Plante

& Gómez, 2018). Critically, this suggestion does not preclude the continued use of explicit instruction but rather suggests that the adoption of activities that emphasise the probabilistic nature of language and literacy may offer a better platform for learning (Apfelbaum et al., 2013). Future research is required to determine whether implicit principles and instruction are beneficial for populations with language and literacy deficits outside of research settings.

In addition to considering the implications of this work for intervention, one should also consider whether the SRTT could ever be used as a risk marker for literacy/language difficulties. Indeed, this thesis set out to understand and improve the psychometric properties so that the SRTT could be a more reliable index of procedural memory and potentially a more sensitive risk marker for language and literacy difficulties. Unlike language and literacy assessments, the oculomotor version of the SRTT has been successfully used in infants (Koch et al., 2020) and would thus be potentially suitable for assessing the procedural memory skills of children at risk of dyslexia and DLD at earlier stages of development. However, based on the correlational evidence available (Chapters 2 and 4) (Lammertink et al., 2020; West, Melby-Lervåg, et al., 2021), it is still unclear whether this task taps into the computations involved in language and literacy abilities, and therefore would be a sensitive risk marker. Furthermore, even though this task has produced robust findings at the group-level, the poor stability of the individual scores does not yet support its use as a risk marker as the performance at one time point is likely to poorly reflect the individuals' true performance. Thus, whilst the oculomotor version of the SRTT could prove useful for the assessment of infants' early procedural learning abilities so that early intervention could ameliorate, or fully compensate, for the language and literacy deficits, future research needs to further investigate which factors interfere with the stability of this task, as well as its validity as an index of procedural memory, particularly in infants, before any efforts are made to use this task for assessment purposes.

Conclusion

This thesis sought to comprehensively examine the psychometric properties of the SRTT to ensure its validity for assessing the predictions of the procedural/declarative model and the procedural deficit hypothesis. These efforts proved fruitless, as despite numerical improvements, the stability of the procedural learning effect remained well-below psychometric standards in both populations with and without dyslexia. The discussion above highlighted some of the potential contributing factors to the SRTT's poor reliability (e.g., attention). Future research will be required to test these hypotheses and to determine whether changes can be made to improve the reliability of this task. Ultimately, these findings suggest that new procedural memory tasks designed specifically for optimising variability between individuals should be created, since this is a wide-ranging issue in the field of procedural memory, and indeed cognitive science more broadly.

Irrespective of the test's psychometric issues, which naturally have an impact on the magnitude of the correlations between procedural memory and language/literacy, we found no evidence to support the procedural deficit hypothesis, as both groups showed comparable procedural learning on the SRTT across sessions. Similarly, contrary to the predictions of the procedural/declarative model, procedural memory showed no, or small, correlations with language and literacy abilities in the TD group. However, there was more evidence for the procedural/declarative model in individuals with DLD and dyslexia. Irrespective of the limited evidence for the procedural/declarative model across populations, the relationship between procedural learning and attention was consistently replicated across studies. Whilst it is tempting to conclude that the link between procedural memory and language/literacy may be less strong than proposed by the procedural/declarative model, it is still possible for this relationship to be more prominent at earlier stages of language and literacy acquisition. Moving forward, large scale registered report studies and meta-analyses which consider the moderating and mediating effects of procedural memory through declarative and attention on language and literacy abilities across development, will help clarify the validity of these hypotheses.

Appendices

Appendix A: Trial Modelling

Using simulated response time data which follow the guidelines in Haines et al. (2020), two research choices were modelled to determine its consequences for test-retest reliability. As in Haines et al. (2020), response time data was simulated for each participant (N = 150) from a lognormal distribution, with 500 iterations. Probable and improbable conditions were simulated at each time point, with faster response times for the probable condition to represent the procedural learning effect. Since test-retest reliability was being modelled, response time data was drawn for two sessions per participant using exactly the same parameters for the lognormal distribution. Thus, the generative parameters had a reliability of 1.0.

Test-retest reliability for the simulated data was computed as the Person's correlation between the difference between improbable and probable trials for each session.

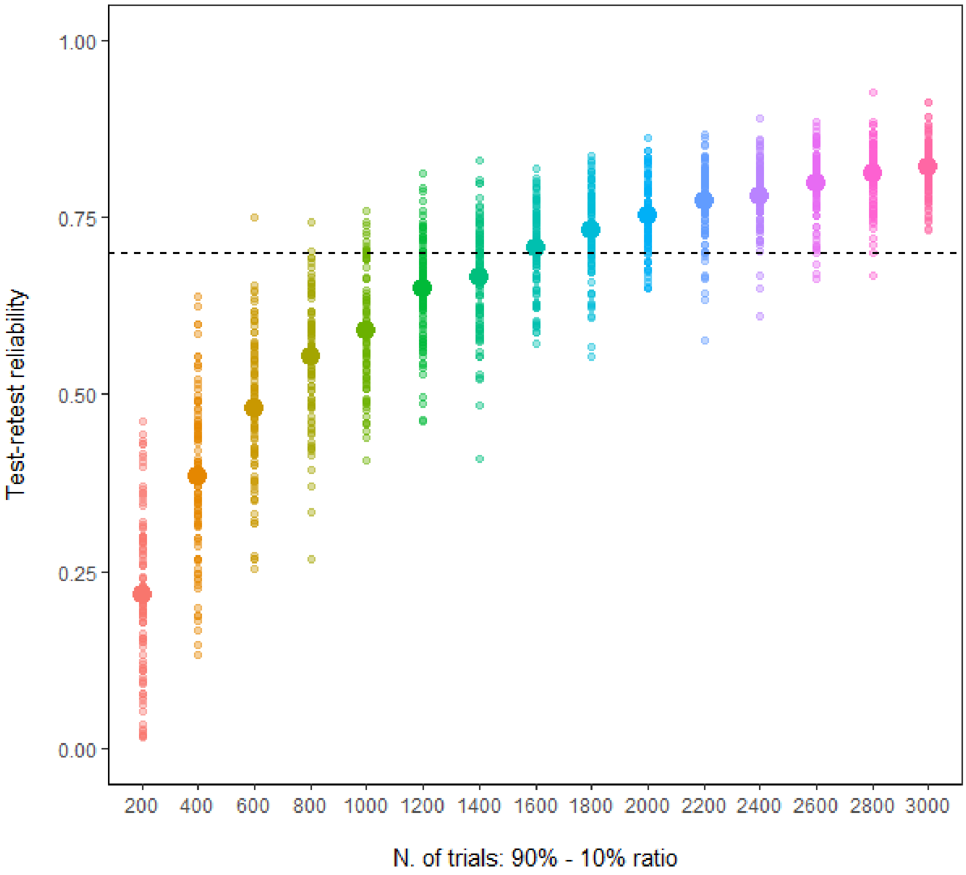
Figure 1A shows the results of modelling the effect on reliability of increasing the number of trials in the SRTT whilst maintaining the same ratio between probable and improbable trials, in this case a ratio of 90-10 (90% of probable to 10% of improbable trials) as used by West and colleagues (West et al., 2018, 2019, 2021). Whilst the increase in the number of trials led to a substantial improvement in the test-retest reliability, it only reached the threshold of acceptable test-retest reliability (*r* > .70, e.g., Burlingame et al., 1995) with 1000+ trials.

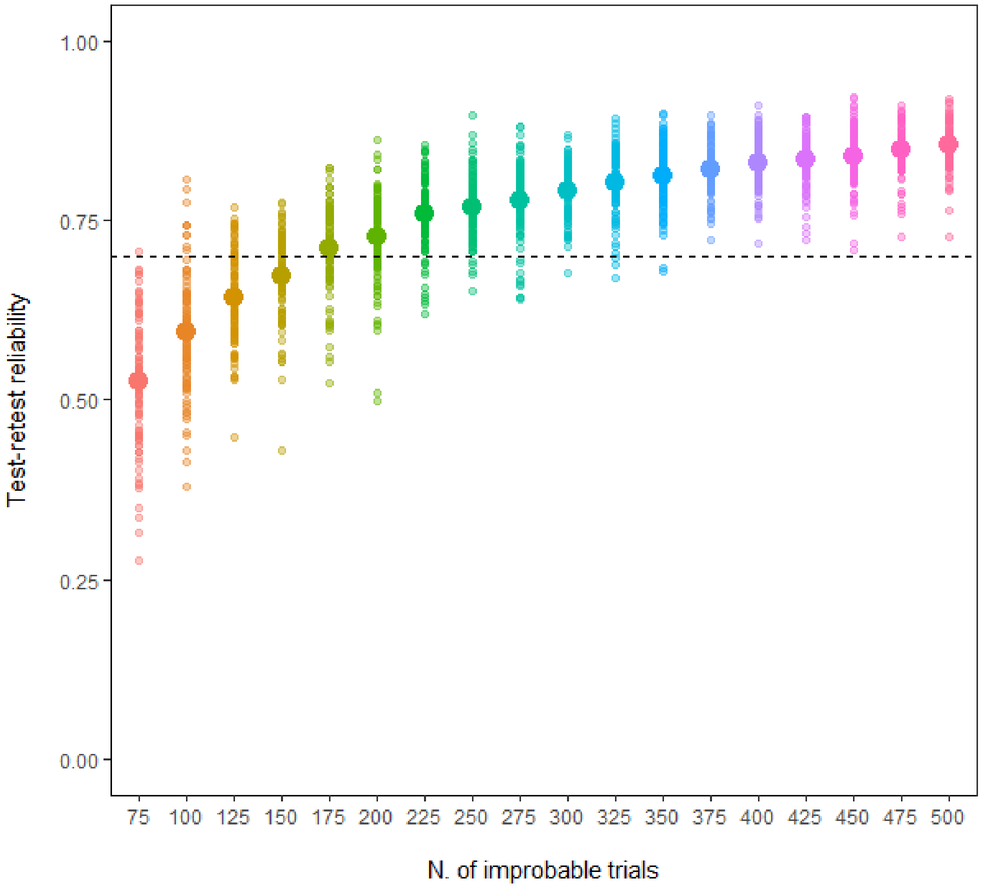
Figure 1B, on the other hand, demonstrates the effect on reliability of progressively increasing the number of improbable trials from 75 to 500 until it reaches a similar number of trials as the probable condition, which was set at 500 trials. Again, a perfect correlation between difference scores is assumed. This simulation reveals that with only 150+ improbable trials and a ratio of 70-30 a reliability above .70 is often observed. Whilst Rouder et al. (2019) has already highlighted critical role of the number of trials on the amount of attenuation of the test-retest reliability, this simulation demonstrates how attempts to increase the number of trials without consideration of the ratio between probable and improbable trials may fail to reach adequate levels of test-retest reliability.

This work suggests that alternating sequences would be expected to show higher reliability, given that the design includes the same number of sequenced and random trials. One must, however, take into consideration that a decrease in the ratio between probable and improbable trials may have consequences in the size of the procedural learning effect, which could also have negative effects on test-retest reliability due to a decrease in between-subject variability (Hedge et al., 2018). Evidence for a decrease in the procedural learning effect as the ratio decreases is noticeable when comparing the magnitude of the effect obtained in probabilistic sequences (e.g., mean procedural learning at test 18.4 ms and retest 37.5 ms, Siegelman & Frost, 2015) and alternating tasks (e.g., mean procedural learning at test 14.4 ms, mean procedural learning at retest 10.1 ms, Stark-Inbar et al., 2017).

Figure 1

Figure A represents the test-retest reliability as a function of the ratio between probable and improbable trials; and Figure B) as a function of the number of improbable trials, assuming a ratio of 90% - 10%. Dotted line indicates a test-retest reliability of .70.

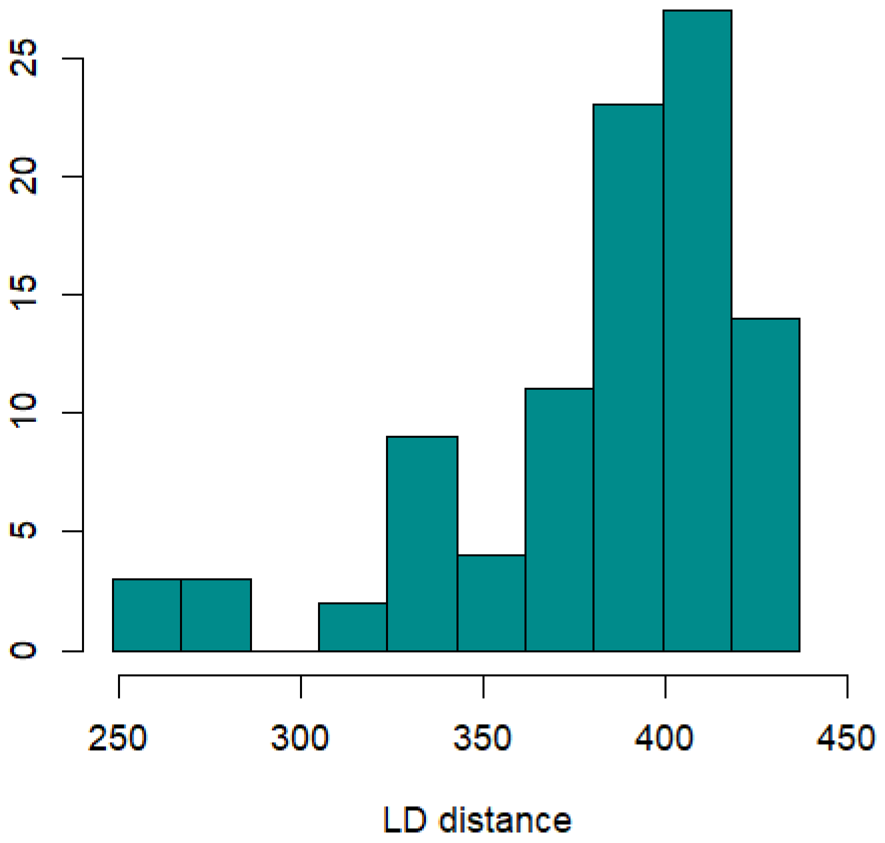



Test-retest reliability
1.00
0.75
0.50
0.25
0.00
75
100
125
150
175
200
225
250
275
300
325
350
375
400
425
450
475
500
N. of improbable trials

Appendix B: Experiment 1 - Distribution Of Similarity Levels

Figure 2

Distribution of similarity levels (computed as the Levenshtein distance, i.e., the number of edits, e.g., insertions, deletions or substitutions, required for two strings to be identical) between sequences (higher LD distance indicates lower similarity between sequences at test and retest)



Appendix C: Explicit Awareness

Generation Tasks

To measure explicit sequence memory, after completion of the SRTT, participants completed a generation task (Wilkinson & Shanks, 2004). This task presented the same four black outlined rectangles as in the main SRTT. Following Wilkinson and Shanks (2004), participants were instructed as follows: first they were asked if they had noticed a pattern in the task; and then informed that a sequence was embedded in the task with participants being asked to “guess” what the pattern was by making 100 key presses using the following instructions adapted from Wilkinson and Shanks (2004):

Inclusion instructions

During the reaction time trials, you may have noticed that the smiley face appeared in a regular repeating sequence. (Do not worry if you did not notice this; the sequence was designed to be very difficult to detect). Now in the final stage of the experiment we will see how much (if anything) you have learned about the sequence. You will have to do a slightly different task in this final block of trials. Your reaction time will no longer be measured. Instead of responding to the position of the smiley face, what we would like you to do is to press the keys 100 times, attempting to freely generate the sequence that you saw in the reaction time phase. Each time you press a key, the smiley face will appear in the appropriate box. It will remain on the screen until you press a further key. Do not worry if your memory of the sequence is poor; just try to generate the sequence as best as you can. Please avoid pressing the same key on successive occasions. The computer will tell you when you have made 100 keypresses. Try to be as accurate as possible.

Press space to start.

In the supplementary experiment, after completing the task previously described, participants were additionally asked to complete an additional task - exclusion condition, in which participants were instructed to avoid generating the sequence acquired during the SRTT. The following instructions were used following the instructions also adapted from Wilkinson and Shanks (2004):

Exclusion Instructions

In the final stage of the experiment we will use a different method to see how much (if anything) you have learned about the sequence. As in the last block you must press the keys 100 times but this time attempting to freely generate a sequence that is as DIFFERENT as possible from the one you saw in the reaction time phase. If you can remember particular parts of the sequence, then you should avoid generating them. Each time you press a key, the smiley face will appear in the appropriate box. It will remain on the screen until you press a further key. Don't worry if your memory of the sequence is poor; just try as best you can to avoid generating the sequence. Please avoid pressing the same key on successive occasions. The computer will tell you when you have made 100 keypresses.

Press space to start

Both in the inclusion and exclusion conditions, the levels of explicit awareness were analysed using the following measures: the number of total and distinct triplets and the longest consecutive string of elements of the sequence. To determine whether participants presented explicit awareness beyond chance levels, chance performance was determined by randomly generating 16000 sequences of 100 keypresses which was compared to participants' scores by conducting one sample t-tests as previously done by Lee and Tomblin (2015). Chance level performance was estimated by randomly generating 16000 sequences of 100 keypresses with the average chance level of triplets being estimated at 29.96, distinct triplets at 10.4 and of longest sequence elements at 5.84.

Results

***Experiment** 1*

40/98 participants indicated noticing a pattern (with 37 participants responding negatively, and 21 unsure).

In the generation task, participants recalled on average 9.31 distinct triplets (SD = 1.96) and 6.48 elements of the 12-item sequence (SD = 1.70). One sample t-tests revealed significant differences between the chance level and participants' performance (total triplets: $t(96) = 7.76$, $p < .001$; distinct triplets: $t(96) = -5.48$, $p < .001$; elements of the sequence: $t(96) = 3.73$, $p < .001$), with participants generating significantly less distinct triplets but more total triplets and elements of the sequence than chance.

Pearson correlations between random slopes for both sessions and performance on the generation task were negligible and non-significant (*p*s < .05) for the total number of triplets generated (session1: r = -.08, p = .452, BF_{10} = .27; session2: r = -.06, p = .566, BF_{10} = .50), different triplets (session1: r = -.04, p = .674, BF_{10} = .29; session2: r = .06, p = .594, BF_{10} = .47) and longest sequence (session 1: r = .04, p = .685, BF_{10} = .43; session2: r = -.01, p = .932, BF_{10} = .25). Similarly, a negligible relationship between similarity and explicit awareness was observed when correlating both measures of explicit awareness with similarity levels, for total triplets (r = .06, p = .57, BF_{10} = .24), different triplets (r = -.003, p = .975, BF_{10} = .30) and longest sequence recalled (r = -.06, p = .541, BF_{10} = .24). This indicates that the levels of explicit awareness are not associated with procedural learning nor with similarity.

As in the present experiment, explicit awareness has been frequently observed in the SRTT after extended training (Shanks et al., 2005; Wilkinson & Shanks, 2004). This has been found to occur despite attempts to reduce explicit awareness by adopting a probabilistic SRTT (Jimenez & Mendez, 1999) and the absence of an interstimulus interval (Destrebecqz & Cleeremans, 2001). Yet, explicit awareness was not associated with the amount of procedural learning in our experiment. It is possible that the low reliability of the SRTT could lead to the underestimation of the true relationship between sequence and explicit learning (Arnon, 2019). However, a substantive interpretation of these results is supported by the findings of Song et al. (2007), who showed that even when participants were exposed to the sequence before performing the SRTT, this did not improve procedural learning. Taken together, this pattern of results strongly suggests that implicit learning develops continuously with its time course being unaffected by explicit knowledge.

Experiment 2

After completion of the SRTT in the second session, participants were asked to answer two questions related to explicit awareness and complete a free generation task. The first question was related to whether participants noticed a pattern in the SRTT, to which 22 participants responded affirmatively, 13 negatively and the remaining 11 were unsure about the presence of a sequence. The second question referred to the timing of awareness of the presence of a pattern, 13 participants indicated that they noticed the sequence in the first session, 12 in session 2 and 7 in session 3. The remaining participants (N = 14) indicated that they never noticed a pattern. In the generation task, participants recalled on average 9.35 distinct triplets (SD = 1.39; max = 12) and 6.46 elements of the 12-item sequence (SD = 1.44; max = 12). One sample t-tests against chance levels (distinct triplets: M = 10.40; elements of the sequence: M = 5.84) revealed a similar pattern to experiment 1 with participants generating significantly less distinct triplets ($t(45)$ = -5.15, p < .001) and more elements of sequence ($t(45)$ = -2.90, p = .006) than chance levels.

Unlike experiment 1, there was evidence of a small negative, non-significant (*p*s < .05) association between procedural learning and the level of explicit awareness, both when this was indexed by the number of total triplets recalled (session 1: *r* = -.005, BF_{10} = .34, session 2: *r* = -.06, BF_{10} = .36; session 3: *r* = .06, BF_{10} = .36), different triplets (session 1: *r* = -.15, BF_{10} = 0.52, session 2: *r* = -.08, BF_{10} = .57; session 3: *r* = -.18, BF_{10} = .64), and by the longest sequence recalled by the participants (session 1: *r* = -.14, BF_{10} = .49 , session 2: *r* = -.08, BF_{10} = .37, session 3: *r* = -.007, BF_{10} = .33).

Appendix D: Test-Retest Reliability Across Distributions

As demonstrated in Haines et al. (2020), decisions about how to characterise a distribution may impact the test-retest reliability of a task. In a simulation, Haines and colleagues (2020) contrasted the test-retest reliability estimated using Pearson's correlation of the Kullback-Leibler divergence against the mean difference score. The Kullback-Leibler (K-L) divergence quantifies the distance between two probability distributions across trials. Unlike the other models we selected, the K-L divergence makes no assumptions about the shape of the RTs. The test-retest reliability for the K-L divergence was better at recovering the true test-retest reliability than the difference between means at test and retest. This was taken by the authors as suggesting that the inclusion of distribution information is important for capturing individual differences. Following from this, and in complement to the measures included in the manuscript we present the test-retest reliability for the ratio scores, random slopes from generalised mixed models with gamma distribution, and Kullback-Leibler divergence for experiments 1-2 and the supplementary experiment, in Table 1.

Table 1

Pairwise test-retest reliability of the procedural learning measures (last 600 trials)

Experiment	***Session***	***Group***	***Test-retest reliability***		
			Ratio	Random slopes	KLD
1	SRT1 - SRT2	-	-.003	.19	-.04
2	SRT1 - SRT2	-	.30	.45***	.22
	SRT2 - SRT3	-	.42	.62***	.42**
Supplementary	SRT1 - SRT2	ISI	.02	.14	-.08
	SRT1 - SRT2	noISI	.27	.52***	.25

Note. *p < .05, **p < .01, ***p < .001; outliers were removed before computing the test-retest reliability for ratio and random slopes but not for KLD

Appendix E: Supplementary Experiment

Abstract

The Serial Reaction Time task (SRTT) is a robust measure of procedural learning despite consistent poor stability across sessions (*r*s < .70), with the exception of (West et al., 2021). In this experiment, we aimed to more closely replicate their design by adopting an interstimulus interval (ISI) of 250 ms and recruiting participants from a wider age range (18-60 years). Counter to our predictions, the test-retest reliability was poor for both groups with and without an ISI (*r*s < .50), suggesting that other design features may have contributed to the superior stability observed by West, Shanks, et al. (2021).

Introduction

Whilst manipulating the similarity between sequences in Experiment 1 and increasing the number of sessions in Experiment 2 contributed to a larger experimental effect for later sessions and decrease in trial variability, the test-retest reliability of the SRTT remained suboptimal. This may be associated with the small variability between participants when compared to within-subject variance in the procedural learning effect, which diminishes the test's capacity to effectively differentiate participants' performance (Hedge et al., 2018; Spearman, 1910). In West, Shanks, et al. (2021), where substantially higher test-rest reliability was observed, both the magnitude of the procedural learning effect and the variability across participants were considerably higher than our previous experiments. Thus, in this experiment we aimed to more closely replicate West and colleagues' (2021), study in adults and examine whether participants' wider age range and the presence of an ISI contributed to the higher test-retest reliability.

