Grammatical generalisation in statistical learning: The contribution of vocabulary, explicit and implicit knowledge

Amanda Jayne Hickey

Doctor of Philosophy

University of York
Psychology

March 2022

Abstract

The learning and generalisation of grammatical regularities is fundamental to successful language acquisition and use. Research into statistical learning has started to consider how this process occurs through the implicit encoding and integration of grammatical regularities. Drawing on lexicalist and multi-componential theories, this thesis uses a statistical learning framework to explore two possible contributors to grammatical generalisation: 1) vocabulary, and 2) explicit and implicit knowledge. Across four studies adults and children learnt an artificial language containing two semantic categories denoted by a co-occurring determiner and suffix. The first two studies examined the role of vocabulary. Unlike child learners, in adult learners increasing the variability of vocabulary items in the learning context enhanced grammatical generalisation. When adults' vocabulary knowledge was reduced to that of the children, the adults' grammatical generalisation was likewise reduced to the same levels as the children's. In the first two studies, the explicit knowledge that emerged in the course of the study influenced adults' generalisation, but not the children's. Using more sensitive measures of explicit knowledge, the final two studies also found a contribution of explicit knowledge in adult learners, but only for regularities that showed the highest level of learning. In addition, the contribution of implicit knowledge was examined using two measures: a confidence rating measure revealed evidence of implicit knowledge of all the regularities tested, but this was not detected when using a novel, reaction time-based measure. To conclude, vocabulary is important for grammatical generalisation, with both explicit and implicit knowledge contributing to adult grammatical generalisation. The theoretical implications of these findings are discussed along with future developmental considerations.

List of Contents

List of Tables

List of Figures

List of Appendices

List of Accompanying Material

The experimental chapters each have associated pages on the Open Science Framework, containing pre-registrations (where included), experimental materials and stimuli, data and analyses.

Chapter 2

Experiment 1: https://osf.io/mrxwy/

Experiment 2: https://osf.io/2xf6v/

Chapter 4

Experiment 1: https://osf.io/wjhg5/

Experiment 2: https://osf.io/p3yvs/ & https://osf.io/xdvyg

Acknowledgements

First and foremost, I would like to thank my supervisors Dr Emma Hayiou-Thomas and Dr Jelena Mirkovic. I feel extremely lucky to have you as my 'academic parents', you have been incredibly supportive and encouraging and I have thoroughly enjoyed working with and learning from you. Your demonstrations of understanding and support as supervisors and mentors is something that has helped me to grow in confidence and is a model of support I will take forward with me.

My PhD journey was also greatly helped through the advice and support given by Dr Lisa Henderson and Dr Tom Hartley as part of my thesis advisory panel. You always seemed to capture exactly the feedback I needed to help move me forward or settle uncertainties I may have had, for which I am very grateful. Thanks also need to go to the schools that allowed me to come in and run my studies and to the hard work of MSc students who helped with the data collection for one of the experiments used in this thesis: Joseph Vogliqi, Ariadni Michail and Alyx Lace. I also wish to thank the many communities I've had the opportunity to be part of within the psychology department, from the ECR committee to knitR, lab journal club meetings and SLAM to my PhD cohort. The support and help I have gained from these communities, both academically and personally has been essential in getting me to the end of my PhD and I am very grateful to all those I have had the pleasure of meeting and working with through these communities. I would like to particularly thank Leah Kelly, Charlotte Ashton and Yujin Lee for their support and friendship along with Catia Ferreira De Oliveira for many interesting conversations and for so kindly stepping in to help me out of data troubles.

Another key person from my academic life to thank is Liz Du Pré. When I first started my undergraduate degree, I had many wrong assumptions about what I was capable of after years in the school system as a dyslexic child. As my dyslexia tutor throughout my degree, you helped to show me how I could help myself as well directly challenging my wrong assumptions. Your support and encouragement all those years ago, still have a direct impact on me today and allowed me to get to this stage in my academic career. Thank you.

Outside of academia I am indebted to many people, who without their support I would not have had the stamina to make it to the end of my PhD. Firstly, to my friends who are always there to cheer me on, particularly Lillian, Hannah, Ifalase and Holly. To my college tutor buddies, Kat and Molly who made my first year in York full of fun and who have been a constant source of support since. Managing a PhD along with a pandemic for the last 2 years would not have been possible without Doug. You have shown endless patience listening to my stressed ramblings, you have always been able to ease my worries, you have been my biggest cheerleader and you have always just been there, no questions asked, in whatever way I needed to help get

me to the finish line. You have supported me in ways I didn't even realise I needed, and I could not have asked for a better person to take on the world with.