In this experiment we compared SRTT performance and reliability in two participant groups using the same sequences as in experiment 2: for one group we included an ISI of 250 ms, as in West, Shanks, et al. (2021), and for the other there was no ISI, as in Experiment 2. To explore the effect of age on the stability of procedural learning, we recruited participants aged 18-60 years as in West, Shanks, et al. (2021), with the groups age-matched. Importantly, this experiment was also run online, allowing us to assess the reliability of an online SRTT. Finally, comparing the relationship between procedural learning and attention in the two ISI groups also offered a further window into the role of attention in the SRTT.

Previous evidence by Destrebecqz and Cleeremans (2001, 2003) has suggested that individuals trained with a 250 ISI develop stronger representations of the underlying sequences, which may explain the larger procedural learning effects observed by West and colleagues' (West et al., 2021). If superior procedural learning (and potential automatisation) leads to earlier independence from attention, one would expect that the ISI group would show a smaller correlation between procedural learning and attention (Seger & Spiering, 2011). Alternatively, given the increased duration of the ISI version, it is possible that individual differences in attention may be related to procedural learning in this longer (potentially less engaging) version (Franklin et al., 2016). Thus, these (speculative) explanations were also explored.

The hypotheses (pre-registered at https://osf.io/t78kf) were as follows:

- H1: Participants are expected to demonstrate evidence of procedural learning in both sessions.
- H2: Both groups (with ISI and no-ISI) will show higher split-half reliability than test-retest reliability.
- H3: Higher test-retest reliability will be expected for the ISI group which replicates West, Shanks, et al.'s (2021) design more closely than the no-ISI group. It was also predicted that the ISI group will show faster RTs and larger procedural learning effects, and subsequently less evidence of practice effects than the group with no-ISI.
- H4: Slower RTs are expected for older participants, despite similar levels of procedural learning for younger and older participants. Older participants will also be expected to show more stable procedural learning than younger participants.[4]
- H5: Both ISI groups will demonstrate a positive correlation between procedural learning and sustained attention, as measured by the PVT; however, if the attentional demands of the SRTT diminish once the sequence becomes more predictable, it would be expected that the ISI group would show a smaller correlation between procedural learning and attention than the group with no-ISI in Session 2.
- H6: Finally, it was predicted that participants in the ISI group will demonstrate higher evidence of explicit awareness by recalling more elements of the sequence, with this difference being more noticeable in the exclusion condition.

[4] The effect of age on reliability was explored in Appendix E.1

Methods

Participants

One hundred and thirty-five participants aged 18-60 years (M = 29.74, SD = 10.01) took part. The sample included monolingual, bilingual and multilingual individuals from 20 distinct nationalities. All participants were proficient English or Portuguese speakers with normal or corrected-to-normal vision. Recruitment was conducted on social media (e.g., Twitter, Reddit and Facebook) and each participant was randomly assigned to the ISI or no-ISI group. The experiment was approved by the Ethics Committee of the Psychology Department at the University of York and each participant gave written informed consent. A power analysis based on the lower bound of the test-retest reliability obtained by West, Shanks, et al. (2021) indicated that 35 participants per group would be required to achieve 80% power.

Measures[5]

Serial Reaction Time task

The SRTT used in Experiment 2 was used here. For the group with ISI, there was an interval between trials of 250 ms, whilst for the no-ISI group the following trial started as soon as the response was made. RT and accuracy were measured for each trial. Participants had to respond to each trial to initiate the next trial, regardless of accuracy.

Psychomotor Vigilance Task

Sustained attention was measured using a shorter (5 minutes) and online version of the Psychomotor Vigilance task (Reifman et al., 2018) used in experiment 2.

Free Generation Tasks

Explicit awareness of the training sequence was assessed through two free generation tasks (Wilkinson & Shanks, 2004). After completing the SRTT, participants were informed of the presence of

[5] Online versions of the SRTT, PVT and generation tasks are available at https://gitlab.pavlovia.org/memory-group

a pattern in the stimuli and were then presented with the outline of four rectangles as in the SRTT and asked to press the same keys to elicit the appearance of the stimuli. The target remained in the corresponding position to the key pressed until a new response was made. In the inclusion task, each participant was asked to generate 100 guesses of the underlying sequence learned in the SRTT; whilst in the exclusion task participants were asked to produce a sequence as different as possible from the training sequence. Each sequence generated by the participants was coded into the number of different triplets recalled. Further details are presented in the Appendix C.

Procedure

The experiment used a mixed-subjects design with each participant randomly assigned to an ISI group and performing the SRTT at two time points each separated by roughly one week (M = 7.85, SD = 2.21, range = [5 - 20]). The two underlying sequences of the SRTT were counterbalanced to avoid order effects. Participants were randomly allocated to one of two groups: (i) no interstimulus interval (no ISI group) or (ii) 250 ms (ISI group). Groups were age-matched to avoid age-related differences in procedural learning. After completion of the SRTT, sustained attention was measured using an adapted version of the Psychomotor Vigilance task (Reifman et al., 2018) to explore the relationship between attention and procedural learning. Additionally, after completion of the SRTT in Session 2 participants performed two free generation tasks aiming to capture explicit knowledge of the sequence. All tasks were programmed using Psychopy 3 (Peirce et al., 2019) and the sessions were run online on the Pavlovia Platform (Bridges et al., 2020).

Statistical analyses

H1 and H4: Mixed effects model

The same procedures as in experiments 1 and 2 were adopted for data treatment and analyses. Two participants were removed from the analyses for both sessions for the ISI group and four participants for the no ISI group, with two of them removed for both sessions and the remaining one from session 1 and the other from session 2.

Given the transition into online testing, a model contrasting online and in-lab results was computed to compare the performance in the SRTT across settings. This model showed similar performance in both settings (in-lab vs online), thus showing that the paradigm is robust to changes in testing conditions (see Appendix E.2). Having established its robustness, the results for this experiment were then analysed by fitting a mixed effects model to the performance of both groups

(ISI vs no-ISI) on the SRTT. This model included *Probability* (probable vs improbable)*, Epoch* (contrasts between successive epochs 2-1, 3-2, 4-3, 5-4)*, Session* (1 vs 2) and *Group* (ISI vs no-ISI) as fixed effects and *Participants* as random effects, with a maximal-fixed-effects structure as all interactions between the predictors were analysed.

A second model was computed to explore age-related changes in procedural learning. This model included the main effects *Probability*, *Session*, *Group* and *Age* and was only estimated on the last 600 trials when procedural learning would be expected to be more salient. Since age is a continuous predictor, this variable was standardised and centred. Two participants were identified as influential for the first model and one for the second.

H2 and H3: Reliability

For split-half and test-retest reliability analysis of the SRTT were computed similarly to previous experiments, yet included Group as a predictor to account for group differences (Lammertink et al., 2020). Model random effects were extracted using the *ranef()* function from the *lme4* package (Bates et al., 2015).

For the PVT, some of the most used measures were computed: lapses (number of response times equal or above 500 ms), mean RT, mean reciprocal response time (mean 1/RT) and median RT (Basner & Dinges, 2011). Additionally, an ex-Gaussian analysis was performed on the PVT for each participant to determine the levels of variability of the response times distribution, i.e., the tau parameter, whose reliability was also analysed.

H6: Explicit awareness

Explicit awareness performance was analysed by comparing the performance of each group in the generation task in both conditions (inclusion and exclusion). A two-way ANOVA with *Group* (ISI vs no-ISI) and *Condition* (inclusion vs exclusion) as predictors and number of triplets recalled as an outcome variable was conducted to determine whether there were significant differences between groups in explicit awareness of the sequence at time 2 in inclusion and exclusion conditions.

Additionally, to determine whether participants presented explicit awareness beyond chance levels, chance performance was determined by randomly generating 16000 sequences of 100 keypresses which was compared to participants' scores by conducting one sample t-tests as previously done by Lee and Tomblin (2015).

H5: Relationship between procedural learning, attention and explicit memory

Pearson correlations between procedural learning, attention and explicit were computed for each session per group. Correlations were compared using the Fisher r-to-z transformation.

Results

RT data was available for 134/135 participants for Session 1 (ISI: N = 68; noISI: N = 66) and for 101/135 participants for Session 2 (ISI: N = 52; noISI: N = 49). One participant was removed from the analysis as they had already taken part in a previous experiment, with three additional participants (ISI: N = 2; noISI: N = 1) being removed from analyses for Session 2 due to an administrative error. The remaining participants failed to return for Session 2.

H1: Procedural learning in the SRTT - Effect of ISI

A linear mixed effects regression model was fitted to the response time data. RTs decreased with practice as evidenced by the significant main effects of *Session* and *Epoch* for Epoch2-1 and Epoch4-3 (Epoch4-3 was no longer significant after correction for multiple comparisons), yet there was a group effect as the ISI group was significantly faster than the noISI group. Additionally, there was evidence of procedural learning - faster response times for probable than improbable trials, with this difference also increasing with practice as evidenced by a significant interaction between *Probability* and *Epoch* for Epoch2-1 and Epoch4-3 (Epoch 2-1 did not survive correction for multiple comparisons), yet there was no evidence of improvements between sessions as evidenced by the non-significant three-way interaction between *Probability*, *Epoch* and *Session*.

Despite significantly faster RTs for the ISI group, there were no significant overall group differences for procedural learning as evidenced by the two-way interaction between *Probability* and *Group*. Yet, the three-way interaction between *Probability*, *Epoch* and *Group* showed significant differences during the learning process, with the ISI group showing stronger evidence of procedural learning for the contrasts between Epoch4-3, whilst the no-ISI group showed better performance for the contrast between Epoch3-2. Only the 3-way interaction for Epoch3-2 survived correction for multiple comparisons. There was no significant difference for the first and last Epochs (Epoch 2 vs Epoch 1 and Epoch 5 vs Epoch 4). When considering the four-way interaction between *Probability*, *Epoch*, *Session* and *Group*, it showed a similar pattern to the three-way interaction, even though the only significant difference was found on the last Epoch contrast with the no-ISI group showing more evidence of procedural learning.

Figure 3

Mean and 95% CI response times for probable and improbable trials per Epoch and Session for ISI and no-ISI groups (Session 1 on the left, Session 2 on the right).

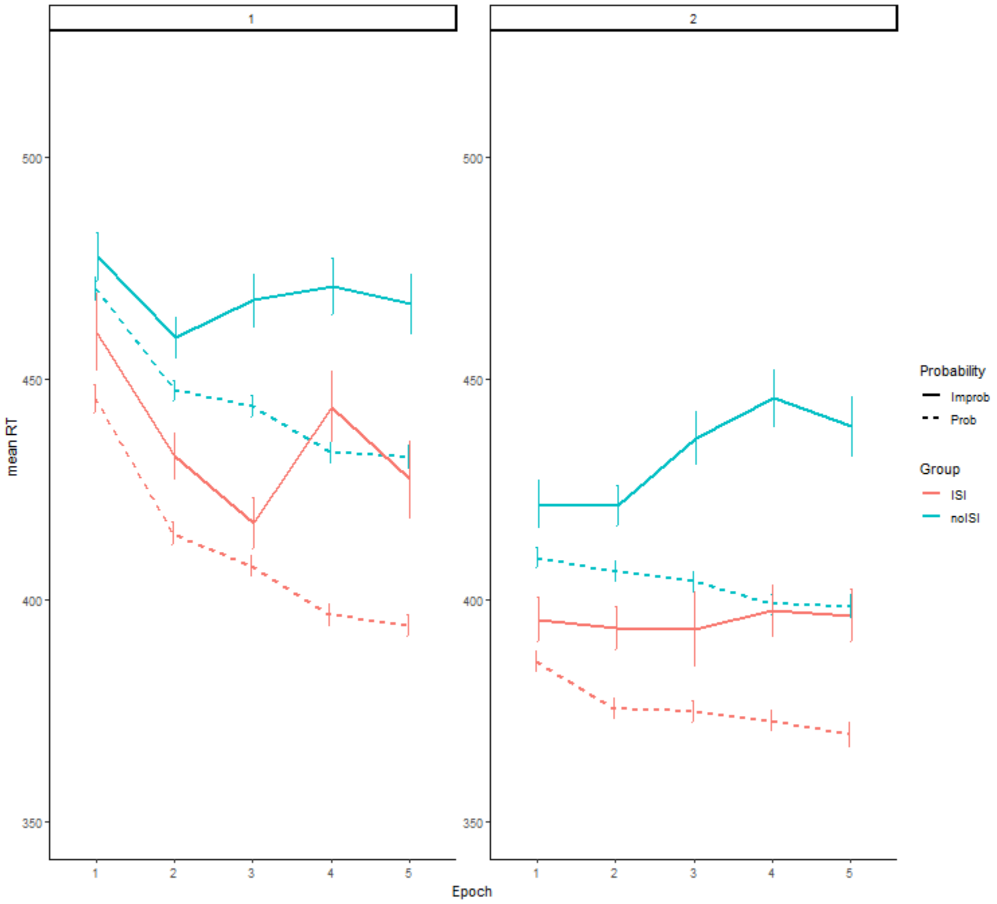


Table 2

Predictors of the group effect on the magnitude of procedural learning

Fixed effects	*b*	*SE*	*t*	*p*	*CI*	
(Intercept)	**6.018**	**0.020**	**297.137**	**<.001**	**5.977**	**6.059**
Probability	**0.026**	**0.002**	**11.783**	**<.001**	**0.022**	**0.030**
Epoch2-1	**-0.026**	**0.006**	**-4.330**	**<.001**	**-0.038**	**-0.014**
Epoch3-2	-0.006	0.005	-1.060	.294	-0.016	0.005
Epoch4-3	0.009	0.004	2.133	.036	0.001	0.018
Epoch5-4	-0.011	0.006	-1.853	.069	-0.023	0.001
Session	**0.042**	**0.005**	**8.361**	**<.001**	**0.032**	**0.052**

Group	**-0.076**	**0.020**	**-3.750**	**.001**	**-0.117**	**-0.035**
Probability x Epoch2-1	0.008	0.003	2.732	.006	0.002	0.013
Probability x Epoch3-2	0.005	0.003	1.748	.080	-0.001	0.010
Probability x Epoch4-3	**0.023**	**0.003**	**7.435**	**<.001**	**0.017**	**0.029**
Probability x Epoch5-4	-0.006	0.003	-1.732	.083	-0.012	0.001
Probability x Session	-0.001	0.002	-0.862	.393	-0.005	0.002
Epoch2-1 x Session1	-0.017	0.006	-2.787	.007	-0.029	-0.005
Epoch3-2 x Session1	-0.007	0.005	-1.243	.219	-0.017	0.004
Epoch4-3 x Session1	-0.003	0.006	-0.512	.610	-0.015	0.009
Epoch5-4 x Session1	0.011	0.007	1.535	.130	-0.003	0.024
Probability x Group	0.001	0.002	0.210	.835	-0.004	0.005
Epoch2-1 x Group	0.003	0.006	0.540	.592	-0.009	0.015
Epoch3-2 x Group	-0.009	0.005	-1.665	.101	-0.019	0.002
Epoch4-3 x Group	0.002	0.004	0.337	.737	-0.007	0.010
Epoch5-4 x Group	-0.006	0.006	-1.023	.310	-0.018	0.006
Session x Group	0.003	0.005	0.533	.597	-0.007	0.013
Probability x Epoch2-1 x Session	0.003	0.003	1.066	.286	-0.002	0.008
Probability x Epoch3-2 x Session	-0.001	0.003	-0.235	.814	-0.006	0.005
Probability x Epoch4-3 x Session	0.002	0.003	0.700	.484	-0.004	0.008
Probability x Epoch5-4 x Session	-0.002	0.003	-0.659	.510	-0.008	0.004
Probability x Epoch2-1 x Group	0.001	0.003	0.248	.804	-0.005	0.006
Probability x Epoch3-2 x Group	**-0.010**	**0.003**	**-3.604**	**<.001**	**-0.016**	**-0.005**
Probability x Epoch4-3 x Group	0.008	0.003	2.512	.012	0.002	0.014
Probability x Epoch5-4 x Group	0.001	0.003	0.218	.828	-0.006	0.007
Probability x Session x Group	-0.001	0.002	-0.436	.665	-0.004	0.003
Epoch2-1 x Session x Group	0.002	0.006	0.284	.777	-0.011	0.014
Epoch3-2 x Session x Group	-0.001	0.005	-0.205	.838	-0.012	0.010
Epoch4-3 x Session x Group	0.000	0.006	0.063	.950	-0.012	0.012
Epoch5-4 x Session x Group	-0.010	0.007	-1.418	.162	-0.024	0.004

Probability x Epoch2-1 x Session x Group	0.000	0.003	-0.003	.998	-0.005	0.005
Probability x Epoch3-2 x Session x Group	0.000	0.003	0.043	.966	-0.005	0.006
Probability x Epoch4-3 x Session x Group	0.005	0.003	1.586	.113	-0.001	0.011
Probability x Epoch5-4 x Session x Group	-0.008	0.003	-2.534	.011	-0.014	-0.002

Random effects	*Variance*	*SD*
Participant (Intercept)	0.017	0.131
Participant: Epoch2-1	0.001	0.034
Participant: Epoch3-2	0.001	0.028
Participant: Epoch4-3	0.000	0.020
Participant: Epoch5-4	0.001	0.032
Participant: Session	0.001	0.032
Participant: Probability	0.000	0.013
Participant: Epoch2-1 x Session	0.001	0.035
Participant: Epoch3-2 x Session	0.001	0.029
Participant: Epoch4-3 x Session	0.001	0.033
Participant: Epoch5-4 x Session	0.002	0.040
Participant: Session x Probability	0.000	0.008

H4: Procedural learning in the SRTT - Age effects

To explore age differences in the performance on the SRTT, a linear mixed effects model was fitted to the RTs for the last 3 Epochs. In agreement with the previous model, RTs were significantly predicted by *Probability*. In addition to the probability effect which showed evidence for procedural learning, the two-way interaction between *Probability* x *Session* indicates that the difference between probable and improbable trials increased with experience. In line with the previous model, the procedural learning effect was similar for both groups with a non-significant interaction between *Probability* and *Group* x *Probability, Group* x *Session*. *Age* was also a significant predictor of RTs indicating that younger participants showed faster RTs than older participants (no longer significant after correction for multiple comparisons), yet there were no significant differences in procedural learning across ages as evidenced by the non-significant two-way interaction between *Probability* x *Age*, the three-way interaction between *Probability*, *Age* and *Session* and between *Probability* x *Group* x *Age* and the four-way interaction between *Probability* x *Age* x *Session* x *Group*.

Table 3

Predictors of the age effect on the magnitude of procedural learning

Fixed effects	*b*	*SE*	*t*	*P*	*CI*	
(Intercept)	**6.010**	**0.020**	**301.991**	**<.001**	**5.970**	**6.051**
Probability	**0.034**	**0.003**	**12.926**	**<.001**	**0.029**	**0.040**
Session	**0.036**	**0.005**	**7.687**	**<.001**	**0.027**	**0.046**
Group	**-0.068**	**0.020**	**-3.398**	**.001**	**-0.108**	**-0.028**
Age	0.049	0.020	2.427	.019	0.008	0.090
Probability x Session	-0.002	0.002	-0.826	.413	-0.007	0.003
Probability x Group	-0.001	0.003	-0.236	.814	-0.006	0.005
Session x Group	-0.001	0.005	-0.181	.857	-0.010	0.009
Probability x Age	0.002	0.003	0.721	.475	-0.004	0.007
Session x Age	0.001	0.005	0.116	.908	-0.009	0.010
Group x Age	-0.020	0.020	-1.004	.321	-0.061	0.020
Probability x Session x Group	0.000	0.002	-0.026	.979	-0.005	0.005
Probability x Session x Age	0.001	0.002	0.558	.580	-0.003	0.006
Probability x Group x Age	0.002	0.003	0.765	.449	-0.003	0.008
Session x Group x Age	-0.013	0.005	-2.768	.008	-0.023	-0.004
Probability x Session x Group x Age	0.001	0.002	0.576	.567	-0.003	0.006

Random effects	*Variance*	*SD*
Participant (Intercept)	0.017	0.129
Participant: Probability	0.001	0.014
Participant: Session	0.001	0.029

Participant: Probability x Session	0.000	0.012

H2 and H3: Reliability

Moderate to high split-half reliability was observed for both groups, particularly when using random slopes. With the exception of the difference scores for the last 600 trials, the no-ISI group showed better split-half reliability than the ISI group. However, only the group contrasts for the random slopes in session 1 for the whole task (session 1: z = -4.19, p < .001) and random slopes (Session 1: z = -3.99, p < .001) remained statistically significant after correction for multiple comparisons.

Overall test-retest reliability was below psychometric standards (i.e., <.70; .01-.48). Test-retest reliability for difference scores was similar for both ISI groups. For the random slopes the no-ISI group showed numerically higher scores when only the last 600 trials were considered (z = -2.00, p = .046) but this did not survive correction for multiple comparisons.

Table 4

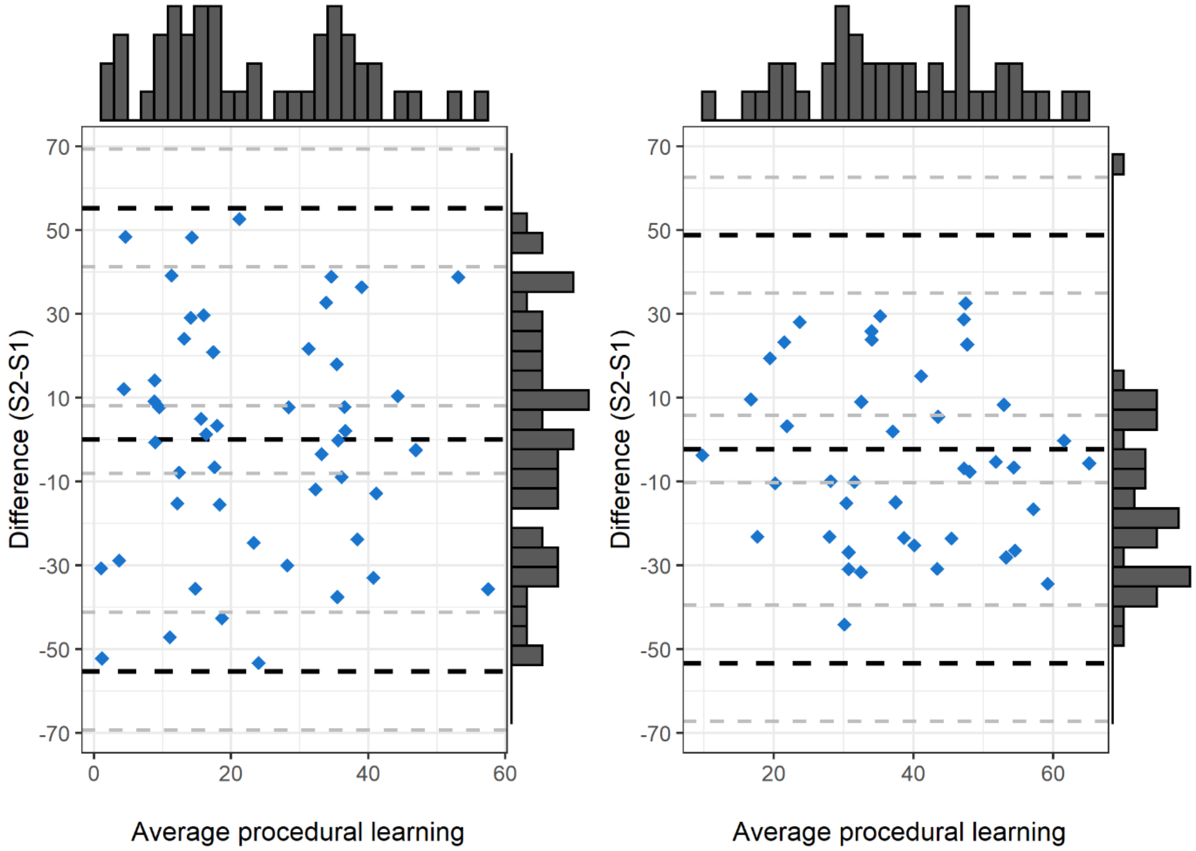
Split-half and test-retest reliability of the procedural learning measures for overall and last 600 trials of the SRTT for sessions 1 (SRTT1) and 2 (SRTT2)

Measure		***Split-half reliability***				***Test-retest reliability***	
		ISI		**no ISI**		**ISI**	**no ISI**
		SRTT1 (N 62 – 65)	SRTT2 (N 48 – 49)	SRTT1 (N 59 – 61)	SRTT2 (N 45 – 45)	S 1-2 (N 49 – 50)	S 1-2 (N 43 – 45)
Difference	Overall	.56	.69	.67	.77	.19	.08
	Last 600 trials	.53	.58	.52	.60	.01	.03
Random Slopes	Overall	.71	.69	.93	.83	.20	.44
	Last 600 trials	.73	.59	.93	.74	.09	.48

Bland-Altman plots were used to compare the agreement between procedural learning for the last 3 epochs for each group. These revealed smaller limits of agreement (lower and upper black lines in Figure 4) for the no-ISI group (-53.43; 48.76) than the ISI-group (-55.33; 55.24), indicating a more consistent performance for individuals in the former group. Irrespective of the previous differences, both groups showed wide limits of agreement, suggesting that there is an unacceptable degree of agreement.

Figure 4

Plot of the mean of the two measurements against the differences between procedural learning in session 1 and session 2 for the ISI (left) and no-ISI (right) groups



Split-half reliability for the Psychomotor Vigilance task varied between .32 to .90 with similar scores for both sessions, with the exception of mean RT. Test-retest reliability ranged from .51 to .76. Median RT shows the highest split-half (*r* ≥.90) and test-retest reliability (*r* = .76) at both time points. Mean RT showed the lowest split-half and test-retest reliability of all measures.

Table 5

Descriptive statistics, split-half and test-retest reliability of the psychomotor vigilance task

PVT measure	***Descriptive statistics***						***Split-half reliability***		***Test-retest reliability***
	Session 1			Session 2			Session 1	Session 2	
	N	M	SD	N	M	SD			
Lapses	97	1.79	2.34	62	2.16	3.15	.62	.70	.65
Mean RT	99	338.59	88.81	63	353.58	80.83	.77	.32	.51
Median	99	306.35	53.68	63	323.83	50.00	.93	.90	.76
Reciprocal	97	3.43	1.33	64	3.16	.52	.65	.84	.60

Ex-gaussian Analysis

Similarly to the findings of experiment 2, there was a pattern for a numerically higher reliability for the procedural learning effect for those who showed lower variability on the PVT (lower tau), especially on the second session. Yet, this effect only emerged for the noISI group (comparable to the conditions in experiment 2).

Table 6

Test-retest reliability for low and high tau groups for each session

Session	***ISI***		***noISI***	
	low-tau	high-tau	low-tau	high-tau
1	*r*(17) = .19	*r*(20) = .11	*r*(17) = .51*	r(15) = .46
2	*r*(13) = -.06	r(11) = -.06	*r*(13) = .61	r(13) = .30

Note. * p<.05

H6: Explicit awareness

Each generated sequence was coded on the number of triplets recalled in common with the training sequence with a maximum score of 98 triplets. Chance level performance was estimated to be 29.98 triplets.

Table 7

Descriptive statistics (means and standard deviations) for the explicit awareness tasks

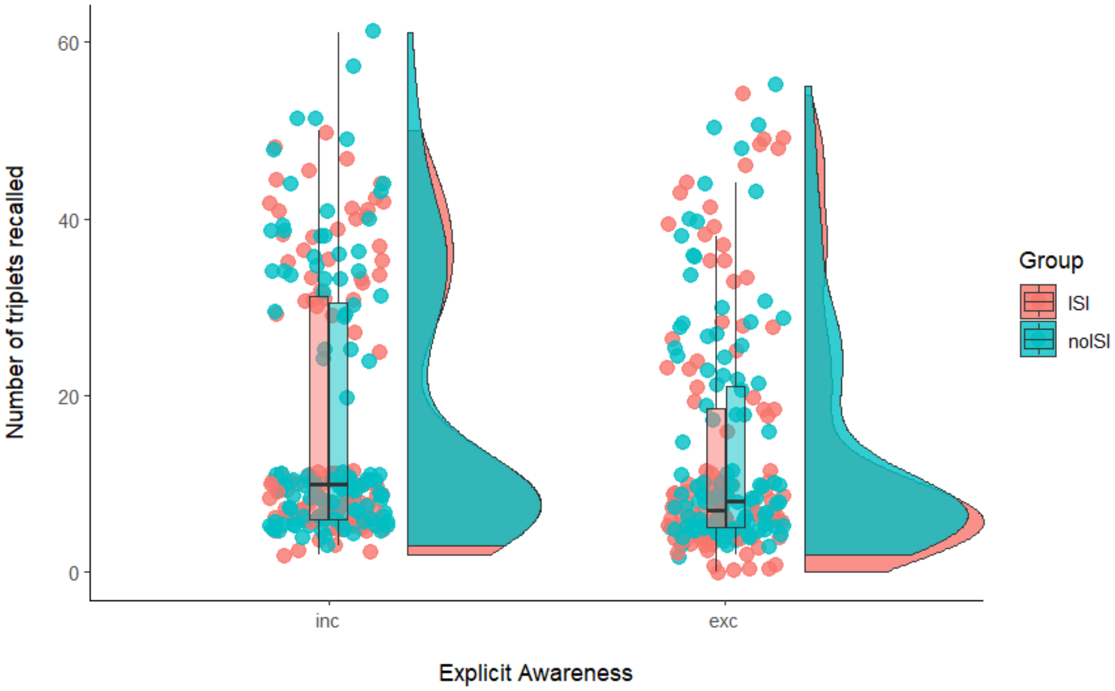
Triplets	***ISI***			***no-ISI***		
	N	Mean	SD	N	Mean	SD
Inclusion	36	36.92	6.23	38	36.89	9.27
Exclusion	37	28.62	14.80	40	28.55	11.90

Both groups showed a similar performance on the explicit awareness tasks (Table X) as evidenced in the two-way ANOVA, with *Group* as a non-significant predictor of performance on these tasks (*F*(1, 144) = .001, *p* = .98). There was a statistically significant effect of *Condition* (inclusion vs exclusion; *F*(1, 144) = 20.95, *p* = <.001), as participants recalled more triplets in the inclusion than in the exclusion condition, but no significant interaction between *Group* x *Condition* (*F*(1, 144) = .00, *p* = .99).

Comparisons against chance level for the inclusion task revealed a statistically significant higher number of triplets recalled by participants when compared to chance level for both ISI (*t*(35) = 6.68, *p* < .001) and no-ISI groups (*t*(36) = 4.54, *p* < .001). For the exclusion task, participants in both groups performed similarly to chance levels (ISI: *t*(36) = -.56, *p* = .581; no-ISI: *t*(37) = -.74, *p* = .466), thus revealing some control over their explicit knowledge of the sequence. Thus, contrary to predictions, the ISI condition did not result in greater levels of explicit awareness when compared to the noISI condition.

Figure 5

Violin and boxplots showing the distribution of number of items recalled in the generation tasks in the inclusion and exclusion conditions by group



H5 and H6: Relationship between procedural learning and attention, and explicit awareness

Pearson correlations revealed a significant negative association between attention on both sessions and procedural learning for the ISI group for session 1 and 2 (session 1: *r* = -.42, session 2: r = -.38), yet only the association for Session 1 remained significant after correction. This association indicates that participants with better attention show more evidence of procedural learning when the sequence is presented more slowly. None of the other correlations between the performance on the PVT and procedural learning were significant for either the ISI or no-ISI groups. Similarly, there were no significant correlations between explicit awareness and procedural learning (*p* > .05). None of the group contrasts were statistically significant *(p* > .05).

Table 8

Pearson correlations between procedural learning in the SRTT and attention and explicit awareness measures

Correlations	***ISI***				***noISI***			
	SRT1		SRT2		SRT1		SRT2	
	N	r	N	r	N	r	N	r
Median 1	50	**-.42*****[a]	41	.12	45	-.02	37	-.23
Median 2	30	-.38*[a]	27	.09	34	.23	31	-.32
Inclusion generation	36	-.16	35	-.03	38	.18	36	.09
Explicit generation	36	.10	35	.03	39	.14	37	-.19

Note: ***p<.001; * p<.05; bold – survived corrections for multiple comparisons using the Holm-Bonferroni Method (Holm, 1979), [a] Correlations with Bayes factor equal or bigger than 3

Discussion

As in Experiments 1 and 2, there was clear evidence of procedural learning for both groups and Sessions. The ISI group showed faster response times than the noISI group as expected (Martini et al., 2013) but this did not translate into differences in procedural learning. However, there was a small, and non-significant, increase in the procedural learning effect from session 1 to session 2 for the noISI group whilst the opposite was observed for the ISI group. Split-half reliability was again higher than stability across sessions for both groups, with significantly better split-half reliability for the no-ISI group. Test-retest reliability was also again below psychometric standards and did not differ statistically for the noISI and ISI groups (r = .42 and r = .21, respectively). Thus, counter to predictions, including an ISI did not lead to improved stability; in fact, there was more evidence for the reverse pattern. Therefore, the presence of an ISI cannot account for the higher levels of stability reported by West, Shanks, et al. (2021). In the current experiment, the ISI group showed lower levels of procedural learning overall, as well as higher within-subject variability, than those reported by West, Shanks, et al. (2021) (Session 1: M = 96.06; SD = 69.73; Session 2: M = 103.33; SD = 55.02). This may be because the ISI task characteristics may have led to increased attentional demands for the ISI group given that the duration of the ISI version of the SRTT lasted approximately 4 minutes longer (representing a 33% increase in task duration) than the noISI version. Corroborating this hypothesis there was evidence of a stronger correlation between attention and procedural learning, and higher response time variability for this group (ISI: 399.83 (132.59); noISI SD = 431.65 (118.48)).

Thus, fluctuations in attention and decreases in motivation would be more likely to occur for the ISI version, which is in line with the poor enjoyment of the task observed by West, Shanks, et al. (2021).

The finding of a significant association between attention and procedural learning for the ISI group only conflicts with Experiment 2, where an association was observed with a 0ms ISI. However, this could be due to the use of the 5-minute PVT (as opposed to the 10-minute task used in Experiment 2). Indeed, a follow-up experiment (Appendix E.3) replicated the correlation between procedural learning and attention for a noISI group (N = 52) when using the 10-minute version online.

Counter to predictions there was no effect of age on procedural learning. Similarly, ISI groups showed comparable performance on the explicit awareness tasks (in contrast with Huang et al., (2017) and Verneau et al. (2014)) and there were no associations between explicit awareness and procedural learning.

In sum, both ISI groups showed comparable procedural learning effect, and there was no evidence that the ISI group showed greater reliability. In fact, the noISI group showed greater within session reliability than the ISI group. Similarly, there was no effect of age on procedural learning or reliability. Thus, neither ISI nor age account for the higher retest reliability previously reported by West, Shanks, et al. (2021). Finally, Experiment 2 examines the use of the SRTT in an online setting, importantly demonstrating that the magnitude and trajectory of procedural learning, as well as its stability, is consistent with data collected in the lab. This suggests that, despite its poor psychometric properties, this task can reveal robust procedural learning effects in these different settings.

Appendix E. 1: Reliability Comparison - Younger Vs Older Participants

Younger and older groups were determined by computing the median split of age per group. Given age's highly positively skewed distribution, a logarithmic transformation was performed. Test-retest reliability for each group is presented in Table 6. Despite the numerically higher reliability for the noISI group, against our predictions, there was no evidence of a clear pattern for higher reliability for the older participants. However due to the small sample size, these results should be interpreted with caution.

Table 8

Pairwise test-retest reliability of the procedural learning measures for younger and older participants

Measure		***Test-retest reliability***			
		ISI		noISI	
		Younger (N = 22)	Older (N = 23)	Younger (N = 21)	Older (N = 18)
Difference scores	Overall	.23	.06	-.18	.34
	Last 600 trials	-.01	-.09	-.08	.18
Random Slopes	Overall	.23	.15	.39	.53
	Last 600 trials	.15	.02	.52	.38

Appendix E. 2: Online Versus In-Lab Testing

The model comparing online and in-lab performance included the participants from experiment 2 and the individuals in the no-ISI group from the supplementary experiment as these participants experienced the same conditions except for the mode of testing. This model included the predictors *Probability*, *Epoch*, *Session* and *Testing Condition* (online vs in lab), with all interactions being considered.

A linear mixed effects model was fitted to the data to determine whether there were differences in the magnitude and trajectory of the procedural learning effect between participants tested online (experiment 2) and in the lab (supplementary experiment - noISI group). As expected, *Epoch* and *Session* were significant predictors, indicating that RTs decreased with practice. *Probability* was a significant predictor of RTs, thus showing evidence of procedural learning with RTs being faster for probable than improbable trials and that this effect increased with practice as evidenced by the interactions between *Probability* x *Epoch* and *Probability* x *Session*. Participants in both conditions showed comparable procedural learning as there was no evidence of an interaction between *Probability* x *Testing Condition, Probability x Block* x *Testing Condition* or *Probability x Block x Session* x *Testing Condition*. Overall, there was no evidence that online testing resulted in differences in performance on the SRTT compared to in-lab testing.

Table 9

Predictors of the testing condition on the magnitude of procedural learning

Fixed effects	*b*	*SE*	*t*	*p*	*CI*	
(Intercept)	**6.062**	**0.013**	**466.204**	**<.001**	**6.037**	**6.088**
Probability	**0.033**	**0.001**	**22.644**	**<.001**	**0.030**	**0.035**
Epoch2-1	**-0.020**	**0.004**	**-4.685**	**<.001**	**-0.029**	**-0.012**
Epoch3-2	0.002	0.003	0.493	.623	-0.005	0.008
Epoch4-3	0.001	0.004	0.373	.710	-0.006	0.008
Epoch5-4	**-0.013**	**0.004**	**-3.430**	**<.001**	**-0.021**	**-0.006**
Session	**0.051**	**0.003**	**15.488**	**<.001**	**0.044**	**0.057**
Condition	0.018	0.013	1.411	.161	-0.007	0.044
Probability x Epoch2-1	**0.011**	**0.002**	**5.569**	<.001	**0.007**	**0.015**
Probability x Epoch3-2	**0.010**	**0.002**	**4.941**	<.001	**0.006**	**0.014**

Probability x Epoch4-3	**0.020**	**0.002**	**9.255**	<.001	**0.015**	**0.024**
Probability x Epoch5-4	-0.005	0.002	-2.517	.012	-0.010	-0.001
Probability x Session	**-0.004**	**0.001**	**-5.246**	<.001	**-0.005**	**-0.002**
Epoch2-1 x Session	**-0.017**	**0.002**	**-8.435**	<.001	**-0.021**	**-0.013**
Epoch3-2 x Session	-0.003	0.002	-1.391	.164	-0.007	0.001
Epoch4-3 x Session	0.000	0.002	-0.032	.975	-0.004	0.004
Epoch5-4 x Session	-0.003	0.002	-1.455	.146	-0.008	0.001
Probability x Condition	0.003	0.001	1.898	.061	0.000	0.006
Epoch2-1 x Condition	0.003	0.004	0.657	.512	-0.006	0.011
Epoch3-2 x Condition	-0.003	0.003	-1.016	.312	-0.010	0.003
Epoch4-3 x Condition	0.003	0.004	0.894	.373	-0.004	0.010
Epoch5-4 x Condition	-0.001	0.004	-0.387	.699	-0.009	0.006
Session x Condition	0.007	0.003	2.238	.028	0.001	0.014
Probability x Epoch2-1 x Session	-0.001	0.002	-0.594	.553	-0.005	0.003
Probability x Epoch3-2 x Session	-0.004	0.002	-2.065	.039	-0.008	0.000
Probability x Epoch4-3 x Session	0.004	0.002	1.786	.074	0.000	0.008
Probability x Epoch5-4 x Session	0.001	0.002	0.666	.505	-0.003	0.006
Probability x Epoch2-1 x Condition	0.003	0.002	1.687	.092	-0.001	0.007
Probability x Epoch3-2 x Condition	-0.003	0.002	-1.605	.108	-0.007	0.001
Probability x Epoch4-3 x Condition	0.002	0.002	1.133	.257	-0.002	0.007
Probability x Epoch5-4 x Condition	0.000	0.002	-0.098	.922	-0.004	0.004
Probability x Session x Condition	-0.001	0.001	-1.026	.305	-0.002	0.001
Epoch2-1 x Session x Condition	0.001	0.002	0.456	.648	-0.003	0.005
Epoch3-2 x Session x Condition	0.003	0.002	1.556	.120	-0.001	0.007
Epoch4-3 x Session x Condition	0.003	0.002	1.477	.140	-0.001	0.007
Epoch5-4 x Session x Condition	-0.003	0.002	-1.333	.183	-0.007	0.001
Probability x Epoch2-1 x Session x Condition	0.000	0.002	0.013	.990	-0.004	0.004
Probability x Epoch3-2 x Session x Condition	-0.003	0.002	-1.377	.169	-0.007	0.001
Probability x Epoch4-3 x Session x Condition	0.003	0.002	1.485	.138	-0.001	0.007

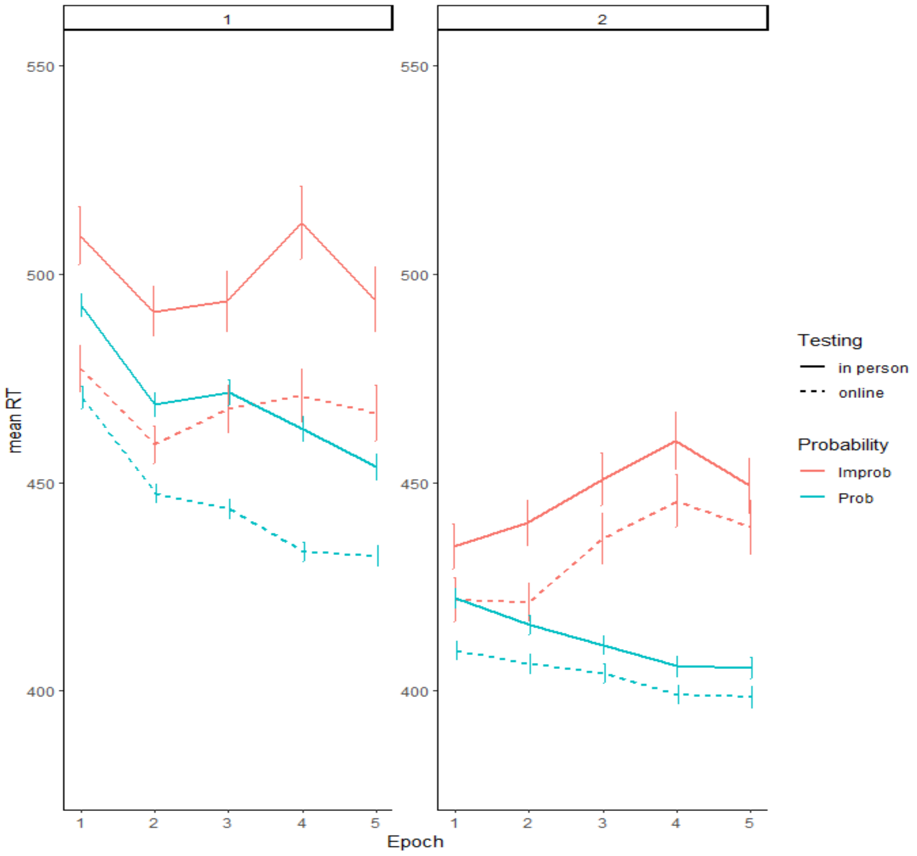
Probability x Epoch5-4 x Session x Condition	0.000	0.002	-0.020	.984	-0.004	0.004

Random effects	Variance	SD
Participant: (Intercept)	0.017	0.130
Participant: Session (slope)	0.001	0.030
Participant: Block2-1 (slope)	0.001	0.038
Participant: Block3-2 (slope)	0.001	0.027
Participant: Block4-3 (slope)	0.001	0.029
Participant: Block5-4 (slope)	0.001	0.032
Participant: Probability (slope)	0.000	0.013

Note. Indicated in bold are the contrasts that survived correction for multiple comparisons using the Holm-Bonferroni method.

Figure 6

Mean response times for probable and improbable trials per epoch and session (session 1 on the left and session 2 on the right) and testing condition (in person vs online). Bars indicate 95 % CI.



Appendix E. 3: Follow-Up experiment

Rationale

The present experiment was conducted as a follow-up to the supplementary experiment and aimed to replicate the relationship between attention and procedural learning observed in experiment 2 in an online setting and determine whether the absence of a relationship between attention and procedural learning for the noISI group in the supplementary experiment was potentially due to the lack of sensitivity of the 5-minute version.

Methods

52 participants with ages between 18 and 35 (M = 21.09; SD = 2.76) were recruited through social media. Out of the 52 participants, 11 were identified as outliers due to poor accuracy (<50%) and 1 was removed as their RTs were more than 2.5 SD above the mean of the group.

These participants completed the SRTT and PVTs in agreement with the supplementary experiment, except for the duration of the PVT which was increased to 10 minutes as in experiment 2. For the SRTT, the sequences used in the supplementary experiment were counterbalanced across participants.

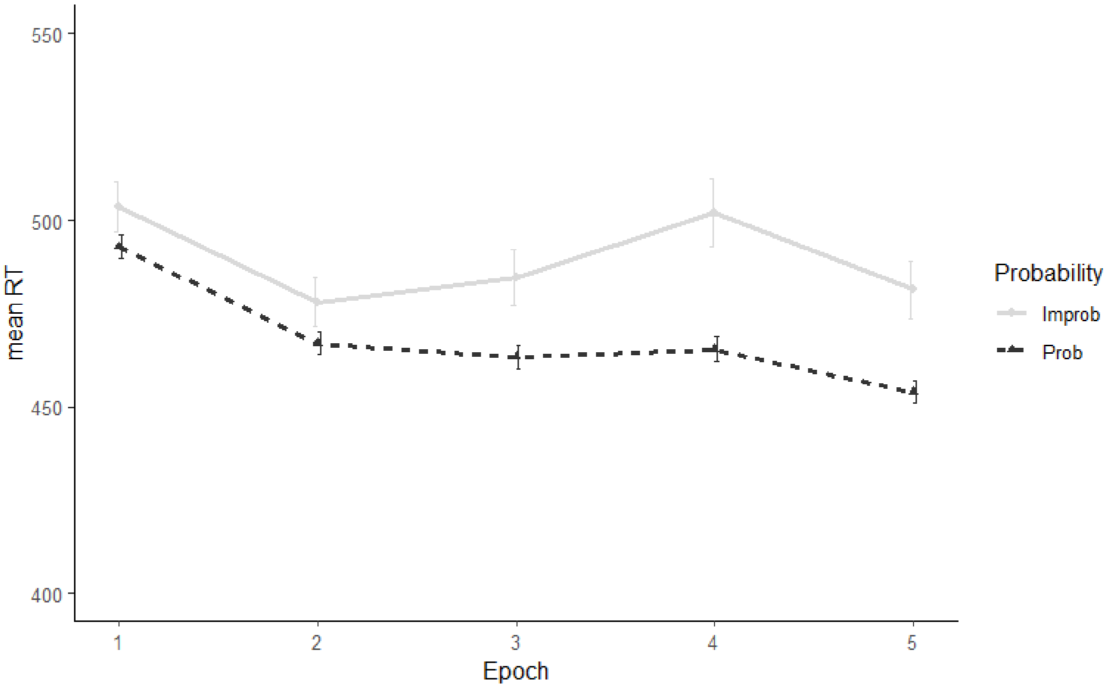
Pearson's correlations were used to assess the reliability and the relationship between procedural learning and attention as previously described.

Results

The group showed evidence of procedural learning as evidenced in figure 7 with faster response times for the probable than improbable trials.

Figure 7

Descriptive statistics - Mean and 95% CI response times for probable and improbable trials per Epoch



Split-half reliability was assessed for the SRTT and PVT using the procedures previously described in experiments 1-2. Split-half reliability was poor for the SRTT (difference scores: *r* = .34; random slopes: *r* = .41), in opposition to the PVT for which split-half reliability was adequate for all measures, with the mean being the exception (see Table 10; *r*s >.70; Burlingame et al., 1995; Nunnally & Bernstein, 1994). A positive and statistically significant relationship between attention (median) and procedural learning (random slopes) was found (*r*(30) = -.48, *p* = .005, 95% CI [-.71; -.16]).

Table 10

Pairwise split-half reliability of the 10 min version of the Psychomotor Vigilance task

Measure	***Split-half reliability***
Lapses	.86
Mean	.42
Median	.97
Reciprocal	.95

Appendix F: Experiment 2 - Cognitive Measures

Table 11

Mean (and SD) scores for all background measures

	Test	***Raw scores***			***Standardised scores***	
		Mean (SD)	Range	Max	Mean (SD)	Range
Age	-	20.09 (2.09)	17 - 34			
Matrix reasoning	WASI – II	21.71 (2.57)	14 - 26	30	52.04 (7.13)	32 - 68
Word reading	WIAT – III	67.78 (3.32)	59 - 73	75	97.75 (7.22)	81 - 115
Nonword reading	WIAT – III	42.54 (4.62)	32 - 49	52	98.11 (11.53)	72 - 114
Spelling	WIAT – III	55.51 (3.98)	47 - 63	63	109 (11.21)	78 - 134
Sentence Recall	CELF 5	69.92 (4.83)	59 - 76	26	11.26 (2.38)	7 - 15
Nonword Repetition	CToPP 2	20.53 (2.33)	16 - 26	30	55.28 (25.15)	9 - 98
Vocabulary	WASI – II	36.28 (3.24)	30 - 43	59	NA[1]	NA[1]

Note. CELF - 5 UK, Clinical Evaluation of Language Fundamentals - Fifth Edition (Wiig et al., 2013); CToPP-2, Comprehensive Test of Phonological Processing 2 (Wagner et al., 2013); WASI - II, Wechsler Abbreviated Scales of Intelligence – Second edition (Wechsler, 2011); WIAT – III, Wechsler Individual Achievement Test - Third UK Edition (Wechsler, 2009).

Appendix G: Bayesian Correlations Between Procedural Learning and Cognitive Measures

Following the recommendation from Dienes (2014) Bayesian correlations were performed to complement Pearson's correlations between procedural learning and cognitive measures, in order to allow for interpretation of non-significant results. The cut-offs suggested by Jeffreys (1998) were adopted, with a Bayes factor less than ⅓ or greater than 3 or indicating substantial evidence for, or against, the null hypothesis, respectively. Values in between were taken as representing weak evidence, though we realise that these cut-offs are only guidelines and Bayes factors should be regarded as a continuum.

Appendix G.1: Experiment 2

Table 12

Experiment 2: Bayes factors (BF_{10}) and credible intervals for pairwise correlations between procedural learning and cognitive measures. Bold indicates BF_{10} > 3 or < ⅓

Measures		**Procedural learning session 1**		**Procedural learning session 2**		**Procedural learning session 3**	
Age		0.85	(-.09, .45)	0.34	(-.30, .23)	0.35	(-.32, .23)
Literacy	Word reading	0.35	(-.25, .31)	0.47	(-.16, .38)	0.34	(-.28, .27)
	Nonword reading	0.34	(-.26, .29)	0.38	(-.20, .34)	0.36	(-.23, .33)
	Spelling	0.70	(-.11, .43)	1.11	(-.06, .46)	0.34	(-.25, .30)
Language	Vocabulary	0.39	(-.34, .21)	0.42	(-.17, .36)	**7.55**	**(.08, .57)**
	Nonword repetition	0.35	(-.31, .25)	0.40	(-.35, .19)	1.04	(-.46, .06)
	Recalling	0.53	(-.40, .15)	0.50	(-.38, .15)	1.68	(-.49, .02)
Nonverbal IQ	Matrix Reasoning	0.38	(-.34, .22)	0.39	(-.20, .35)	0.87	(-.09, .45)
Attention	Reciprocal	1.90	(-.02, .51)	**64.40**	**(.19, .64)**	1.15	(-.06, .47)
	median	1.46	(-.49, .04)	**29.88**	**(-.62, -.15)**	1.11	(-.47, .06)
	tau	.63	(-.42, .13)	.48	(-.38, .16)	.68	(-.43, .12)

Appendix G.2: Supplementary experiment

Table 13

Experiment 2: Bayes factors (BF_{10}) and credible intervals for pairwise correlations between procedural learning and cognitive measures. Bold indicates BF_{10} > 3 or < ⅓

Measures		***ISI***		***noISI***	
		Procedural learning Session 1	Procedural learning Session 2	Procedural learning Session 1	Procedural learning Session 2
Attention	PVT median Session 1	**28.85 (-.59, -.14)**	0.41 (-.21, .36)	1.47 (-.13, .66)	0.55 (-.43, .43)
	PVT median Session 2	**3.01 (-.60, -.01)**	0.44 (-.29, .39)	1.35 (-.19, .69)	0.68 (-.56, .40)
	PVT tau Session 1	0.45 (-.36, .16)	1.33 (-.49, .05)	0.70 (-.55, .29)	0.58 (-.37, .49)
	PVT tau Session 2	0.59 (-.45, .18)	0.66 (-.18, .48)	0.62 (-.42, .52)	0.71 (-.38, .58)
Explicit awareness	Inclusion	0.58 (-.43, .16)	0.38 (-.33, .28)	0.73 (-.30, .57)	0.60 (-.40, .51)
	Exclusion	0.42 (-.23, .37)	0.38 (-.28, .33)	0.56 (-.39, .47)	0.84 (-.60, .27)

Appendix G.3: Dyslexic experiment

Table 14

Dyslexic experiment: Bayes factors (BF_{10}) and credible intervals for pairwise correlations between procedural learning and cognitive measures. Bold indicates BF_{10} > 3 or < ⅓

Measures		**TD group**			**Dyslexic group**		
		Procedural learning Session 1	Procedural learning Session 2	Procedural learning Session 3	Procedural learning Session 1	Procedural learning Session 2	Procedural learning Session 3
Age		2.44 (-.54, .00)	1.18 (-.07, .49)	.39 (-.25, .34)	.42 (-.36, .19)	.36 (-.32, .24)	.94 (-.46, .08)
Literacy	Word Reading	.93 (-.09, .48)	.59 (-.42, .15)	.67 (-.45, .14)	.36 (-.25, .32)	.61 (-.14, .43)	.37 (-.26, .32)
	Nonword Reading	.67 (-.13, .44)	.43 (-.21, .37)	.36 (-.30, .30)	.36 (-.32, .23)	.35 (-.26, .31)	.37 (-.22, .34)
	Spelling	.36 (-.28, .31)	.45 (-.19, .38)	.45 (-.20, .39)	.40 (-.20, .35)	.39 (-.21, .35)	.61 (-.13, .42)
Language	Vocabulary	.69 (-.15, .46)	.99 (-.09, .49)	1.32 (-.07, .52)	.47 (-.19, .39)	.36 (-.28, .30)	.42 (-.22, .37)
	Nonword Repetition	.43 (-.22, .38)	.54 (-.16, .41)	.52 (-.18, .41)	.35 (-.26, .30)	.38 (-.34, .23)	1.93 (-.02, .52)
	Sentence Recall	.40 (-.24, .36)	.67 (-.13, .44)	.47 (-.40, .20)	.41 (-.19, .36)	1.61 (-.03, .50)	.35 (-.31, .25)
Nonverbal IQ	Matrix Reasoning	.37 (-.29, .31)	.41 (-.22, .36)	.87 (-.11, .47)	.37 (-.24, .34)	.36 (-.28, .30)	.39 (-.23, .34)
Attention	PVT median Session 1	.40 (-.36, .24)	.42 (-.21, .37)	1.26 (-.51, .06)	.40 (-.20, .36)	**3.38 (.02, .55)**	.83 (-.10, .45)

	PVT median Session 2	.37 (-.29, .32)	.45 (-.39, .20)	2.20 (-.55, .02)	.36 (-.32, .23)	.40 (-.21, .36)	.38 (-.23, .34)
	PVT median Session 3	.39 (-.34, .25)	.51 (-.40, .17)	**15.66 (-.63, -.12)**	.37 (-.33, .23)	.38 (-.23, .33)	.36 (-.24, .32)
	PVT Reciprocal Session 1	.67 (-.14, .45)	.42 (-.37, .21)	1.29 (-.06, .51)	.74 (-.11, .44)	.38 (-.34, .22)	.35 (-.30, .26)
	PVT Reciprocal Session 2	.39 (-.26, .34)	.53 (-.17, .41)	1.91 (-.03, .54)	.56 (-.40, .14)	1.44 (-.49, .04)	.50 (-.39, .16)
	PVT Reciprocal Session 3	1.03 (-.09, .49)	.36 (-.27, .31)	**10.99 (.10, .62)**	.38 (-.34, .22)	.50 (-.40, .17)	.61 (-.42, .13)
	PVT tau Session 1	.50 (-.41, .19)	.37 (-.28, .31)	**3.48 (-.58, -.03)**	.39 (-.24, .35)	.40 (-.23, .35)	.37 (-.33, .27)
	PVT tau Session 2	.99 (-.49, .09)	**5.94 (-.59, -.06)**	2.54 (-.56, .00)	.40 (-.35, .20)	1.05 (-.47, .07)	.67 (-.12, .43)
	PVT tau Session 3	.41 (-.27, .36)	.37 (-.30, .30)	.57 (-.43, .17)	1.04 (-.08, .47)	.60 (-.15, .43)	.37 (-.32, .25)
	Conners	.47 (-.39, .20)	.47 (-.39, .19)	.41 (-.36, .23)	.48 (.39, .17)	.47 (-.18, .39)	.46 (-.18, .38)
ARQ		.53 (-.17, .41)	.71 (-.44, .12)	.87 (-.10, .47)	.35 (-.31, .25)	.50 (-.16, .39)	.40 (-.35, .21)
Explicit awareness	Inclusion	**5.28 (-.60, -.05)**	.40 (-.36, .23)	.37 (-.29, .31)	.49 (-.39, .17)	.36 (-.28, .29)	.35 (-.30, .28)
	Exclusion	.47 (-.40, .20)	2.60 (-.55, .00)	**10.87 (-.62, -.10)**	.35 (-.25, .31)	.44 (-.38, .19)	.89 (-.46, .09)
Enjoyment		**3.18 (.02, .57)**	2.51 (.00, .55)	.76 (-.46, .12)	.37 (-.33, .23)	.35 (-.30, .27)	.36 (-.31, .26)

Appendix H: Descriptive Statistics for RTs

Table 15

Experiment 1: RT means and standard deviations for the SRTT per epoch and sessions (SRT1 refers to session 1 and SRT2 refers to session 2)

Epoch	***Session***			
	SRTT1		**SRTT2**	
	Prob	Improb	Prob	Improb
1	494.85	500.88	414.13	421.88
	(134.54)	(135.8)	(92.97)	(98.88)
2	472.86	493.44	408.94	421.82
	(131.71)	(139.81)	(92.92)	(96.08)
3	470.35	501.10	410.35	434.04
	(135.81)	(146.05)	(99.17)	(102.52)
4	459.72	483.52	407.29	440.61
	(129.42)	(132.77)	(100.12)	(110.43)
5	445.16	474.61	402.03	436.81
	(123.87)	(132.91)	(101.07)	(105.82)

Table 16

Experiment 2: RT means and standard deviations for the SRTT per epoch and sessions (SRTT1 refers to session 1 and SRTT2 refers to session 2, SRTT3 refers to session 3)

Epoch	***Session***					
	SRTT1		**SRTT2**		**SRTT3**	
	Prob	Improb	Prob	Improb	Prob	Improb
1	492.56	509.38	422.31	434.75	398.94	415.08
	(126.88)	(133.13)	(96.84)	(101.70)	(92.76)	(94.43)
2	468.83	491.09	415.84	440.41	399.91	425.65
	(123.53)	(125.37)	(102.67)	(110.05)	(100.22)	(103.90)
3	471.59	493.58	410.89	450.78	393.90	435.28
	(133.86)	(135.69)	(103.25)	(121.90)	(97.13)	(112.20)
4	462.84	512.40	405.91	460.01	391.33	445.60
	(135.50)	(152.50)	(109.84)	(125.36)	(100.28)	(110.75)
5	453.76	493.95	405.52	449.32	384.66	431.95
	(137.66)	(141.27)	(111.58)	(117.72)	(102.37)	(109.40)

Table 17

Supplementary experiment: RT means and standard deviations for the SRTT per epoch and sessions (SRTT1 refers to session 1 and SRTT2 refers to session 2)

Epoch	***ISI***				***noISI***			
	SRTT1		**SRTT2**		**SRTT1**		**SRTT2**	
	Prob	Improb	Prob	Improb	Prob	Improb	Prob	Improb
1	445.51	460.47	386.17	395.68	470.41	477.46	409.63	421.81
	(168.84)	(214.45)	(111.87)	(105.50)	(132.35)	(129.15)	(99.76)	(105.94)
2	415.08	432.58	375.70	393.76	447.30	459.13	406.53	421.33
	(127.60)	(127.27)	(100.65)	(106.42)	(114.52)	(111.06)	(98.98)	(94.28)
3	407.74	417.42	374.90	393.53	443.71	467.64	404.24	436.60
	(124.39)	(129.88)	(106.07)	(165.54)	(122.53)	(129.90)	(102.86)	(114.70)
4	396.79	443.65	372.91	397.77	433.38	470.94	399.13	445.66
	(129.15)	(161.75)	(113.36)	(100.36)	(118.12)	(126.45)	(103.25)	(107.39)
5	394.27	427.32	369.83	396.61	432.40	466.73	398.62	439.22
	(124.71)	(180.86)	(118.86)	(106.81)	(127.25)	(137.92)	(113.59)	(114.81)

Table 18

Dyslexic experiment: RT means and standard deviations for the SRTT per epoch and sessions (SRTT1 refers to session 1, SRTT2 refers to session 2 and SRTT3 refers to session 3)

Epoch	***TD group***						***Dyslexic group***					
	SRTT1		SRTT2		SRTT3		SRTT1		SRTT2		SRTT3	
	Prob	Improb	Prob	Improb	Prob	Improb	Prob	Improb	Prob	Improb	Prob	Improb
1	530.52	533.00	439.77	451.74	409.95	424.63	546.48	578.69	454.68	466.88	433.64	450.94
	(181.59)	(172.05)	(109.88)	(117.42)	(101.42)	(104.59)	(214.45)	(271.71)	(108.63)	(104.53)	(101.52)	(108.33)
2	496.08	515.59	433.03	459.17	411.16	436.58	522.29	537.43	451.44	469.27	430.78	456.93
	(165.65)	(165.86)	(120.37)	(122.66)	(105.82)	(112.86)	(191.92)	(174.47)	(111.91)	(109.05)	(101.23)	(110.57)
3	481.03	506.73	425.46	458.25	404.69	434.88	515.27	540.67	448.75	475.49	427.86	456.33
	(155.02)	(168.72)	(111.92)	(120.33)	(94.57)	(100.11)	(168.98)	(173.28)	(114.41)	(134.18)	(107.90)	(114.73)
4	472.12	516.37	423.63	469.07	398.78	450.35	508.83	545.49	445.11	493.08	435.47	484.60
	(144.85)	(162.82)	(120.29)	(133.42)	(94.85)	(113.21)	(164.64)	(161.59)	(115.04)	(134.46)	(124.07)	(134.35)
5	470.15	497.99	420.91	470.41	407.66	450.23	507.62	548.88	451.48	483.14	425.28	475.06
	(150.42)	(151.27)	(118.00)	(144.23)	(106.32)	(115.25)	(165.84)	(212.20)	(126.53)	(118.07)	(109.21)	(132.44)

Appendix I: Experiment 1 - Enjoyment

Enjoyment of performance on the SRTT was analysed for the second session by asking participants to rate their levels of enjoyment using a Likert scale from 1 to 10, with 1 representing “not enjoyable at all” and 10 representing “very enjoyable”. The mean rating of enjoyment was 6.70.

Correlations between procedural learning for both sessions and self-rated enjoyment of the SRTT were conducted. There was no association between enjoyment and procedural learning at both time points (session 1: r = .07, p = .527, BF_{10} = .28; session 2: r = -.07, p = .477, BF_{10} = .30).

References

*References marked with a dagger (†) indicate studies included in the meta-analysis on the reliability of the SRTT (**Chapter 3**), whilst those marked with an asterisk (*) were included in the meta-analysis on the relationship between procedural learning and language/literacy (**Chapter 4**).*

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alexander-Passe, N. (2006). How dyslexic teenagers cope: An investigation of self-esteem, coping and depression. *Dyslexia*, *12*(4), 256–275. https://doi.org/10.1002/dys.318

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Brooks/Cole Publishing Company.

Ambrus, G. G., Vékony, T., Janacsek, K., Trimborn, A. B. C., Kovács, G., & Nemeth, D. (2020). When less is more: Enhanced statistical learning of non-adjacent dependencies after disruption of bilateral DLPFC. *Journal of Memory and Language*, *114*, 104144. https://doi.org/10.1016/j.jml.2020.104144

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). *Gorilla in our midst: An online behavioral experiment builder*. 20. https://doi.org/10.3758/s13428-019-01237-x

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. https://doi.org/10.1037/a0029839

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

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

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

Arfé, B., Cona, E., & Merella, A. (2018). Training Implicit Learning of Spelling in Italian Children With Developmental Dyslexia. *Topics in Language Disorders*, *38*(4), 299–315.

https://doi.org/10.1097/TLD.0000000000000163

Arnon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods*, *52*(1), 68–81. https://doi.org/10.3758/s13428-019-01205-5

Ashby, F. G., & Maddox, W. T. (2011). Human Category Learning 2.0. *Annals of the New York Academy of Sciences*, *1124*, 147–161. https://doi.org/10.1111/j.1749-6632.2010.05874.x

Ashby, F. G., Turner, B. O., & Horvitz, J. C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Sciences*, *14*(5), 208–215. https://doi.org/10.1016/j.tics.2010.02.001

Astle, D. E., Bathelt, J., The CALM Team, & Holmes, J. (2019). Remapping the cognitive and neural profiles of children who struggle at school. *Developmental Science*, *22*(1), e12747. https://doi.org/10.1111/desc.12747

Attout, L., Ordonez Magro, L., Szmalec, A., & Majerus, S. (2020). The developmental neural substrates of Hebb repetition learning and their link with reading ability. *Human Brain Mapping*, *41*(14), 3956–3969. https://doi.org/10.1002/hbm.25099

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, *16*(8), 437–443. https://doi.org/10.1016/j.tics.2012.06.010

Badcock, N. A., Bishop, D. V. M., Hardiman, M. J., Barry, J. G., & Watkins, K. E. (2012). Brain & Language Co-localisation of abnormal brain structure and function in specific language impairment. *Brain and Language*, *120*(3), 310–320. https://doi.org/10.1016/j.bandl.2011.10.006

Baird, G., Dworzynski, K., Slonims, V., & Simonoff, E. (2010). Memory impairment in children with language impairment. *Developmental Medicine and Child Neurology*, *52*(6), 535–540. https://doi.org/10.1111/j.1469-8749.2009.03494.x

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*(3), 295–314. https://doi.org/10.1037/met0000337

Baker, R. S. J. d., D’Mello, S. K., Rodrigo, Ma. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners’ cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241.

https://doi.org/10.1016/j.ijhcs.2009.12.003

Barker, L. (2012). Defining the Parameters of Incidental Learning on a Serial Reaction Time (SRT) Task: Do Conscious Rules Apply? *Brain Sciences*, *2*(4), 769–789. https://doi.org/10.3390/brainsci2040769

Barnes, K. A., Howard, J. H., Howard, D. V., Kenealy, L., & Vaidya, C. J. (2010). Two Forms of Implicit Learning in Childhood ADHD. *Developmental Neuropsychology*, *35*(5), 494–505. https://doi.org/10.1080/87565641.2010.494750

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. https://doi.org/10.1016/j.jml.2012.11.001

Basner, M., & Dinges, D. F. (2011). Maximizing sensitivity of PVT to Sleep Loss (Basner, Dinges). *Sleep*, *34*(5), 581–591.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Batterink, L. J., Paller, K. A., & Reber, P. J. (2019). Understanding the Neural Bases of Implicit and Statistical Learning. *Topics in Cognitive Science*, tops.12420. https://doi.org/10.1111/tops.12420

Bauer, P. J. (2008). Toward a neuro-developmental account of the development of declarative memory. *Developmental Psychobiology*, *50*(1), 19–31. https://doi.org/10.1002/dev.20265

Baugh, F. (2002). Correcting Effect Sizes for Score Reliability: A Reminder that Measurement and Substantive Issues are Linked Inextricably. *Educational and Psychological Measurement*, *62*(2), 254–263.

Begg, C. B., & Mazumdar, M. (1994). Operating Characteristics of a Rank Correlation Test for Publication Bias. *Biometrics*, *50*(4), 1088. https://doi.org/10.2307/2533446

Beglinger, L., Gaydos, B., Tangphaodaniels, O., Duff, K., Kareken, D., Crawford, J., Fastenau, P., & Siemers, E. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology*, *20*(4), 517–529. https://doi.org/10.1016/j.acn.2004.12.003

Bennett, I. J., Romano, J. C., Howard, Jr, J. H., & Howard, D. V. (2008). Two Forms of Implicit Learning in Young Adults with Dyslexia. *Annals of the New York Academy of Sciences*, *1145*(1), 184–198. https://doi.org/10.1196/annals.1416.006

Berchtold, A. (2016). Test–retest: Agreement or reliability? *Methodological Innovations*, *9*, 205979911667287. https://doi.org/10.1177/2059799116672875

Berger, B., Waterman, M. S., & Yu, Y. W. (2021). Levenshtein Distance, Sequence Comparison and Biological Database Search. *IEEE Transactions on Information Theory*, *67*(6), 3287–3294. https://doi.org/10.1109/TIT.2020.2996543

Bischoff-Grethe, A., Goedert, K. M., Willingham, D. T., & Grafton, S. T. (2004). Neural Substrates of Response-based Sequence Learning using fMRI. *Journal of Cognitive Neuroscience*, *16*(1), 127–138. https://doi.org/10.1162/089892904322755610

Bitan, T., & Karni, A. (2004). Procedural and declarative knowledge of word recognition and letter decoding in reading an artificial script. *Cognitive Brain Research*, *19*(3), 229–243. https://doi.org/10.1016/j.cogbrainres.2004.01.001

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *1*(8476), 308–310.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Bland, J. M., & Altman, D. G. (2003). Applying the right statistics: Analyses of measurement studies. *Ultrasound in Obstetrics and Gynecology*, *22*(1), 85–93. https://doi.org/10.1002/uog.122

Bland, J. M., & Altman, D. G. (2010). Statistical methods for assessing agreement between two methods of clinical measurement. *International Journal of Nursing Studies*, 6. https://doi.org/10.1016/j.ijnurstu.2009.10.001

Bo, J., Jennett, S., & Seidler, R. D. (2011). Working memory capacity correlates with implicit serial reaction time task performance. *Experimental Brain Research*, *214*(1), 73–81. https://doi.org/10.1007/s00221-011-2807-8

Boen, R., Ferschmann, L., Vijayakumar, N., Overbye, K., Fjell, A. M., Espeseth, T., & Tamnes, C. K. (2021). Development of attention networks from childhood to young adulthood: A study of performance, intraindividual variability and cortical thickness. *Cortex*, *138*, 138–151. https://doi.org/10.1016/j.cortex.2021.01.018

Bogaerts, L., Siegelman, N., Ben-Porat, T., & Frost, R. (2018). Is the Hebb repetition task a reliable measure of individual differences in sequence learning? *Quarterly Journal of Experimental Psychology (2006)*, *71*(4), 892–905. https://doi.org/10.1080/17470218.2017.1307432

Bogaerts, L., Siegelman, N., & Frost, R. (2021). Statistical Learning and Language Impairments: Toward More Precise Theoretical Accounts. *Perspectives on Psychological Science*, *16*(2), 319–337. https://doi.org/10.1177/1745691620953082

Borella, E., Chicherio, C., Re, A. M., Sensini, V., & Cornoldi, C. (2011). Increased intraindividual variability is a marker of ADHD but also of dyslexia: A study on handwriting. *Brain and Cognition*, *77*(1), 33–39. https://doi.org/10.1016/j.bandc.2011.06.005

Borestein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of Research Synthesis and Meta-Analysis* (pp. 221–236). Russell Sage Foundation.

Borragán, G., Slama, H., Destrebecqz, A., & Peigneux, P. (2016). Cognitive Fatigue Facilitates Procedural Sequence Learning. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00086

Brewer, M. J., Butler, A., & Cooksley, S. L. (2016). The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, *7*(6), 679–692. https://doi.org/10.1111/2041-210X.12541

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414–e9414. https://doi.org/10.7717/peerj.9414

Brookman, A., McDonald, S., McDonald, D., & Bishop, D. V. M. (2013). Fine motor deficits in reading disability and language impairment: Same or different? *PeerJ*, *1*, e217–e217. https://doi.org/10.7717/peerj.217

Brosowsky, N. P., Murray, S., Schooler, J. W., & Seli, P. (2021). Attention need not always apply: Mind wandering impedes explicit but not implicit sequence learning. *Cognition*, *209*, 104530. https://doi.org/10.1016/j.cognition.2020.104530

† Brown, J. (2010). *An Analysis of Functional Differences in Implicit Learning* [University of Cambridge]. https://doi.org/10.17863/CAM.16467

Brown, R. M., Robertson, E. M., & Press, D. Z. (2009). Sequence Skill Acquisition and Off-Line Learning in Normal Aging. *PLoS ONE*, *4*(8), e6683. https://doi.org/10.1371/journal.pone.0006683

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(1), 1–20. https://doi.org/10.5334/joc.10

Buchholz, J., & McKone, E. (2004). Adults with dyslexia show deficits on spatial frequency doubling and visual attention tasks. *Dyslexia*, *10*(1), 24–43. https://doi.org/10.1002/dys.263

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology*, *40*(3), 227–259. https://doi.org/10.1006/cogp.1999.0731

Buffington, J., Demos, A. P., & Morgan-Short, K. (2021). The Reliability And Validity Of Procedural

Memory Assessments Used In Second Language Acquisition Research. *Studies in Second Language Acquisition*, *43*(3), 635–662. https://doi.org/10.1017/S0272263121000127

Burlingame, G. M., Lambert, M. J., Reisinger, C. W., Neff, W. M., & Mosier, J. (1995). Pragmatics of tracking mental health outcomes in a managed care setting. *The Journal of Mental Health Administration*, *22*(3), 226–236. https://doi.org/10.1007/BF02521118

Burton, L., Nunes, T., & Evangelou, M. (2021). Do children use logic to spell logician? Implicit versus explicit teaching of morphological spelling rules. *British Journal of Educational Psychology*, *91*(4), 1231–1248. https://doi.org/10.1111/bjep.12414

Butterworth, B., & Kovas, Y. (2013). Understanding neurocognitive developmental disorders can improve education for all. *Science (New York, N.Y.)*, *340*(6130), 300–305. https://doi.org/10.1126/science.1231022

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring Higher the Second Time Around: Meta-Analyses of Practice Effects in Neuropsychological Assessment. *The Clinical Neuropsychologist*, *26*(4), 543–570. https://doi.org/10.1080/13854046.2012.680913

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test-retest correlations. *Clinical Neuropsychologist*, *27*(7), 1077–1105. https://doi.org/10.1080/13854046.2013.809795

Castellanos, F. X., Sonuga-Barke, E. J. S., Milham, M. P., & Tannock, R. (2006). Characterizing cognition in ADHD: Beyond executive dysfunction. *Trends in Cognitive Sciences*, *10*(3), 117–123. https://doi.org/10.1016/j.tics.2006.01.011

Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Castro-schilo, L., & Grimm, K. J. (2018). *Using residualized change versus difference scores for longitudinal research*. *35*(1), 32–58. https://doi.org/10.1177/0265407517718387

Chen, G., Pine, D. S., Brotman, M. A., Smith, A. R., Cox, R. W., & Haller, S. P. (2021). Trial and error: A hierarchical modeling approach to test-retest reliability. *NeuroImage*, *245*, 118647. https://doi.org/10.1016/j.neuroimage.2021.118647

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

Chomsky, N. (1980). *Rules and representations*. Columbia University Press.

Cicchetti, D. V. (1994a). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cicchetti, D. V. (1994b). Interreliability Standards in Psychological Evaluations. *Psychological Assessment*, *4*, 284–290.

Cicchetti, D. V., & Sparrow, S. S. (1990). Assessment of adaptive behavior in young children. In *Developmental assessment in clinical child psychology: A handbook.* (pp. 173–196). Pergamon Press.

Clark, G. M., & Lum, J. A. G. (2017a). Procedural learning in Parkinson's disease, specific language impairment, dyslexia, schizophrenia, developmental coordination disorder, and autism spectrum disorders: A second-order meta-analysis. *Brain and Cognition*. https://doi.org/10.1016/j.bandc.2017.07.004

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

* Clark, G. M., & Lum, J. A. G. (2017c). First-Order and higher order sequence learning in specific language impairment. Neuropsychology, 31(2), 149–159. https://doi.org/10.1037/neu0000316

Clark, G. M., Lum, J. A. G., & Ullman, M. T. (2014). A meta-analysis and meta-regression of serial reaction time task performance in Parkinson's disease. *Neuropsychology*, *28*(6), 945–958. https://doi.org/10.1037/neu0000121

Cleeremans, A. (2008). Consciousness: The radical plasticity thesis. *Progress in Brain Research*, *168*, 19–33. https://doi.org/10.1016/S0079-6123(07)68003-0

Cleeremans, A. (2011). The Radical Plasticity Thesis: How the Brain Learns to be Conscious. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00086

Cleeremans, A., & Jiménez, L. (2002). Implicit learning and concsciousness: A graded, dynamic perspective. *Implicit Learning and Concsciousness. An Empirical, Philosophical and Computational Consensus in the Making*, 1–40.

Cleeremans, A., & Sarrazin, J.-C. (2007). Time, action, and consciousness. *Human Movement Science*, *26*(2), 180–202. https://doi.org/10.1016/j.humov.2007.01.009

Clegg, J., Hollis, C., Mawhood, L., & Rutter, M. (2005). Developmental language disorders—A follow-up in later adult life. Cognitive, language and psychosocial outcomes. *Journal of Child Psychology and Psychiatry*, *46*(2), 128–149. https://doi.org/10.1111/j.1469-7610.2004.00342.x

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129. https://doi.org/10.2307/3001666

Cohen, J. D., McClelland, J. L., & Dunbar, K. (1990). On the Control of Automatic Processes: A Parallel Distributed Processing Account of the Stroop Effect. *Psychological Review*, *97*(3), 332–361. https://doi.org/10.1037/0033-295X.97.3.332

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science (New York, N.Y.)*, *210*(4466), 207–210. https://doi.org/10.1126/science.7414331

Conners, C. K., Epstein, J. N., Angold, A., & Klaric, J. (2003). *Continuous Performance Test Performance in a Normative Epidemiological Sample*. 8.

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. https://doi.org/10.1111/1460-6984.12338

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

Conway, C. M., Arciuli, J., Lum, J. A. G., & Ullman, M. T. (2019). *Seeing problems that may not exist: A reply to West et al.'s (2018) questioning of the procedural deficit hypothesis*. 6.

Coomans, D., Vandenbossche, J., & Deroost, N. (2014). The effect of attentional load on implicit sequence learning in children and young adults. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00465

Coppola, D. M., Purves, H. R., McCoy, A. N., & Purves, D. (1998). The distribution of oriented contours in the real world. *Proceedings of the National Academy of Sciences*, *95*(7), 4002–4006. https://doi.org/10.1073/pnas.95.7.4002

Cordewener, K. A. H., Bosman, A. M. T., & Verhoeven, L. (2015). Implicit and explicit instruction: The case of spelling acquisition. *Written Language & Literacy*, *18*(1), 121–152. https://doi.org/10.1075/wll.18.1.06cor

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review. *International Journal of Psychological Research*, *3*(1), 58–67. https://doi.org/10.21500/20112084.844

Coynel, D., Marrelec, G., Perlbarg, V., Pélégrini-Issac, M., Van de Moortele, P.-F., Ugurbil, K., Doyon, J., Benali, H., & Lehéricy, S. (2010). Dynamics of motor-related functional integration during motor sequence learning. *NeuroImage*, *49*(1), 759–766. https://doi.org/10.1016/j.neuroimage.2009.08.048

Cronbach, L. J. (1990). *Essentials of Psychological Testing* (5th ed.). Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423. https://doi.org/10.1016/S1364-6613(03)00197-9

Cycowicz, Y. M., Friedman, D., Snodgrass, J. G., & Duff, M. (2001). Recognition and source memory for pictures in children and adults. *Neuropsychologia*, *39*(3), 255–267. https://doi.org/10.1016/S0028-3932(00)00108-1

Dahms, C., Brodoehl, S., Witte, O. W., & Klingner, C. M. (2020). The importance of different learning stages for motor sequence learning after stroke. *Human Brain Mapping*, *41*(1), 270–286. https://doi.org/10.1002/hbm.24793

Danner, D., Hagemann, D., & Funke, J. (2017). Measuring Individual Differences in Implicit Learning with Artificial Grammar Learning Tasks: Conceptual and Methodological Conundrums. *Zeitschrift Für Psychologie*, *225*(1), 5–19. https://doi.org/10.1027/2151-2604/a000280

de Diego-Balaguer, R., Martinez-Alvarez, A., & Pons, F. (2016). Temporal Attention as a Scaffold for Language Development. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00044

de Guibert, C., Maumet, C., Jannin, P., Ferré, J.-C., Tréguier, C., Barillot, C., Le Rumeur, E., Allaire, C., & Biraben, A. (2011). Abnormal functional lateralization and activity of language brain areas in typical specific language impairment (developmental dysphasia). *Brain*, *134*(10), 3044–3058. https://doi.org/10.1093/brain/awr141

Del Re, A., & Hoyt, W. T. (2018). *MAc: Meta-Analysis with Correlations*. https://CRAN.R-project.org/package=MAc

Delage, H., & Frauenfelder, U. H. (2020). Relationship between working memory and complex syntax in children with Developmental Language Disorder. *Journal of Child Language*, *47*(3), 600–632. https://doi.org/10.1017/S0305000919000722

* 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. https://doi.org/10.1080/13803390903313556

* Desmottes, L., Maillart, C., & Meulemans, T. (2017). Memory consolidation in children with specific language impairment: Delayed gains and susceptibility to interference in implicit sequence

learning. *Journal of Clinical and Experimental Neuropsychology*, *39*(3), 265–285. https://doi.org/10.1080/13803395.2016.1223279

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

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

Destrebecqz, A., & Cleeremans, A. (2001). Can sequence learning be implicit? New evidence with the process dissociation procedure. *Psychonomic Bulletin & Review*, *8*(2), 343–350. https://doi.org/10.3758/BF03196171

Destrebecqz, A., & Cleeremans, A. (2003). Temporal effects in sequence learning. In *Attention and implicit learning.* (pp. 181–213). John Benjamins Publishing Company. https://doi.org/10.1075/aicr.48.11des

Destrebecqz, A., Peigneux, P., Laureys, S., Degueldre, C., Del Fiore, G., Aerts, J., Luxen, A., Van Der Linden, M., Cleeremans, A., & Maquet, P. (2005). The neural correlates of implicit and explicit sequence learning: Interacting networks revealed by the process dissociation procedure. *Learning & Memory*, *12*(5), 480–490. https://doi.org/10.1101/lm.95605

Devoto, F., Carioti, D., Danelli, L., & Berlingeri, M. (2021). A meta-analysis of functional neuroimaging studies on developmental dyslexia across European orthographies: The ADOD model. *Language, Cognition and Neuroscience*, 1–30. https://doi.org/10.1080/23273798.2021.1970200

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z., Broadbent, D., & Berry, D. (1991). Implicit and Explicit Knowledge Bases in Artificial Grammar Learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *17*(5), 875–887. https://doi.org/10.1037/0278-7393.17.5.875

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test–retest reliability and practice effects of Expanded Halstead–Reitan Neuropsychological Test Battery. *Journal of the International Neuropsychological Society*, *5*(4), 346–356. https://doi.org/10.1017/S1355617799544056

Dorrian, J., Rogers, N. L., & Dinges, D. F. (2005). Psychomotor vigilance performance: Neurocognitive assay sensitive to sleep loss. In C. A. Kushida (Ed.), *Sleep Deprivation: Clinical Issues, Pharmacology, and Sleep Loss Effects* (pp. 39–70). Marcel Dekker.

Doyon, J., Bellec, P., Amsel, R., Penhune, V., Monchi, O., Carrier, J., Lehéricy, S., & Benali, H. (2009). Contributions of the basal ganglia and functionally related brain structures to motor learning. *Behavioural Brain Research*, *199*(1), 61–75. https://doi.org/10.1016/j.bbr.2008.11.012

Doyon, J., & Benali, H. (2005). Reorganization and plasticity in the adult brain during learning of motor skills. *Current Opinion in Neurobiology*, *15*(2), 161–167. https://doi.org/10.1016/j.conb.2005.03.004

Doyon, J., Owen, A. M., Petrides, M., Sziklas, V., & Evans, A. C. (1996). Functional Anatomy of Visuomotor Skill Learning in Human Subjects Examined with Positron Emission Tomography. *European Journal of Neuroscience*, *8*(4), 637–648. https://doi.org/10.1111/j.1460-9568.1996.tb01249.x

Doyon, J., Song, A. W., Karni, A., Lalonde, F., Adams, M. M., & Ungerleider, L. G. (2002). Experience-dependent changes in cerebellar contributions to motor sequence learning. *Proceedings of the National Academy of Sciences*, *99*(2), 1017–1022. https://doi.org/10.1073/pnas.022615199

Draganski, B., Kherif, F., Kloppel, S., Cook, P. A., Alexander, D. C., Parker, G. J. M., Deichmann, R., Ashburner, J., & Frackowiak, R. S. J. (2008). Evidence for Segregated and Integrative Connectivity Patterns in the Human Basal Ganglia. *Journal of Neuroscience*, *28*(28), 7143–7152. https://doi.org/10.1523/JNEUROSCI.1486-08.2008

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. https://doi.org/10.1007/s11881-012-0077-1

Duff, K. (2012). Evidence-Based Indicators of Neuropsychological Change in the Individual Patient: Relevant Concepts and Methods. *Archives of Clinical Neuropsychology*, *27*(3), 248–261. https://doi.org/10.1093/arclin/acr120

Eadie, P., Conway, L., Hallenstein, B., Mensah, F., McKean, C., & Reilly, S. (2018). Quality of life in children with developmental language disorder: Quality of life in children with DLD. *International Journal of Language & Communication Disorders*, *53*(4), 799–810. https://doi.org/10.1111/1460-6984.12385

Earle, F. S., & Ullman, M. T. (2021). Deficits of Learning in Procedural Memory and Consolidation in Declarative Memory in Adults With Developmental Language Disorder. *Journal of Speech,* *Language, and Hearing Research*, *64*(February), 1–11. https://doi.org/10.1044/2020_jslhr-20-00292

Ebert, K. D., & Kohnert, K. (2011). Sustained Attention in Children With Primary Language Impairment: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1372–1384. https://doi.org/10.1044/1092-4388(2011/10-0231)

Eckert, M. (2004). Neuroanatomical Markers for Dyslexia: A Review of Dyslexia Structural Imaging Studies. *The Neuroscientist*, *10*(4), 362–371. https://doi.org/10.1177/1073858404263596

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

Eichenbaum, H. (2002). *The cognitive neuroscience of memory: An introduction*. Oxford University Press.

Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain*. Oxford Univ.

Emberson, L. L., Conway, C. M., & Christiansen, M. H. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *Quarterly Journal of Experimental Psychology*, *64*(5), 1021–1040. https://doi.org/10.1080/17470218.2010.538972

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(12), 5472–5477. https://doi.org/10.1073/pnas.1818430116

Erickson, L. C., Kaschak, M. P., Thiessen, E. D., & Berry, C. A. S. (2016). Individual Differences in Statistical Learning: Conceptual and Measurement Issues. *Collabra*, *2*(1), 14. https://doi.org/10.1525/collabra.41

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

Eriksen, N., & Tougaard, J. (2006). Analysing differences among animal songs quantitatively by means of the Levenshtein distance measure. *Behaviour*, *143*(2), 239–252. https://doi.org/10.1163/156853906775900685

Esser, S., & Haider, H. (2017). The Emergence of Explicit Knowledge in a Serial Reaction Time Task: The Role of Experienced Fluency and Strength of Representation. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00502

Esser, S., Lustig, C., & Haider, H. (2021). What triggers explicit awareness in implicit sequence learning? Implications from theories of consciousness. *Psychological Research*. https://doi.org/10.1007/s00426-021-01594-3

Faes, J., Gillis, J., & Gillis, S. (2016). Phonemic accuracy development in children with cochlear implants up to five years of age by using Levenshtein distance. *Journal of Communication Disorders*, *59*, 40–58. https://doi.org/10.1016/j.jcomdis.2015.09.004

Failing, M., & Theeuwes, J. (2018). Selection history: How reward modulates selectivity of visual attention. *Psychonomic Bulletin & Review*, *25*(2), 514–538. https://doi.org/10.3758/s13423-017-1380-y

† Feldman, J., Kerr, B., & Streissguth, A. P. (1995). Correlational analyses of procedural and declarative learning performance. *Intelligence*, *20*(1), 87–114. https://doi.org/10.1016/0160-2896(95)90007-1

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In *Educational measurement, 3rd ed.* (pp. 105–146). American Council on Education.

Finn, A. S., Kalra, P. B., Goetz, C., Leonard, J. A., Sheridan, M. A., & Gabrieli, J. D. E. (2016). Developmental dissociation between the maturation of procedural memory and declarative memory. *Journal of Experimental Child Psychology*, *142*, 212–220. https://doi.org/10.1016/j.jecp.2015.09.027

Fischer, S., Drosopoulos, S., Tsen, J., & Born, J. (2006). Implicit learning—Explicit knowing: A role for sleep in memory system interaction. *Journal of Cognitive Neuroscience*, *18*(3), 311–319.

Fisher, Z., & Tipton, E. (2015). robumeta: An R-package for robust variance estimation in meta-analysis. *ArXiv:1503.02220 [Stat]*. http://arxiv.org/abs/1503.02220

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. John Wiley & Sons.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT Press.

Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences*, *103*(31), 11778–11783. https://doi.org/10.1073/pnas.0602659103

Forkstam, C., Hagoort, P., Fernandez, G., Ingvar, M., & Petersson, K. M. (2006). Neural correlates of artificial syntactic structure classification. *NeuroImage*, *32*(2), 956–967. https://doi.org/10.1016/j.neuroimage.2006.03.057

Fortenbaugh, F. C., DeGutis, J., & Esterman, M. (2017). Recent theoretical, neural, and clinical advances in sustained attention research: Sustained attention. *Annals of the New York Academy of Sciences*, *1396*(1), 70–91. https://doi.org/10.1111/nyas.13318

Fortenbaugh, F. C., DeGutis, J., Germine, L., Wilmer, J. B., Grosso, M., Russo, K., & Esterman, M.

(2015). Sustained Attention Across the Life Span in a Sample of 10,000: Dissociating Ability and Strategy. *Psychological Science*, *26*(9), 1497–1510. https://doi.org/10.1177/0956797615594896

Fostick, L., & Revah, H. (2018). Dyslexia as a multi-deficit disorder: Working memory and auditory temporal processing. *Acta Psychologica*, *183*(April 2017), 19–28. https://doi.org/10.1016/j.actpsy.2017.12.010

Franklin, M. S., Smallwood, J., Zedelius, C. M., Broadway, J. M., & Schooler, J. W. (2016). Unaware yet reliant on attention: Experience sampling reveals that mind-wandering impedes implicit learning. *Psychonomic Bulletin and Review*, *23*(1), 223–229. https://doi.org/10.3758/s13423-015-0885-5

Frensch, P. A., Buchner, A., & Lin, J. (1994). Implicit Learning of Unique and Ambiguous Serial Transitions in the Presence and Absence of a Distractor Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 567–584. https://doi.org/10.1037/0278-7393.20.3.567

Frensch, P. A., Haider, H., Rünger, D., Neugebauer, U., Voigt, S., & Werg, J. (2003). Verbal report of incidentally experienced environmental regularity: The route from implicit learning to verbal expression of what has been learned. In L. Jimenez (Ed.), *Attention and implicit learning* (Ed., pp. 335–366). Benjamins.

Frensch, P. A., Lin, J., & Buchner, A. (1998). Learning versus behavioral expression of the learned: The effects of a secondary tone-counting task on implicit learning in the serial reaction task. *Psychological Research*, *61*(2), 83–98. https://doi.org/10.1007/s004260050015

Frensch, P. A., & Miner, C. S. (1994). Effects of presentation rate and individual differences in short-term memory capacity on an indirect measure of serial learning. *Memory & Cognition*, *22*(1), 95–110. https://doi.org/10.3758/BF03202765

Frensch, P. A., Wenke, D., & Riinger, D. (1999). A Secondary Tone-Counting Task Suppresses Expression of Knowledge in the Serial Reaction Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 260–274. https://psycnet.apa.org/doi/10.1037/0278-7393.25.1.260

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. https://doi.org/10.1037/bul0000210

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction* (pp. xvi, 349). Sage Publications, Inc.

Gabay, Y., Schiff, R., & Vakil, E. (2012a). Attentional requirements during acquisition and consolidation of a skill in normal readers and developmental dyslexics. *Neuropsychology*, *26*(6), 744–757. https://doi.org/10.1037/a0030235

* Gabay, Y., Schiff, R., & Vakil, E. (2012b). Dissociation between the procedural learning of letter names and motor sequences in developmental dyslexia. *Neuropsychologia*, *50*(10), 2435–2441. https://doi.org/10.1016/j.neuropsychologia.2012.06.014

* Gabriel, A., Maillart, C., Guillaume, M., Stefaniak, N., & Meulemans, T. (2011). Exploration of Serial Structure Procedural Learning in Children with Language Impairment. *Journal of the International Neuropsychological Society*, *17*(2), 336–343. https://doi.org/10.1017/S1355617710001724

* Gabriel, A., Maillart, C., Stefaniak, N., Lejeune, C., Desmottes, L., & Meulemans, T. (2013). Procedural Learning in Specific Language Impairment: Effects of Sequence Complexity. *Journal of the International Neuropsychological Society*, *19*(3), 264–271. https://doi.org/10.1017/S1355617712001270

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

* Gabriel, A., Stefaniak, N., Maillart, C., Schmitz, X., & Meulemans, T. (2012). Procedural Visual Learning in Children With Specific Language Impairment. *American Journal of Speech-Language Pathology*, *21*(4), 329–341. https://doi.org/10.1044/1058-0360(2012/11-0044)

Gaffan, D. (1974). Recognition impaired and association intact in the memory of monkeys after transection of the fornix. *Journal of Comparative and Physiological Psychology*, *86*(6), 1100–1109. https://doi.org/10.1037/h0037649

Gagnon, S., Foster, J., Turcotte, J., & Jongenelis, S. (2004). Involvement of the hippocampus in implicit learning of supra-span sequences: The case of sj. *Cognitive Neuropsychology*, *21*(8), 867–882. https://doi.org/10.1080/02643290342000609

Gaillard, V., Destrebecqz, A., Michiels, S., & Cleeremans, A. (2009). Effects of age and practice in sequence learning: A graded account of ageing, learning, and control. *European Journal of Cognitive Psychology*, *21*(2–3), 255–282. https://doi.org/10.1080/09541440802257423

Gais, S., & Born, J. (2004). Declarative memory consolidation: Mechanisms acting during human sleep. *Learning & Memory*, *11*(6), 679–685. https://doi.org/10.1101/lm.80504

Galea, J. M., Albert, N. B., Ditye, T., & Miall, R. C. (2010). Disruption of the Dorsolateral Prefrontal Cortex Facilitates the Consolidation of Procedural Skills. *Journal of Cognitive Neuroscience*, *22*(6), 1158–1164. https://doi.org/10.1162/jocn.2009.21259

Gavril, L., Roșan, A., & Szamosközi, Ștefan. (2021). The role of visual-spatial attention in reading development: A meta-analysis. *Cognitive Neuropsychology*, *38*(6), 387–407. https://doi.org/10.1080/02643294.2022.2043839

Gingras, M., & Sénéchal, M. (2019). Evidence of Statistical Learning of Orthographic Representations in Grades 1–5: The Case of Silent Letters and Double Consonants in French. *Scientific Studies of Reading*, *23*(1), 37–48. https://doi.org/10.1080/10888438.2018.1482303

Godwin, K. E., Almeda, Ma. V., Seltman, H., Kai, S., Skerbetz, M. D., Baker, R. S., & Fisher, A. V. (2016). Off-task behavior in elementary school children. *Learning and Instruction*, *44*, 128–143. https://doi.org/10.1016/j.learninstruc.2016.04.003

Gomes, H., Molholm, S., Christodoulou, C., Ritter, W., & Cowan, N. (2000). The development of auditory attention in children. *Frontiers in Bioscience : A Journal and Virtual Library*, *5*, D108-120. https://doi.org/10.2741/gomes

Gooch, D., Snowling, M. J., & Hulme, C. (2012). Reaction Time Variability in Children With ADHD Symptoms and/or Dyslexia. *Developmental Neuropsychology*, *37*(5), 453–472. https://doi.org/10.1080/87565641.2011.650809

Gottlieb, J. (2012). Attention, Learning, and the Value of Information. *Neuron*, *76*(2), 281–295. https://doi.org/10.1016/j.neuron.2012.09.034

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. M. (2020). Theoretically Informed Generative Models Can Advance the Psychological and Brain Sciences: Lessons from the Reliability Paradox. *PsyArXiv*. https://doi.org/10.31234/osf.io/xr7y3

Hamrick, P., Lum, J. A. G., & 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 of the United States of America*, *115*(7), 1487–1492. https://doi.org/10.1073/pnas.1713975115

Hardwick, R. M., Rottschy, C., Miall, R. C., & Eickhoff, S. B. (2013). A quantitative meta-analysis and review of motor learning in the human brain. *NeuroImage*, *67*, 283–297. https://doi.org/10.1016/j.neuroimage.2012.11.020

Hauptmann, B., Reinhart, E., Brandt, S. A., & Karni, A. (2005). The predictive value of the leveling off of within-session performance for procedural memory consolidation. *Cognitive Brain Research*, *24*(2), 181–189. https://doi.org/10.1016/j.cogbrainres.2005.01.012

Hay, J. F., Moscovitch, M., & Levine, B. (2002). Dissociating habit and recollection: Evidence from

Parkinson’s disease, amnesia and focal lesion patients. *Neuropsychologia*, *40*(8), 1324–1334. https://doi.org/10.1016/S0028-3932(01)00214-7

Hazeltine, E. (1997). Attention and stimulus characteristics determine the locus of motor- sequence encoding. A PET study. *Brain*, *120*(1), 123–140. https://doi.org/10.1093/brain/120.1.123

* Hedenius, M., Lum, J. A. G., & Bölte, S. (2021). Alterations of procedural memory consolidation in children with developmental dyslexia. *Neuropsychology*, *35*(2), 185–196. https://doi.org/10.1037/neu0000708

* Hedenius, M., Persson, J., Alm, P. A., Ullman, M. T., Howard, 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. https://doi.org/10.1016/j.ridd.2013.08.014

Hedenius, M., Persson, J., Tremblay, A., Adi-Japha, E., Veríssimo, J., Dye, C. D., Alm, P., Jennische, M., Bruce Tomblin, J., & Ullman, M. T. (2011). Grammar predicts procedural learning and consolidation deficits in children with Specific Language Impairment. *Research in Developmental Disabilities*, *32*(6), 2362–2375. https://doi.org/10.1016/j.ridd.2011.07.026

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

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

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

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

Henríquez-Henríquez, M. P., Billeke, P., Henríquez, H., Zamorano, F. J., Rothhammer, F., & Aboitiz, F. (2015). Intra-individual response variability assessed by ex-Gaussian analysis may be a new endophenotype for attention-deficit/hyperactivity disorder. *Frontiers in Psychiatry*, *6*(JAN), 1–8. https://doi.org/10.3389/fpsyt.2014.00197

Herbert, M. R., Ziegler, D. A., Makris, N., Bakardjiev, A., Hodgson, J., Adrien, K. T., Kennedy, D. N., Filipek, P. A., & Caviness, V. S. (2003). Larger brain and white matter volumes in children with developmental language disorder. *Developmental Science*, *6*(4), F11–F22. https://doi.org/10.1111/1467-7687.00291

Heuer, H., & Schmidtke, V. (1996). Secondary-task effects on sequence learning. *Psychological Research*, *59*(2), 119–133. https://doi.org/10.1007/BF01792433

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

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. https://doi.org/10.1080/13682820010019874

Hodel, A. S., Markant, J. C., Van Den Heuvel, S. E., Cirilli-Raether, J. M., & Thomas, K. M. (2014). Developmental differences in effects of task pacing on implicit sequence learning. *Frontiers in Psychology*, *5*(FEB), 1–10. https://doi.org/10.3389/fpsyg.2014.00153

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Holz, J., Piosczyk, H., Landmann, N., Feige, B., Spiegelhalder, K., Riemann, D., Nissen, C., & Voderholzer, U. (2012). The Timing of Learning before Night-Time Sleep Differentially Affects Declarative and Procedural Long-Term Memory Consolidation in Adolescents. *PLoS ONE*, *7*(7), e40963. https://doi.org/10.1371/journal.pone.0040963

Howard, D. V., & Howard, J. H. (1989). Age Differences in Learning Serial Patterns: Direct Versus Indirect Measures. *Psychology and Aging*, *4*(3), 357–364. https://doi.org/10.1037/0882-7974.4.3.357

Howard, D. Y., & Howard, J. H. (1992). Adult Age Differences in the Rate of Learning Serial Patterns: Evidence From Direct and Indirect Tests. *Psychology and Aging*, *7*(2), 232–241. https://doi.org/10.1037//0882-7974.7.2.232

Howard, 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. https://doi.org/10.1016/j.neuropsychologia.2005.10.015

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

Huang, J., Li, Y., Zhang, J., Wang, X., Huang, C., Chen, A., & Liu, D. (2017). FMRI Investigation on Gradual Change of Awareness States in Implicit Sequence Learning. *Scientific Reports*, *7*(1), 16731. https://doi.org/10.1038/s41598-017-16340-2

Hulme, C., Bowyer-Crane, C., Carroll, J. M., Duff, F. J., & Snowling, M. J. (2012). The Causal Role of

Phoneme Awareness and Letter-Sound Knowledge in Learning to Read: Combining Intervention Studies With Mediation Analyses. *Psychological Science*, *23*(6), 572–577. https://doi.org/10.1177/0956797611435921

Jackson, E., Leitão, S., Claessen, M., & Boyes, M. (2020). Working, declarative, and procedural memory in children with developmental language disorder. *Journal of Speech, Language, and Hearing Research*, *63*(12), 4162–4178. https://doi.org/10.1044/2020_JSLHR-20-00135

Janacsek, K., & Nemeth, D. (2013). Implicit sequence learning and working memory: Correlated or complicated? *Cortex*, *49*(8), 2001–2006. https://doi.org/10.1016/j.cortex.2013.02.012

Janacsek, K., Shattuck, K. F., Tagarelli, K. M., Lum, J. A. G., Turkeltaub, P. E., & Ullman, M. T. (2020). Sequence learning in the human brain: A functional neuroanatomical meta-analysis of serial reaction time studies. *NeuroImage*, *207*. https://doi.org/10.1016/j.neuroimage.2019.116387

Jeffreys, H. (1961). *Theory of Probability* (3rd Edition). Clarendon Press.

Jernigan, T. L., Hesselink, J. R., Sowell, E., & Tallal, P. A. (1991). Cerebral Structure on Magnetic Resonance Imaging in Language- and Learning-Impaired Children. *Archives of Neurology*, *48*(5), 539–545. https://doi.org/10.1001/archneur.1991.00530170103028

Jimenez, L., & Mendez, C. (1999). Which Attention Is Needed for Implicit Sequence Learning?? *Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 236–259. https://doi.org/10.1037/0278-7393.25.1.236

Jiménez, L., & Vázquez, G. A. (2005). Sequence learning under dual-task conditions: Alternatives to a resource-based account. *Psychological Research Psychologische Forschung*, *69*(5–6), 352–368. https://doi.org/10.1007/s00426-004-0210-9

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. https://doi.org/10.1007/s11881-010-0048-3

Juhasz, D., Nemeth, D., & Janacsek, K. (2019). Is there more room to improve? The lifespan trajectory of procedural learning and its relationship to the between? The within-group differences in average response times. *PLoS ONE*, *14*(7), 1–20. https://doi.org/10.1371/journal.pone.0215116

Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall.

Kalm, K., Davis, M. H., & Norris, D. (2013). Individual Sequence Representations in the Medial Temporal Lobe. *Journal of Cognitive Neuroscience*, *25*(7), 1111–1121. https://doi.org/10.1162/jocn_a_00378

† Kalra, P. B., Gabrieli, J. D. E., & Finn, A. S. (2019). Evidence of stable individual differences in

implicit learning. *Cognition*, *190*(July 2018), 199–211. https://doi.org/10.1016/j.cognition.2019.05.007

Karatekin, C., White, T., & Bingham, C. (2009). Incidental and intentional sequence learning in youth-onset psychosis and attention-deficit/hyperactivity disorder (ADHD). *Neuropsychology*, *23*(4), 445–459. https://doi.org/10.1037/a0015562

Karmiloff-Smith, A. (2015). An Alternative to Domain-general or Domain-specific Frameworks for Theorizing about Human Evolution and Ontogenesis. *AIMS Neuroscience*, *2*(2), 91–104. https://doi.org/10.3934/Neuroscience.2015.2.91

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*(3), 321–340. https://doi.org/10.1016/j.cognition.2010.05.011

Keele, S. W., Ivry, R., Mayr, U., Hazeltine, E., & Heuer, H. (2003). The cognitive and neural architecture of sequence representation. *Psychological Review*, *110*(2), 316–339. https://doi.org/10.1037/0033-295X.110.2.316

Kelly, S. W., Griffiths, S., & Frith, U. (2002). Evidence for implicit sequence learning in dyslexia. *Dyslexia*, *8*(1), 43–52. https://doi.org/10.1002/dys.208

Kemény, F. (2014). Self-insight in probabilistic categorization—Not implicit in children either. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00233

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

* Kidd, E., & 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. https://doi.org/10.1080/01690965.2010.493735

Kinder, A., Rolfs, M., & Kliegl, R. (2008). Sequence learning at optimal stimulus-response mapping: Evidence from a serial reaction time task. *Quarterly Journal of Experimental Psychology (2006)*, *61*(2), 203–209. https://doi.org/10.1080/17470210701557555

King, B. R., Hoedlmoser, K., Hirschauer, F., Dolfen, N., & Albouy, G. (2017). Sleeping on the motor engram: The multifaceted nature of sleep-related motor memory consolidation. *Neuroscience and Biobehavioral Reviews*, *80*, 1–22. https://doi.org/10.1016/j.neubiorev.2017.04.026

Kirkham, J. J., Dwan, K. M., Altman, D. G., Gamble, C., Dodd, S., Smyth, R., & Williamson, P. R. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*, *340*(feb15 1), c365–c365. https://doi.org/10.1136/bmj.c365

Kirov, R., Kolev, V., Verleger, R., & Yordanova, J. (2015). Labile sleep promotes awareness of abstract knowledge in a serial reaction time task. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01354

Kiss, M., Nemeth, D., & Janacsek, K. (2019). Stimulus presentation rates affect performance but not the acquired knowledge – Evidence from procedural learning. *BioRxiv*, 650598–650598. https://doi.org/10.1101/650598

Knowlton, B. J., Mangels, J. A., Squire, L. R., Series, N., & Sep, N. (2006). *A Neostriatal Habit Learning System in Humans A Neostriatal Habit Learning System in Humans*. *273*(5280), 1399–1402.

Knowlton, B. J., & Moody, T. D. (2008). Procedural learning in humans. In J. Byrne (Ed.), *Learning and Memory: A Comprehensive Reference, vol 3, Memory Systems* (pp. 321–340). Elsevier Press.

Knowlton, B. J., Ramus, S. J., & Squire, L. R. (1992). Intact Artificial Grammar Learning in Amnesia: Dissociation of Classification Learning and Explicit Memory for Specific Instances. *Psychological Science*, *3*(3), 172–179. https://doi.org/10.1111/j.1467-9280.1992.tb00021.x

Knowlton, B. J., & Squire, L. R. (1996). Artificial Grammar Learning Depends on Implicit Acquisition of Both Abstract and Exemplar-Specific Information. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*(1), 169–181.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory*, *1*(2), 106–120. https://doi.org/10.1101/lm.1.2.106

Koch, F.-S., Sundqvist, A., Thornberg, U. B., Nyberg, S., Lum, J. A. G., Ullman, M. T., Barr, R., Rudner, M., & Heimann, M. (2020). Procedural memory in infancy: Evidence from implicit sequence learning in an eye-tracking paradigm. *Journal of Experimental Child Psychology*, *191*, 104733. https://doi.org/10.1016/j.jecp.2019.104733

Konstantinidis, S. (2005). Computing the Levenshtein distance of a regular language. *IEEE Information Theory Workshop, 2005.*, 4 pp. https://doi.org/10.1109/ITW.2005.1531868

Kottner, J., & Streiner, D. L. (2011). The difference between reliability and agreement. *Journal of Clinical Epidemiology*, *64*(6), 701–702. https://doi.org/10.1016/j.jclinepi.2010.12.001

Krishnan, S., Asaridou, S. S., Cler, G. J., Smith, H. J., Willis, H. E., Healy, M. P., Thompson, P. A., Bishop, D. V. M., & Watkins, K. E. (2021). Functional organisation for verb generation in children with developmental language disorder. *NeuroImage*, *226*, 117599. https://doi.org/10.1016/j.neuroimage.2020.117599

Krok, W. C., & Leonard, L. B. (2015). Past Tense Production in Children With and Without Specific Language Impairment Across Germanic Languages: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *58*(4), 1326–1340. https://doi.org/10.1044/2015_JSLHR-L-

14-0348

* Kuppuraj, S., Rao, P., & Bishop, D. V. (2016). Declarative capacity does not trade-off with procedural capacity in children with specific language impairment. *Autism & Developmental Language Impairments*, *1*, 239694151667441–239694151667441. https://doi.org/10.1016/j.oooo.2016.01.005

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

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. https://doi.org/10.1007/s11881-013-0083-y

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

†* Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2020). 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. *Language Learning and Development*, *16*(4), 426–450. https://doi.org/10.1080/15475441.2020.1820340

Lauer, T., Cornelissen, T. H. W., Draschkow, D., Willenbockel, V., & Võ, M. L.-H. (2018). The role of scene summary statistics in object recognition. *Scientific Reports*, *8*(1), 14666. https://doi.org/10.1038/s41598-018-32991-1

Law, J., Garrett, Z., & Nye, C. (2003). Speech and language therapy interventions for children with primary speech and language delay or disorder. *Cochrane Database of Systematic Reviews*, *2015*(5). https://doi.org/10.1002/14651858.CD004110

LeBel, E. P., & Paunonen, S. V. (2011). Sexy But Often Unreliable: The Impact of Unreliability on the Replicability of Experimental Findings With Implicit Measures. *Personality and Social Psychology Bulletin*, *37*(4), 570–583. https://doi.org/10.1177/0146167211400619

Leber, A. B., & Egeth, H. E. (2006). It's under control: Top-down search strategies can override attentional capture. *Psychonomic Bulletin & Review*, *13*(1), 132–138. https://doi.org/10.3758/BF03193824

Lee, J. C., Mueller, K. L., & Tomblin, J. B. (2016). Examining procedural learning and corticostriatal pathways for individual differences in language: Testing endophenotypes of DRD2/ANKK1. *Language, Cognition and Neuroscience*, *31*(9), 1098–1114. https://doi.org/10.1080/23273798.2015.1089359

Lee, J. C., Nopoulos, P. C., & Bruce Tomblin, J. (2013). Abnormal subcortical components of the corticostriatal system in young adults with DLI: A combined structural MRI and DTI study. *Neuropsychologia*, *51*(11), 2154–2161. https://doi.org/10.1016/j.neuropsychologia.2013.07.011

* Lee, J. C., & Tomblin, J. B. (2015). Procedural Learning and Individual Differences in Language. *Language Learning and Development*, *11*(3), 215–236. https://doi.org/10.1080/15475441.2014.904168

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*, 707–710.

Levy, F. (1980). The development of sustained attention (vigilance) and inhibition in children: Some normative data. *Journal of Child Psychology and Psychiatry*, *21*(1), 77–84. https://doi.org/10.1111/j.1469-7610.1980.tb00018.x

Lieberman, M. D., Chang, G. Y., Chiao, J., Bookheimer, S. Y., & Knowlton, B. J. (2004). An event-related fMRI study of artificial grammar learning in a balanced chunk strength design. *Journal of Cognitive Neuroscience*, *16*(3), 427–438. https://doi.org/10.1162/089892904322926764

Liégeois, F., Mayes, A., & Morgan, A. (2014). Neural Correlates of Developmental Speech and Language Disorders: Evidence from Neuroimaging. *Current Developmental Disorders Reports*, *1*(3), 215–227. https://doi.org/10.1007/s40474-014-0019-1

Lin, H.-Y., Hwang-Gu, S.-L., & Gau, S. S.-F. (2015). Intra-individual reaction time variability based on ex-Gaussian distribution as a potential endophenotype for attention-deficit/hyperactivity disorder. *Acta Psychiatrica Scandinavica*, *132*(1), 39–50. https://doi.org/10.1111/acps.12393

* Llompart, M., & Dąbrowska, E. (2020). Explicit but Not Implicit Memory Predicts Ultimate Attainment in the Native Language. *Frontiers in Psychology*, *11*(September), 1–14. https://doi.org/10.3389/fpsyg.2020.569586

Lohvansuu, K., Torppa, M., Ahonen, T., Eklund, K., Hämäläinen, J. A., Leppänen, P. H. T., & Lyytinen, H. (2021). Unveiling the Mysteries of Dyslexia—Lessons Learned from the Prospective Jyväskylä Longitudinal Study of Dyslexia. *Brain Sciences*, *11*(4), 427. https://doi.org/10.3390/brainsci11040427

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

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

Lum, J. A. G., & Bleses, D. (2012). Declarative and procedural memory in Danish speaking children with specific language impairment. *Journal of Communication Disorders*, *45*(1), 46–58. https://doi.org/10.1016/j.jcomdis.2011.09.001

Lum, J. A. G., & Conti-Ramsden, G. (2013). Long-Term Memory: A Review and Meta-Analysis of Studies of Declarative and Procedural Memory in Specific Language Impairment. *Topics in Language Disorders*, *33*(4), 282–297. https://doi.org/10.1097/01.TLD.0000437939.01237.6a

Lum, J. A. G., 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*. https://doi.org/10.1016/j.cortex.2013.10.011

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

Lum, J. A. G., Gelgic, C., & Conti-Ramsden, G. (2010). Procedural and declarative memory in children with and without specific language impairment. *International Journal of Language & Communication Disorders*, *45*(1), 96–107. https://doi.org/10.3109/13682820902752285

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

Lum, J. A. G., Lammertink, I., Clark, G. M., Fuelscher, I., Hyde, C., Enticott, P. G., & Ullman, M. T. (2019). Visuospatial sequence learning on the serial reaction time task modulates the P1 event-related potential: XXXX. *Psychophysiology*, *56*(2), e13292. https://doi.org/10.1111/psyp.13292

Lum, J. A. G., 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. https://doi.org/10.1016/j.ridd.2013.07.017

Maisog, J. M., Einbinder, E. R., Flowers, D. L., Turkeltaub, P. E., & Eden, G. F. (2008). A Meta-analysis of Functional Neuroimaging Studies of Dyslexia. *Annals of the New York Academy of Sciences*, *1145*(1), 237–259. https://doi.org/10.1196/annals.1416.024

Mandler, J. M., & Robinson, C. A. (1978). Developmental changes in picture recognition. *Journal of*

Experimental Child Psychology, *26*(1), 122–136. https://doi.org/10.1016/0022-0965(78)90114-5

Marini, A., Piccolo, B., Taverna, L., Berginc, M., & Ozbič, M. (2020). The Complex Relation between Executive Functions and Language in Preschoolers with Developmental Language Disorders. *International Journal of Environmental Research and Public Health*, *17*(5), 1772. https://doi.org/10.3390/ijerph17051772

Martin, A., Kronbichler, M., & Richlan, F. (2016). Dyslexic brain activation abnormalities in deep and shallow orthographies: A meta-analysis of 28 functional neuroimaging studies. *Human Brain Mapping*, *37*(7), 2676–2699. https://doi.org/10.1002/hbm.23202

Martini, M., Furtner, M. R., & Sachse, P. (2013). Working Memory and Its Relation to Deterministic Sequence Learning. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0056166

Massidda, D. (2013). *retimes: Reaction Time Analysis*. https://CRAN.R-project.org/package=retimes

Mathews, R. C. (1997). Is research painting a biased picture of implicit learning? The dangers of methodological purity in scientific debate. *Psychonomic Bulletin & Review*, *4*(1), 38–42. https://doi.org/10.3758/BF03210771

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. (2017). Bayesian Inference for Correlations in the Presence of Measurement Error and Estimation Uncertainty. *Collabra: Psychology*, *3*(1), 25. https://doi.org/10.1525/collabra.78

Maughan, B., Rutter, M., & Yule, W. (2020). The Isle of Wight studies: The scope and scale of reading difficulties. *Oxford Review of Education*, *46*(4), 429–438. https://doi.org/10.1080/03054985.2020.1770064

May, K., & Hittner, J. B. (2003). On the Relation between Power and Reliability of Difference Scores. *Percept Mot Skills*, *97*(3 Pt 1), 905–908. https://doi.org/10.2466/pms.2003.97.3.905

* Mayor-Dubois, C., Zesiger, P., Van der Linden, M., & Roulet-Perez, E. (2014). Nondeclarative learning in children with Specific Language Impairment: Predicting regularities in the visuomotor, phonological, and cognitive domains. *Child Neuropsychology*, *20*(1), 14–22. https://doi.org/10.1080/09297049.2012.734293

McAvinue, L. P., Habekost, T., Johnson, K. A., Kyllingsbæk, S., Vangkilde, S., Bundesen, C., & Robertson, I. H. (2012). Sustained attention, attentional selectivity, and attentional capacity across the lifespan. *Attention, Perception, & Psychophysics*, *74*(8), 1570–1582. https://doi.org/10.3758/s13414-012-0352-6

McCaffrey, R. J., Duff, K., & Westervelt, H. J. (2000). *Practitioner's guide to evaluating change with neuropsychological assessment instruments.* Kluwer Academic, Plenum Publishers.

McClelland, J. L., & O'Reilly, R. C. (1995). Why There Are Complementary Learning Systems in the Hippocampus and Neocortex:InsightsFrom the Successesand Failuresof Connectionist Models of Learning and Memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295x.102.3.419

McGregor, K. K. (2020). How We Fail Children With Developmental Language Disorder. *Language, Speech, and Hearing Services in Schools*, *51*(4), 981–992. https://doi.org/10.1044/2020_LSHSS-20-00003

Medimorec, S., Milin, P., & Divjak, D. (2021a). Working memory affects anticipatory behavior during implicit pattern learning. *Psychological Research*, *85*(1), 291–301. https://doi.org/10.1007/s00426-019-01251-w

Medimorec, S., Milin, P., & Divjak, D. (2021b). Inhibition of Eye Movements Disrupts Spatial Sequence Learning. *Experimental Psychology*, *68*(4), 221–228. https://doi.org/10.1027/1618-3169/a000528

Melby-Lervåg, M., Lyster, S.-A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin*, *138*(2), 322–352. https://doi.org/10.1037/a0026744

* 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. https://doi.org/10.1016/j.neuropsychologia.2009.11.003

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. https://doi.org/10.1016/j.neuroimage.2006.08.024

Menghini, D., Hagberg, G. E., Petrosini, L., Bozzali, M., Macaluso, E., Caltagirone, C., & Vicari, S. (2008). Structural Correlates of Implicit Learning Deficits in Subjects with Developmental Dyslexia. *Annals of the New York Academy of Sciences*, *1145*(1), 212–221. https://doi.org/10.1196/annals.1416.010

Milin, P., Divjak, D., & Baayen, R. H. (2017). A learning perspective on individual differences in skilled reading: Exploring and exploiting orthographic and semantic discrimination cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1730–1751. https://doi.org/10.1037/xlm0000410

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

https://doi.org/10.1080/20445911.2016.1223082

Mirzakhany - Araghi, N., Yasaei, R., Khoshalipanah, M., Nejati, V., Pashazadeh - Azari, Z., & Tabatabaee, S. M. (2013). Motor Learning in children with ADHD and Normal Childre: Comparison of Implicit and Explicit Motor Sequence. *Motor Behavior*, *2*(Shahid Beheshti University), 12–23. https://doi.org/10.22037/jcpr.2017.04

Monroy, C., Meyer, M., Gerson, S., & Hunnius, S. (2017). Statistical learning in social action contexts. *PLOS ONE*, *12*(5), e0177261. https://doi.org/10.1371/journal.pone.0177261

Mugnaini, D., Lassi, S., La Malfa, G., & Albertini, G. (2009). Internalizing correlates of dyslexia. *World Journal of Pediatrics*, *5*(4), 255–264. https://doi.org/10.1007/s12519-009-0049-7

Naismith, S. L., Lagopoulos, J., Ward, P. B., Davey, C. G., Little, C., & Hickie, I. B. (2010). Fronto-striatal correlates of impaired implicit sequence learning in major depression: An fMRI study. *Journal of Affective Disorders*, *125*(1–3), 256–261. https://doi.org/10.1016/j.jad.2010.02.114

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383. https://doi.org/10.1016/0010-0285(77)90012-3

Nemeth, D., Janacsek, K., & Fiser, J. (2013). *Age-dependent and coordinated shift in performance between implicit and explicit skill learning*. *7*(October), 1–13. https://doi.org/10.3389/fncom.2013.00147

Nemeth, D., Janacsek, K., Londe, Z., Ullman, M. T., Howard, D. V., & Howard, J. H. (2010). Sleep has no critical role in implicit motor sequence learning in young and old adults. *Experimental Brain Research*, *201*(2), 351–358. https://doi.org/10.1007/s00221-009-2024-x

Nemeth, D., Janacsek, K., Polner, B., & Kovacs, Z. A. (2013). Boosting Human Learning by Hypnosis. *Cerebral Cortex*, *23*(4), 801–805. https://doi.org/10.1093/cercor/bhs068

Nicolson, R. I., & Fawcett, A. J. (1994). Comparison of deficits in cognitive and motor skills among children with dyslexia. *Annals of Dyslexia*, *44*(1), 147–164. https://doi.org/10.1007/BF02648159

Nieuwenhuis, R., Te Grotenhuis, M., & Pelzer, B. (2012). Influence.ME: Tools for Detecting Influential Data in Mixed Effects Models. *R Journal*, *4*(2), 38–47.

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

Norman, E., & Price, M. (2012). Social intuition as a form of implicit learning: Sequences of body movements are learned less explicitly than letter sequences. *Advances in Cognitive Psychology*, *8*(2), 121–131. https://doi.org/10.5709/acp-0109-x

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2

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

Obeid, R., Brooks, P. J., Powers, K. L., Gillespie-Lynch, K., & Lum, J. A. G. (2016). Statistical learning in specific language impairment and autism spectrum disorder: A meta-analysis. *Frontiers in Psychology*, *7*(AUG), 1–18. https://doi.org/10.3389/fpsyg.2016.01245

Ofen, N. (2012). The development of neural correlates for memory formation. *Neuroscience & Biobehavioral Reviews*, *36*(7), 1708–1717. https://doi.org/10.1016/j.neubiorev.2012.02.016

Ofen, N., Kao, Y. C., Sokol-Hessner, P., Kim, H., Whitfield-Gabrieli, S., & Gabrieli, J. D. E. (2007). Development of the declarative memory system in the human brain. *Nature Neuroscience*, *10*(9), 1198–1205. https://doi.org/10.1038/nn1950

Oliveira, C. M., Hayiou-Thomas, M. E., & Henderson, L. M. (in prep). *Limited evidence of an association between language, literacy and procedural learning in typical and atypical development: A meta-analysis*.

†* Oliveira, C. M., Hayiou-Thomas, M. E., & Henderson, L. M. (submitted). Reliability of the Serial Reaction Time task: If at first you don’t succeed, try try try again. *PsyArxiv*. https://doi.org/10.31234/osf.io/hqmy7

† Oliveira, C. M., Henderson, L., & Hayiou-Thomas, M. E. (in prep). *Procedural learning in the serial reaction time task in individuals with and without dyslexia: Group-level and individual differences*.

Packard, M. G., & Knowlton, B. J. (2002). Learning and Memory Functions of the Basal Ganglia. *Annual Review of Neuroscience*, *25*(1), 563–593. https://doi.org/10.1146/annurev.neuro.25.112701.142937

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. https://doi.org/10.1037/0096-3445.130.3.401

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ*, *372*. https://doi.org/10.1136/bmj.n71

Palmer, C. E., Langbehn, D., Tabrizi, S. J., & Papoutsi, M. (2018). Test-retest reliability of measures commonly used to measure striatal dysfunction across multiple testing sessions: A longitudinal study. *Frontiers in Psychology*, *8*(JAN), 1–13. https://doi.org/10.3389/fpsyg.2017.02363

Pan, S. C., & Rickard, T. C. (2015). Sleep and motor learning: Is there room for consolidation? *Psychological Bulletin*, *141*(4), 812–834. https://doi.org/10.1037/bul0000009

Parent, A., & Hazrati, L.-N. (1995). The cortico-basal ganglia-thalamo-corticalloop. *Abstract Brain Research Reviews*, *20*, 91–127.

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

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

Pascual-Leone, A., Wassermann, E. M., Grafman, J., & Hallett, M. (1996). The role of the dorsolateral prefrontal cortex in implicit procedural learning. *Experimental Brain Research*, *107*(3), 479–485. https://doi.org/10.1007/BF00230427

Patterson, T. K., & Knowlton, B. J. (2018). Subregional specificity in human striatal habit learning: A meta-analytic review of the fMRI literature. *Current Opinion in Behavioral Sciences*, *20*, 75–82. https://doi.org/10.1016/j.cobeha.2017.10.005

Paulesu, E., Danelli, L., & Berlingeri, M. (2014). Reading the dyslexic brain: Multiple dysfunctional routes revealed by a new meta-analysis of PET and fMRI activation studies. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00830

Pedersen, A., & Ohrmann, P. (2018). Impaired Behavioral Inhibition in Implicit Sequence Learning in Adult ADHD. *Journal of Attention Disorders*, *22*(3), 250–260. https://doi.org/10.1177/1087054712464392

Peigneux, P., Orban, P., Balteau, E., Degueldre, C., Luxen, A., Laureys, S., & Maquet, P. (2006). Offline Persistence of Memory-Related Cerebral Activity during Active Wakefulness. *PLoS Biology*, *4*(4), 12. https://doi.org/10.1371/journal.pbio.0040100

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pennington, B. F. (2006). From single to multiple deficit models of developmental disorders.

Cognition, *101*(2), 385–413. https://doi.org/10.1016/j.cognition.2006.04.008

Perlant, A. S., & Largy, P. (2011). Are implicit learning abilities sensitive to the type of material to be processed? Study on typical readers and children with dyslexia: ARE IMPLICIT LEARNING ABILITIES SENSITIVE? *Journal of Research in Reading*, *34*(3), 298–314. https://doi.org/10.1111/j.1467-9817.2010.01464.x

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, *61*(10), 991–996. https://doi.org/10.1016/j.jclinepi.2007.11.010

Petersson, K.-M., Folia, V., & Hagoort, P. (2012). What artificial grammar learning reveals about the neurobiology of syntax. *Brain and Language*, *120*(2), 83–95. https://doi.org/10.1016/j.bandl.2010.08.003

Pinker, S. (1994). *The Language Instinct: The New Science of Language and Mind*. The Penguin Press.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*(11), 456–463. https://doi.org/10.1016/S1364-6613(02)01990-3

Plante, E., & Gómez, R. L. (2018). Learning Without Trying: The Clinical Relevance of Statistical Learning. *Language, Speech, and Hearing Services in Schools*, *49*(3S). https://doi.org/10.1044/2018_lshss-stlt1-17-0131

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, *38*(1), 43–102. https://doi.org/10.1016/0010-0277(91)90022-V

Polanczyk, G. V., Willcutt, E. G., Salum, G. A., Kieling, C., & Rohde, L. A. (2014). ADHD prevalence estimates across three decades: An updated systematic review and meta-regression analysis. *International Journal of Epidemiology*, *43*(2), 434–442. https://doi.org/10.1093/ije/dyt261

Poldrack, R. A., & Packard, M. G. (2003). *Competition among multiple memory systems: Converging evidence from animal and human brain studies*. *41*, 245–251.

Poldrack, R. A., Prabhakaran, V., Seger, C. A., & Gabriel, J. D. E. (1999). Striatal Activation During Acquisition of a Cognitive Skill. *Neuropsychology*, *13*(4), 564–574. https://doi.org/10.1037/0894-4105.13.4.564

Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the Detection of Signals. *J Exp Psychol.*, *109*(2), 160–174. https://doi.org/10.1037/0096-3445.109.2.160

Press, D. Z., Casement, M. D., Pascual-Leone, A., & Robertson, E. M. (2005). The time course of offline motor sequence learning. *Cognitive Brain Research*, *25*(1), 375–378. https://doi.org/10.1016/j.cogbrainres.2005.05.010

Purdy, J. D., Leonard, L. B., Weber-Fox, C., & Kaganovich, N. (2014). Decreased sensitivity to long-distance dependencies in children with a history of specific language impairment: Electrophysiological evidence. *Journal of Speech, Language, and Hearing Research : JSLHR*, *57*(3), 1040–1059. https://doi.org/10.1044/2014_JSLHR-L-13-0176

Pustejovsky, J. (2021). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections*. https://CRAN.R-project.org/package=clubSandwich

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Prevention Science*, *23*(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

Rah, S. K., Reber, A. S., & Hsiao, A. T. (2000). Another wrinkle on the dual-task SRT experiment: It's probably not dual task. *Psychonomic Bulletin & Review*, *7*(2), 309–313. https://doi.org/10.3758/BF03212986

Ramus, F. (2003). Theories of developmental dyslexia: Insights from a multiple case study of dyslexic adults. *Brain*, *126*(4), 841–865. https://doi.org/10.1093/brain/awg076

Ramus, F., & Ahissar, M. (2012). Developmental dyslexia: The difficulties of interpreting poor performance, and the importance of normal performance. *Cognitive Neuropsychology*, *29*(1–2), 104–122. https://doi.org/10.1080/02643294.2012.677420

Raschle, N. M., Chang, M., & Gaab, N. (2011). Structural brain alterations associated with dyslexia predate reading onset. *NeuroImage*, *57*(3), 742–749. https://doi.org/10.1016/j.neuroimage.2010.09.055

Raschle, N. M., Zuk, J., & Gaab, N. (2012). Functional characteristics of developmental dyslexia in left-hemispheric posterior brain regions predate reading onset. *Proceedings of the National Academy of Sciences*, *109*(6), 2156–2161. https://doi.org/10.1073/pnas.1107721109

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855–863. https://doi.org/10.1016/S0022-5371(67)80149-X

Reber, A. S. (1989). Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious. *Journal of Experimental Psychology: General*, *118*(3), 219–235. https://doi.org/10.1037/0096-3445.118.3.219

Reber, A. S., Walkenfeld, F. F., & Hernstadt, R. (1991). Implicit and Explicit Learning: Individual Differences and IQ. *Ournal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 888–896. https://psycnet.apa.org/doi/10.1037/0278-7393.17.5.888

Reber, A. S. (1993). Implicit learning and tacit knowledge: An essay on the cognitive unconscious. Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780195106589.001.0001

Reber, P. J. (2008). Cognitive Neuroscience of Declarative and Nondeclarative Memory. In M. Guadagnoli, A. S. Benjamin, J. S. De Belle, B. Etnyre, & T. A. Polk (Eds.), *Human Learning: Biology, Brain, and Neuroscience* (Eds., Vol. 139, pp. 113–123). Elsevier. DOI: 10.1016/S0166-4115(08)10010-3

Reber, P. J., & Squire, L. R. (1994). Parallel brain systems for learning with and without awareness. *Learning & Memory*, *1*(4), 217–229. https://doi.org/10.1101/lm.1.4.217

Reber, P. J., & Squire, L. R. (1998). Encapsulation of Implicit and Explicit Memory in Sequence Learning. *Journal of Cognitive Neuroscience*, *10*(2), 248–263. https://doi.org/10.1162/089892998562681

Reber, P. J., & Squire, L. R. (1999). Intact Learning of Artificial Grammars and Intact Category Learning by Patients With Parkinson’s Disease. *Behavioral Neuroscience*, *113*(2), 235–242. https://doi.org/10.1037/0735-7044.113.2.235

Reda, F., Gorgoni, M., D’Atri, A., Scarpelli, S., Carpi, M., Di Cola, E., Menghini, D., Vicari, S., Stella, G., & De Gennaro, L. (2021). Sleep-Related Declarative Memory Consolidation in Children and Adolescents with Developmental Dyslexia. *Brain Sciences*, *11*(1), 73. https://doi.org/10.3390/brainsci11010073

Reed, J., & Johnson, P. (1994). Assessing Implicit Learning With Indirect Tests: Determining What Is Learned About Sequence Structure. *Journal of Experimental Psychology: Learning Memory and Cognition*, *20*(3), 585–594. https://doi.org/10.1037/0278-7393.20.3.585

Reifman, J., Kumar, K., Khitrov, M. Y., Liu, J., & Ramakrishnan, S. (2018). PC-PVT 2.0: An updated platform for psychomotor vigilance task testing, analysis, prediction, and visualization. *Journal of Neuroscience Methods*, *304*, 39–45. https://doi.org/10.1016/j.jneumeth.2018.04.007

Richlan, F., Kronbichler, M., & Wimmer, H. (2009). Functional abnormalities in the dyslexic brain: A quantitative meta-analysis of neuroimaging studies. *Human Brain Mapping*, *30*(10), 3299–3308. https://doi.org/10.1002/hbm.20752

Richlan, F., Kronbichler, M., & Wimmer, H. (2013). Structural abnormalities in the dyslexic brain: A meta-analysis of voxel-based morphometry studies: Meta-Analysis Developmental Dyslexia. *Human Brain Mapping*, *34*(11), 3055–3065. https://doi.org/10.1002/hbm.22127

Robertson, E. M., Pascual-Leone, A., & Miall, R. C. (2004). Current concepts in procedural consolidation. *Nature Reviews Neuroscience*, *5*(7), 576–582.

https://doi.org/10.1038/nrn1426

Robertson, E. M., Pascual-Leone, A., & Press, D. Z. (2004). Awareness Modifies the Skill-Learning Benefits of Sleep. *Current Biology*, *14*(3), 208–212. https://doi.org/10.1016/j.cub.2004.01.027

Robertson, E. M., Tormos, J. M., & Maeda, F. (2001). *The Role of the Dorsolateral Prefrontal Cortex during Sequence Learning is Specific for Spatial Information*. 628–635.

Romani, C., Tsouknida, E., di Betta, A. M., & Olson, A. (2011). Reduced attentional capacity, but normal processing speed and shifting of attention in developmental dyslexia: Evidence from a serial task. *Cortex*, *47*(6), 715–733. https://doi.org/10.1016/j.cortex.2010.05.008

Röttger, E., Haider, H., Zhao, F., & Gaschler, R. (2019). Implicit sequence learning despite multitasking: The role of across-task predictability. *Psychological Research*, *83*(3), 526–543. https://doi.org/10.1007/s00426-017-0920-4

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Rouder, J. N., & Haaf, J. M. (2021). Are There Reliable Qualitative Individual Differences in Cognition? *Journal of Cognition*, *4*(1), 46. https://doi.org/10.5334/joc.131

Rouder, J. N., Kumar, A., & Haaf, J. M. (2019). *Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail*. 46.

Rstudio Team. (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. http://www.rstudio.com/

Ruffman, T., Taumoepeau, M., & Perkins, C. (2012). Statistical learning as a basis for social understanding in children: Statistical learning and social understanding. *British Journal of Developmental Psychology*, *30*(1), 87–104. https://doi.org/10.1111/j.2044-835X.2011.02045.x

Rumelhart, D. E., & McClelland, J. L. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In *Mechanisms of language aquisition.* (pp. 195–248). Lawrence Erlbaum Associates, Inc.

Rünger, D., & Frensch, P. A. (2008). How incidental sequence learning creates reportable knowledge: The role of unexpected events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1011–1026. https://doi.org/10.1037/a0012942

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.

https://doi.org/10.1080/13803390591001007

Saffran, J. R. (2018). Statistical learning as a window into developmental disabilities. *Journal of Neurodevelopmental Disorders*, *10*(1). https://doi.org/10.1186/s11689-018-9252-y

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(December), 1926–1928.

Sarter, M., Givens, B., & Bruno, J. P. (2001). The cognitive neuroscience of sustained attention: Where top-down meets bottom-up. *Brain Research Reviews*, *35*(2), 146–160. https://doi.org/10.1016/S0165-0173(01)00044-3

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 501–518. https://doi.org/10.1037/0278-7393.13.3.501

Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The Necessity of the Medial Temporal Lobe for Statistical Learning. *Journal of Cognitive Neuroscience*, *26*(8), 1736–1747. https://doi.org/10.1162/jocn_a_00578

Scharfen, J., Jansen, K., & Holling, H. (2018). Retest effects in working memory capacity tests: A meta-analysis. *Psychonomic Bulletin and Review*, *25*(6), 2175–2199. https://doi.org/10.3758/s13423-018-1461-6

Schendan, H. E., Searl, M. M., Melrose, R. J., & Stern, C. E. (2003). *An fMRI Study of the Role of the Medial Temporal Lobe in Implicit and Explicit Sequence Learning*. *37*, 1013–1025.

* 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. https://doi.org/10.1080/10888438.2018.1482304

Schmalz, X., Treccani, B., & Mulatti, C. (2021). Developmental Dyslexia, Reading Acquisition, and Statistical Learning: A Sceptic’s Guide. *Brain Sciences*, *11*(9), 1143. https://doi.org/10.3390/brainsci11091143

Schmidtke, V., & Heuer, H. (1997). Task integration as a factor in secondary-task effects on sequence learning. *Psychological Research*, *60*(1–2), 53–71. https://doi.org/10.1007/BF00419680

Schuerger, J. M., & Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology*, *45*(2), 294–302. https://doi.org/10.1002/1097-4679(198903)45:2<294::AID-JCLP2270450218>3.0.CO;2-N

Schumacher, E. H., & Schwarb, H. (2009). Parallel response selection disrupts sequence learning under dual-task conditions. *Journal of Experimental Psychology: General*, *138*(2), 270–290. https://doi.org/10.1037/a0015378

Schwarb, H., & Schumacher, E. (2012). Generalized lessons about sequence learning from the study of the serial reaction time task. *Advances in Cognitive Psychology*, *8*(2), 165–178. https://doi.org/10.5709/acp-0113-1

Seger, C. A., & Spiering, B. J. (2011). A Critical Review of Habit Learning and the Basal Ganglia. *Frontiers in Systems Neuroscience*, *5*(August), 1–9. https://doi.org/10.3389/fnsys.2011.00066

Sengottuvel, K., & Rao, P. K. S. (2013a). An Adapted Serial Reaction Time Task for Sequence Learning Measurements. *Psychological Studies*, *58*(3), 276–284. https://doi.org/10.1007/s12646-013-0204-z

Sengottuvel, K., & Rao, P. K. S. (2013b). Aspects of grammar sensitive to procedural memory deficits in children with specific language impairment. *Research in Developmental Disabilities*, *34*(10), 3317–3331. https://doi.org/10.1016/j.ridd.2013.06.036

Sengottuvel, K., & Rao, P. K. S. (2014). Sequence learning pattern in children with specific language impairment. *International Journal on Disability and Human Development*, *13*(1), 55–62. https://doi.org/10.1515/ijdhd-2013-0003

Shanks, D., & Channon, S. (2002). Effects of a secondary task on “implicit” sequence learning: Learning or performance? *Psychological Research*, *66*(2), 99–109. https://doi.org/10.1007/s00426-001-0081-2

Shanks, D. R., Rowland, L. A., & Ranger, M. S. (2005). Attentional load and implicit sequence learning. *Psychological Research*, *69*(5–6), 369–382. https://doi.org/10.1007/s00426-004-0211-8

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*(3), 367–395. https://doi.org/10.1017/S0140525X00035032

Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., Mencl, W. E., Fulbright, R. K., Skudlarski, P., Constable, R. T., Marchione, K. E., Fletcher, J. M., Lyon, G. R., & Gore, J. C. (2002). Disruption of Posterior Brain Systems for Reading in Children with Developmental Dyslexia. *Biological Psychiatry*, *52*, 10.

Shaywitz, S. E., Morris, R., & Shaywitz, B. A. (2008). The Education of Dyslexic Children from Childhood to Young Adulthood. *Annual Review of Psychology*, *59*(1), 451–475. https://doi.org/10.1146/annurev.psych.59.103006.093633

Shaywitz, S. E., & Shaywitz, B. A. (2005). Dyslexia (Specific Reading Disability). *Biological Psychiatry*, *57*(11), 1301–1309. https://doi.org/10.1016/j.biopsych.2005.01.043

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711). https://doi.org/10.1098/rstb.2016.0059

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

†* Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001

Silveri, M. C. (2021). Contribution of the Cerebellum and the Basal Ganglia to Language Production: Speech, Word Fluency, and Sentence Construction—Evidence from Pathology. *The Cerebellum*, *20*(2), 282–294. https://doi.org/10.1007/s12311-020-01207-6

Simor, P., Zavecz, Z., Horváth, K., Éltető, N., Török, C., Pesthy, O., Gombos, F., Janacsek, K., & Nemeth, D. (2019). Deconstructing Procedural Memory: Different Learning Trajectories and Consolidation of Sequence and Statistical Learning. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02708

Smith, F. R. H., Gaskell, M. G., Weighall, A. R., Warmington, M., Reid, A. M., & Henderson, L. M. (2018). Consolidation of vocabulary is associated with sleep in typically developing children, but not in children with dyslexia. *Developmental Science*, *21*(5), e12639. https://doi.org/10.1111/desc.12639

Smith, J., Siegert, R. J., McDowall, J., & Abernethy, D. (2001). Preserved implicit learning on both the serial reaction time task and artificial grammar in patients with Parkinson's disease. *Brain and Cognition*, *45*(3), 378–391. https://doi.org/10.1006/brcg.2001.1286

Smolak, E., McGregor, K. K., Arbisi-Kelm, T., & Eden, N. (2020). Sustained Attention in Developmental Language Disorder and Its Relation to Working Memory and Language. *Journal of Speech, Language, and Hearing Research*, *63*(12), 4096–4108. https://doi.org/10.1044/2020_JSLHR-20-00265

Snowling, M. (2000). *Dyslexia* (2nd ed). Blackwell.

Snowling, M., Dawes, P., Nash, H., & Hulme, C. (2012). Validity of a Protocol for Adult Self-Report of Dyslexia and Related Difficulties. *Dyslexia*, *18*(1), 1–15. https://doi.org/10.1002/dys.1432

Snowling, M. J., Hulme, C., & Nation, K. (2020). Defining and understanding dyslexia: Past, present and future. *Oxford Review of Education*, *46*(4), 501–513. https://doi.org/10.1080/03054985.2020.1765756

Soetens, E., Melis, A., & Notebaert, W. (2004). Sequence learning and sequential effects.

Psychological Research, *69*(1–2), 124–137. https://doi.org/10.1007/s00426-003-0163-4

Song, S., Howard, J. H., & Howard, D. V. (2007). Implicit probabilistic sequence learning is independent of explicit awareness. *Learning & Memory*, *14*(3), 167–176. https://doi.org/10.1101/lm.437407

Soriano-Mas, C., Pujol, J., Ortiz, H., Deus, J., López-Sala, A., & Sans, A. (2009). Age-related brain structural alterations in children with specific language impairment. *Human Brain Mapping*, *30*(5), 1626–1636. https://doi.org/10.1002/hbm.20620

Spearman, C. (1904). “General Intelligence,” Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292.

Spencer, M., Kaschak, M. P., Jones, J. L., & Lonigan, C. J. (2015). Statistical learning is related to early literacy-related skills. *Reading and Writing*. https://doi.org/10.1007/s11145-014-9533-0

* Spit, S., & Rispens, J. (2019). On the Relation Between Procedural Learning and Syntactic Proficiency in Gifted Children. Journal of Psycholinguistic Research, 48(2), 417–429. https://doi.org/10.1007/s10936-018-9611-6

Squire, L., & Dede, A. (2015). Conscious and Unconscious Memory Systems. *Cold Spring Harbor Perspectives in Biology*, *7*. https://doi.org/10.1101/cshperspect.a021667

Squire, L. R. (1984). *Nondeclarative Memory: Multiple Brain Systems Supporting Learning*. *4*(3).

Squire, L. R. (1994). Declarative and Nondeclarative Memory: Multiple Brain Systems Supporting Learning and Memory. In D. L. Schacter & E. Tulving (Eds.), *Memory Systems 1994* (pp. 203–231). The MIT Press. https://doi.org/10.7551/mitpress/4545.003.0008

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177. https://doi.org/10.1016/j.nlm.2004.06.005

Squire, L. R., & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: A neurobiological perspective. *Current Opinion in Neurobiology*, *5*(2), 169–177. https://doi.org/10.1016/0959-4388(95)80023-9

Squire, L. R., Genzel, L., Wixted, J. T., & Morris, R. G. (2015). Memory consolidation. *Cold Spring Harbor Perspectives in Biology*, *7*(8), a021766–a021766. PubMed. https://doi.org/10.1101/cshperspect.a021766

Squire, L. R., Stark, C. E. L., & Clark, R. E. (2004). The medial temporal lobe. *Annual Review of Neuroscience*, *27*(1), 279–306. https://doi.org/10.1146/annurev.neuro.27.070203.144130

Squire, L. R., & Wixted, J. T. (2011). The Cognitive Neuroscience of Human Memory Since H.M. *Annual Review of Neuroscience*, *34*(1), 259–288. https://doi.org/10.1146/annurev-neuro-

061010-113720

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13515–13522. https://doi.org/10.1073/pnas.93.24.13515

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

† Stark-Inbar, A., Raza, M., Taylor, J. A., & Ivry, R. B. (2017). Individual differences in implicit motor learning: Task specificity in sensorimotor adaptation and sequence learning. *Journal of Neurophysiology*, *117*(1), 412–428. https://doi.org/10.1152/jn.01141.2015

Stefaniak, N., Willems, S., Adam, S., & Meulemans, T. (2008). *What is the impact of the explicit knowledge of sequence regularities on both deterministic and probabilistic serial reaction time task performance?* 16.

Sterne, J. A. C., & Egger, M. (2005). Regression Methods to Detect Publication and Other Bias in Meta-Analysis. In H. R. Rothstein, A. J. Sutton, & M. Borestein (Eds.), *Publication Bias in Meta-Analysis* (pp. 99–110). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470870168.ch6

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rucker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*(jul22 1), d4002–d4002. https://doi.org/10.1136/bmj.d4002

* Stoodley, C. J., Harrison, E. P. D., & Stein, J. F. (2006). Implicit motor learning deficits in dyslexic adults. *Neuropsychologia*, *44*(5), 795–798. https://doi.org/10.1016/j.neuropsychologia.2005.07.009

Sun, R., Slusarz, P., & Terry, C. (2005). The Interaction of the Explicit and the Implicit in Skill Learning: A Dual-Process Approach. *Psychological Review*, *112*(1), 159–192. https://doi.org/10.1037/0033-295X.112.1.159

Takács, Á., Shilon, Y., Janacsek, K., Kóbor, A., Tremblay, A., Németh, D., & Ullman, M. T. (2017). Procedural learning in Tourette syndrome, ADHD, and comorbid Tourette-ADHD: Evidence from a probabilistic sequence learning task. *Brain and Cognition*, *117*, 33–40. https://doi.org/10.1016/j.bandc.2017.06.009

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling Complex Meta-analytic Data Structures Using Robust Variance Estimates: A Tutorial in R. *Journal of Developmental and*

Life-Course Criminology, *2*(1), 85–112. https://doi.org/10.1007/s40865-016-0026-5

Taylor, J. S. H., Rastle, K., & Davis, M. H. (2013). Can cognitive models explain brain activation during word and pseudoword reading? A meta-analysis of 36 neuroimaging studies. *Psychological Bulletin*, *139*(4), 766–791. https://doi.org/10.1037/a0030266

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society*, *5*(4), 357–369. https://doi.org/10.1017/S1355617799544068

Teng, E., Stefanacci, L., Squire, L. R., & Zola, S. M. (2000). Contrasting Effects on Discrimination Learning after Hippocampal Lesions and Conjoint Hippocampal–Caudate Lesions in Monkeys. *The Journal of Neuroscience*, *20*(10), 3853–3863. https://doi.org/10.1523/JNEUROSCI.20-10-03853.2000

Theeuwes, J. (2010). Top-down and bottom-up control of visual selection. *Acta Psychologica*, *135*(2), 77–99. https://doi.org/10.1016/j.actpsy.2010.02.006

Thomas, K. M., Hunt, R. H., Vizueta, N., Sommer, T., Durston, S., Yang, Y., & Worden, M. S. (2004). *Evidence of Developmental Differences in Implicit Sequence Learning: An fMRI Study of Children and Adults*. *1*, 1339–1351.

Tiego, J., Testa, R., Bellgrove, M. A., Pantelis, C., & White, S. (2018). A Hierarchical Model of Inhibitory Control. *Frontiers in Psychology*, *9*, 25.

Tillmann, B., & McAdams, S. (2004). Implicit Learning of Musical Timbre Sequences: Statistical Regularities Confronted With Acoustical (Dis)Similarities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(5), 1131–1142. https://doi.org/10.1037/0278-7393.30.5.1131

Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: Consciousness as an unconscious re-description process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1412–1423. https://doi.org/10.1098/rstb.2011.0421

Tipton, E., & Pustejovsky, J. E. (2015). Small-Sample Adjustments for Tests of Moderators and Model Fit Using Robust Variance Estimation in Meta-Regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634. https://doi.org/10.3102/1076998615606099

Tomblin, J. B., Mainela-Arnold, E., & Zhang, X. (2007). Procedural Learning in Adolescents With and Without Specific Language Impairment. *Language Learning and Development*, *3*(4), 269–293. https://doi.org/10.1080/15475440701377477

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of Specific Language Impairment in Kindergarten Children. *Journal of Speech, Language, and Hearing Research*, *40*(6), 1245–1260. https://doi.org/10.1044/jslhr.4006.1245

Torriero, S., Oliveri, M., Koch, G., Caltagirone, C., & Petrosini, L. (2004). Interference of left and right cerebellar rTMS with procedural learning. *Journal of Cognitive Neuroscience*, *16*(9), 1605–1611. https://doi.org/10.1162/0898929042568488

Tóth-Fáber, E., Janacsek, K., & Németh, D. (2021). Statistical and sequence learning lead to persistent memory in children after a one-year offline period. *Scientific Reports*, *11*(1), 12418. https://doi.org/10.1038/s41598-021-90560-5

Trafimow, D. (2015). A defense against the alleged unreliability of difference scores. *Cogent Mathematics*, *2*(1), 1064626. https://doi.org/10.1080/23311835.2015.1064626

Treiman, R., & Kessler, B. (2006). Spelling as statistical learning: Using consonantal context to spell vowels. *Journal of Educational Psychology*, *98*(3), 642–652. https://doi.org/10.1037/0022-0663.98.3.642

Treiman, R., & Kessler, B. (2011). Similarities among the shapes of writing and their effects on learning. *Written Language & Literacy*, *14*(1), 39–57. https://doi.org/10.1075/wll.14.1.03tre

Tulving, E., & Markowitsch, H. J. (1998). Episodic and declarative memory: Role of the hippocampus. *Hippocampus*, *8*(3), 198–204. https://doi.org/10.1002/(SICI)1098-1063(1998)8:3<198::AID-HIPO2>3.0.CO;2-G

Turner, R. M., Bird, S. M., & Higgins, J. P. T. (2013). The Impact of Study Size on Meta-analyses: Examination of Underpowered Studies in Cochrane Reviews. *PLoS ONE*, *8*(3), e59202. https://doi.org/10.1371/journal.pone.0059202

Tzvi, E., Stoldt, A., Witt, K., & Krämer, U. M. (2015). Striatal–cerebellar networks mediate consolidation in a motor sequence learning task: An fMRI study using dynamic causal modelling. *NeuroImage*, *122*, 52–64. https://doi.org/10.1016/j.neuroimage.2015.07.077

Ullman, M. T. (2001a). *The Declarative/Procedural Model of Lexicon and Grammar*. 33.

Ullman, M. T. (2001b). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience*, *2*(10), 717–726. https://doi.org/10.1038/35094573

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*(1–2), 231–270. https://doi.org/10.1016/j.cognition.2003.10.008

Ullman, M. T. (2014). The role of declarative and procedural memory in disorders of language. *Linguistic Variation*, *13*(2), 133–154. https://doi.org/10.1075/lv.13.2.01ull

Ullman, M. T. (2015). The Declarative/Procedural Model: A Neurobiologically Motivated Theory of

First and Second Language. In B. VanPatten & J. Williams (Eds.), *Theories in Second Language Acquisition* (2nd edition, pp. 135–158). Routledge.

Ullman, M. T. (2016a). *The Declarative / Procedural Model: A Neurobiological Model of Language*. 953–968.

Ullman, M. T. (2016b). The Declarative/Procedural Model. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 953–968). Elsevier. https://doi.org/10.1016/B978-0-12-407794-2.00076-6

Ullman, M. T. (2016c). Chapter 76—The Declarative/Procedural Model: A Neurobiological Model of Language Learning, Knowledge, and Use. In G. Hickok & S. L. Small (Eds.), *Neurobiology of Language* (pp. 953–968). Academic Press. https://doi.org/10.1016/B978-0-12-407794-2.00076-6

Ullman, M. T., Earle, F. S., Walenski, M., & Janacsek, K. (2020). The Neurocognition of Developmental Disorders of Language. *Annual Review of Psychology*, *71*(1), 389–417. https://doi.org/10.1146/annurev-psych-122216-011555

Ullman, M. T., & Pierpont, E. I. (2005). Specific Language Impairment is not Specific to Language: The Procedural Deficit Hypothesis. *Cortex*, *41*(3), 399–433. https://doi.org/10.1016/S0010-9452(08)70276-4

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

Unsworth, N., & Engle, R. W. (2005). Individual differences in working memory capacity and learning: Evidence from the serial reaction time task. *Memory & Cognition*, *33*(2), 213–220. https://doi.org/10.3758/BF03195310

Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(4), 601–615. https://doi.org/10.3758/s13415-016-0417-4

Unsworth, N., & Robison, M. K. (2018). Tracking arousal state and mind wandering with pupillometry. *Cognitive, Affective and Behavioral Neuroscience*. https://doi.org/10.3758/s13415-018-0594-4

Vakil, E., Bloch, A., & Cohen, H. (2017). Anticipation Measures of Sequence Learning: Manual versus Oculomotor Versions of the Serial Reaction Time Task. *Quarterly Journal of Experimental Psychology*, *70*(3), 579–589. https://doi.org/10.1080/17470218.2016.1172095

* 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. https://doi.org/10.1177/0022219413508981

van Belle, J., van Raalten, T., Bos, D. J., Zandbelt, B. B., Oranje, B., & Durston, S. (2015). Capturing the dynamics of response variability in the brain in ADHD. *NeuroImage: Clinical*, *7*, 132–141. https://doi.org/10.1016/j.nicl.2014.11.014

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. https://doi.org/10.1080/10888438.2018.1491582

van der Lely, H. K. J. (2005). Domain-specific cognitive systems: Insight from Grammatical-SLI. *Trends in Cognitive Sciences*, *9*(2), 53–59. https://doi.org/10.1016/j.tics.2004.12.002

van der Lely, H. K. J., & Pinker, S. (2014). The biological basis of language: Insight from developmental grammatical impairments. *Trends in Cognitive Sciences*, *18*(11), 586–595. https://doi.org/10.1016/j.tics.2014.07.001

van der Lely, H. K. J., Rosen, S., & McClelland, A. (1998). Evidence for a grammar-specific deficit in children. *Current Biology*, *8*(23), 1253–1258. https://doi.org/10.1016/S0960-9822(07)00534-9

van Moorselaar, D., & Slagter, H. A. (2019). Learning What Is Irrelevant or Relevant: Expectations Facilitate Distractor Inhibition and Target Facilitation through Distinct Neural Mechanisms. *The Journal of Neuroscience*, *39*(35), 6953–6967. https://doi.org/10.1523/JNEUROSCI.0593-19.2019

Van Selst, M., & Jolicoeur, P. (1994). A Solution to the Effect of Sample Size on Outlier Elimination. *The Quarterly Journal of Experimental Psychology Section A*, *47*(3), 631–650. https://doi.org/10.1080/14640749408401131

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

* van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2019). Statistical learning abilities of children with dyslexia across three experimental paradigms. *PLoS ONE*, *14*(8), 1–29. https://doi.org/10.1371/journal.pone.0220041

† van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2021). The contribution of individual differences in statistical learning to reading and spelling performance in children with and without dyslexia. *Dyslexia*, *27*(2), 168–186. https://doi.org/10.1002/dys.1678

Vandenberghe, M., Schmidt, N., Fery, P., & Cleeremans, A. (2006). Can amnesic patients learn without awareness? *Neuropsychologia*, *44*(10), 1629–1641. https://doi.org/10.1016/j.neuropsychologia.2006.03.022

Verhoeven, B., Van der Steeg, A., Scherpbier, A., Muijtjens, A., Verwijnen, G., & van der Vleuten, C. (1999). Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. In *MEDICAL EDUCATION* (Vol. 33, Issue 11, pp. 832–837). BLACKWELL SCIENCE LTD.

Verneau, M., Van Der Kamp, J., Savelsbergh, G. J. P., & De Looze, M. P. (2014). Age and time effects on implicit and explicit learning. *Experimental Aging Research*, *40*(4), 477–511. https://doi.org/10.1080/0361073X.2014.926778

Verstynen, T., Phillips, J., Braun, E., Workman, B., Schunn, C., & Schneider, W. (2012). Dynamic Sensorimotor Planning during Long-Term Sequence Learning: The Role of Variability, Response Chunking and Planning Errors. *PLoS ONE*, *7*(10), e47336. https://doi.org/10.1371/journal.pone.0047336

Vicari, S. (2005). Do children with developmental dyslexia have an implicit learning deficit? *Journal of Neurology, Neurosurgery & Psychiatry*, *76*(10), 1392–1397. https://doi.org/10.1136/jnnp.2004.061093

Vicari, S., Marotta, L., Menghini, D., Molinari, M., & Petrosini, L. (2003). Implicit learning deficit in children with developmental dyslexia. *Neuropsychologia*, *41*(1), 108–114. https://doi.org/10.1016/S0028-3932(02)00082-9

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. https://doi.org/10.1002/jrsm.11

Virag, M., Janacsek, K., Horvath, A., Bujdoso, Z., Fabo, D., & Nemeth, D. (2015). Competition between frontal lobe functions and implicit sequence learning: Evidence from the long-term effects of alcohol. *Experimental Brain Research*, *233*(7), 2081–2089. https://doi.org/10.1007/s00221-015-4279-8

Virtala, P., Partanen, E., Tervaniemi, M., & Kujala, T. (2018). Neural discrimination of speech sound changes in a variable context occurs irrespective of attention and explicit awareness. *Biological Psychology*, *132*, 217–227. https://doi.org/10.1016/j.biopsycho.2018.01.002

Vogel, E. K., & Luck, S. J. (2000). The visual N1 component as an index of a discrimination process. *Psychophysiology*, *37*(2), 190–203. https://doi.org/10.1111/1469-8986.3720190 von Bastian, C. C., Blais, C., Brewer, G. A., Gyurkovics, M., Hedge, C., Kałamała, P., Meier, M. E., Oberauer, K., Rey-Mermet, A., Rouder, J. N., Souza, A. S., Bartsch, L. M., Conway, A. R. A., Draheim, C., Engle, R. W., Friedman, N. P., Frischkorn, G. T., Gustavson, D. E., Koch, I., … Wiemers, E. A. (2020). *Advancing the understanding of individual differences in attentional control: Theoretical, methodological, and analytical considerations* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/x3b9k

von Hippel, P. T. (2015). The heterogeneity statistic I2 can be biased in small meta-analyses. *BMC Medical Research Methodology*, *15*(1), 35. https://doi.org/10.1186/s12874-015-0024-z

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. https://doi.org/10.1016/S0022-0965(03)00030-4

Wagenmakers, E. J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of 1/fα noise in human cognition. *Psychonomic Bulletin and Review*, *11*(4), 579–615. https://doi.org/10.3758/BF03196615

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2013). *CTOPP-2 Comprehensive Test of Phonological Processing – Second Edition.* Pearson Clinical.

Walker, S., Gaskell, M. G., Knowland, V. C. P., Fletcher, F. E., Cairney, S. A., & Henderson, L. M. (2020). Growing up with interfering neighbours: The influence of time of learning and vocabulary knowledge on written word learning in children. *Royal Society Open Science*, *7*(3), 191597. https://doi.org/10.1098/rsos.191597

Wang, B., & Theeuwes, J. (2018). Statistical regularities modulate attentional capture independent of search strategy. *Attention, Perception, & Psychophysics*, *80*(7), 1763–1774. https://doi.org/10.3758/s13414-018-1562-3

Warbrick, T., Arrubla, J., Boers, F., Neuner, I., & Shah, N. J. (2014). Attention to Detail: Why Considering Task Demands Is Essential for Single-Trial Analysis of BOLD Correlates of the Visual P1 and N1. *Journal of Cognitive Neuroscience*, *26*(3), 529–542. https://doi.org/10.1162/jocn_a_00490

Ward, E. V., Berry, C. J., & Shanks, D. R. (2013). Age effects on explicit and implicit memory. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00639

Watkins, K. E., Vargha-Khadem, F., Ashburner, J., Passingham, R. E., Connelly, A., Friston, K. J., Frackowiak, R. S. J., Mishkin, M., & Gadian, D. G. (2002). MRI analysis of an inherited speech and language disorder: Structural brain abnormalities. *Brain*, *125*(3), 465–478.

https://doi.org/10.1093/brain/awf057

Wechsler, D. (2009). *Wechsler Individual Achievement Test – Third UK Edition (WIAT-III UK)*. Pearson Assessment.

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence – Second Edition (WASI-II)*. NCS Pearson.

West, G. (2018). Procedural and Declarative Memory and Language Ability in Children [Unpublished doctoral dissertation]. *University College London*, 340.

* West, G., Clayton, F. J., Shanks, D. R., & Hulme, C. (2019). Procedural and declarative learning in dyslexia. *Dyslexia*. https://doi.org/10.1002/dys.1615

West, G., Melby-Lervåg, M., & Hulme, C. (2021). Is a procedural learning deficit a causal risk factor for developmental language disorder or dyslexia? A meta-analytic review. *Developmental Psychology*, *57*(5), 749–770. https://doi.org/10.1037/dev0001172

†* West, G., Shanks, D. R., & Hulme, C. (2021). Sustained Attention, Not Procedural Learning, is a Predictor of Reading, Language and Arithmetic Skills in Children. *Scientific Studies of Reading*, *25*(1), 47–63. https://doi.org/10.1080/10888438.2020.1750618

†* West, G., Vadillo, M. A., Shanks, D. R., & Hulme, C. (2018). The procedural learning deficit hypothesis of language learning disorders: We see some problems. *Developmental Science*, *21*(2), e12552. https://doi.org/10.1111/desc.12552

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wiernik, B. M., & Dahlke, J. A. (2019). *Obtaining unbiased results in meta-analysis: The importance of correcting for statistical artefacts* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/9mpbn

Wierzchoń, M., Gaillard, V., Asanowicz, D., & Cleeremans, A. (2012). Manipulating attentional load in sequence learning through random number generation. *Advances in Cognitive Psychology*, *8*(2), 179–195. https://doi.org/10.5709/acp-0114-0

Wiig, E. H., Semel, E., & Secord, W. A. (2013). *Clinical Evaluation of Language Fundamentals—Fifth UK Edition*. Pearson Assessment.

Wilhelm, I., Rose, M., Imhof, K. I., Rasch, B., Büchel, C., & Born, J. (2013). The sleeping child outplays the adult’s capacity to convert implicit into explicit knowledge. *Nature Neuroscience*, *16*(4), 391–393. https://doi.org/10.1038/nn.3343

Wilkinson, L., & Shanks, D. R. (2004). Intentional Control and Implicit Sequence Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 354–369.

https://doi.org/10.1037/0278-7393.30.2.354

Williams, J. N. (2020). The Neuroscience of Implicit Learning. *Language Learning*, *70*(S2), 255–307. https://doi.org/10.1111/lang.12405

Willingham, D. B. (1999). Implicit motor sequence learning is not purely perceptual. *Memory & Cognition*, *27*(3), 561–572. https://doi.org/10.3758/BF03211549

Willingham, D. B., Greenberg, A. R., & Thomas, R. C. (1997). Response-to-stimulus interval does not affect implicit motor sequence learning, but does affect performance. *Memory and Cognition*, *25*(4), 534–542. https://doi.org/10.3758/BF03201128

Willingham, D. B., & Koroshetz, W. J. (1993). *Evidence for dissociable motor skills In Huntington's disease patients*. 10.

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *15*(6), 1047–1060. https://doi.org/10.1037//0278-7393.15.6.1047

Witt, K., Nuhsman, A., & Deuschl, G. (2002). Dissociation of Habit-Learning in Parkinson's and Cerebellar Disease. *Journal of Cognitive Neuroscience*, *14*(3), 493–499. https://doi.org/10.1162/089892902317362001

Woollams, A. M., & Patterson, K. (2012). The consequences of progressive phonological impairment for reading aloud. *Neuropsychologia*, *50*(14), 3469–3477. https://doi.org/10.1016/j.neuropsychologia.2012.09.020

Woollams, A. M., Ralph, M. A. L., Plaut, D. C., & Patterson, K. (2007). SD-squared: On the association between semantic dementia and surface dyslexia. *Psychological Review*, *114*(2), 316–339. https://doi.org/10.1037/0033-295X.114.2.316

Yan, X., Jiang, K., Li, H., Wang, Z., Perkins, K., & Cao, F. (2021). Convergent and divergent brain structural and functional abnormalities associated with developmental dyslexia. *ELife*, *10*, e69523. https://doi.org/10.7554/eLife.69523

Yang, Y., Bi, H.-Y., Long, Z.-Y., & Tao, S. (2013). Evidence for cerebellar dysfunction in Chinese children with developmental dyslexia: An fMRI study. *International Journal of Neuroscience*, *123*(5), 300–310. https://doi.org/10.3109/00207454.2012.756484

Yang, Y., & Hong-Yan, B. (2011). Unilateral implicit motor learning deficit in developmental dyslexia. *International Journal of Psychology*, *46*(1), 1–8. https://doi.org/10.1080/00207594.2010.509800

Zhao, J., & Luo, Y. (2017). Statistical regularities guide the spatial scale of attention. *Attention, Perception, & Psychophysics*, *79*(1), 24–30. https://doi.org/10.3758/s13414-016-1233-1

Zhao, S., Bury, G., Milne, A., & Chait, M. (2019). Pupillometry as an Objective Measure of Sustained Attention in Young and Older Listeners. *Trends in Hearing*, *23*, 233121651988781. https://doi.org/10.1177/2331216519887815

Zola-Morgan, S., & Squire, L. (1991). The Medial Temporal Lobe Memory System. *Science*, *253*(5026), 1380–1386. https://doi.org/10.1126/science.1896849