Finally, but definitely not least, I would like to thank my family. Although no longer with us, I want to add a thank you to my grandparents, Pam, Brian, Joyce and Jack. They all, in many different ways throughout my childhood and young adulthood have helped me in my journey to this point. My mum and dad have always been my advocates; challenging teachers, getting the support I needed at school, as well as helping and encouraging me with whatever I wanted to do. Thank you for believing I could do this even when others said I could not.

Author's declaration

This thesis is a presentation of original work completed solely by the author, under the supervision of Dr Emma Hayiou-Thomas and Dr Jelena Mirkovic. This work has not been previously presented for an award at this or any other university. All sources are acknowledged as references. Some of the data collected were by MSc students under the author's co-supervision (Chapter 5, Experiment 1).

Publications

This thesis features one chapter that has been published in conference proceedings:

Chapter 3

Hickey, A., Hayiou-Thomas, E., & Mirkovic, J. (2019). Grammatical Generalisation in Statistical Learning: Is it implicit and invariant across development? In *CogSci* (pp. 1908-1914).

Conference Presentations:

Several findings have also been presented at conferences as follows:

Chapter 3

Hickey, A. J., Hayiou-Thomas, M. E., & Mirković, J. (2019, June). *Grammatical Generalisation in Statistical Learning: Is it Implicit and Invariant across Development?* Paper presented at the Interdisciplinary Advances in Statistical Learning, San Sebastián, Spain.

Hickey, A. J., Hayiou-Thomas, M. E., & Mirković, J. (2019, June). *Grammatical Generalisation in Statistical Learning: Is it Implicit and Invariant across Development?* Poster session presented at LuCiD's 5th Annual International Language and Communicative Development Conference.

Hickey, A. J., Hayiou-Thomas, M. E., & Mirković, J. (2019, July). *Grammatical Generalisation in Statistical Learning: Is it Implicit and Invariant across Development?* Poster session presented at the Child Language Symposium, Sheffield, UK.

Hickey, A. J., Hayiou-Thomas, M. E., & Mirković, J. (2019, July). *Grammatical Generalisation in Statistical Learning: Is it Implicit and Invariant across Development?* Poster session presented at the 41st Annual Meeting of the Cognitive Science Society, Montreal, Canada.

Hickey, A. J., Hayiou-Thomas, M. E., & Mirković, J. (2020, January). *Grammatical Generalisation in Statistical Learning: Is it Implicit and Invariant across Development?* Paper presented at the Experimental Psychological Society Meeting, London, UK.

Hickey, A. J., Hayiou-Thomas, M. E., & Mirković, J. (2020, June). *Grammatical Generalisation in Statistical Learning: Is it Implicit and Invariant across Development?* Poster presented at the 2nd Interdisciplinary Approaches in Language Science, IMPRS Conference, Online.

Chapter 1. Introduction

This review considers the generalisation of grammatical regularities within a statistical learning framework. The first section of this review outlines the literature that proposes statistical learning as a domain-general learning mechanism involved in the acquisition and learning of grammar. The literature detailed will be from a connectionist and usage-based approach to grammar acquisition and use, considered in relation to the nativist viewpoint the statistical learning literature aims to counter. The second part of this review will then explore factors which may affect grammatical generalisation within a statistical learning framework, namely the contribution of vocabulary, explicit and implicit grammatical knowledge. The potential implications of these on our understanding of grammatical generalisation from the current literature will be discussed.

1.1 Part One: Statistical learning and grammatical generalisation.

1.1.1 Language acquisition, grammar, and the role of language input.

The ability of young infants to assimilate and generalise the complex grammatical rules of their native language has been of great interest to researchers. Infants can achieve this without explicit instruction, in a relatively short amount of time and in what appears to be an impoverished language environment. Solving this 'logical problem of language acquisition' (Baker & McCarthy, 1981) has been a contentious issue between nativist and constructivist viewpoints. Put simply, nativists posit that grammatical knowledge is innate with input playing a small role in the learning process, and constructivists propose that it is the cognitive mechanisms used to learn grammar that are innate and that input plays a major role (Chomsky, 1965; Gathercole & Hoff, 2007; Gómez & Gerken, 2000; Saffran & Thiessen, 2007).

Theoretical perspectives which emphasise the role of the input in language acquisition can be found both within psychological and linguistic traditions. Usage-based theory is grounded in a linguistic tradition and is built on the idea that language is developed through its use and experience of it. It opposes the nativist idea of interpreting a developing child's grammatical productions through the lens of adult grammatical use, which this theory views as the end result of a usage-based learning process rather than the emergence of already present knowledge structures (Goldberg, 2005, 2009; Tomasello, 2000). Connectionism on the other hand is rooted in a broader context of cognitive and neural mechanisms of learning and memory, and views the development of language through use and experience which increases or decreases weighted connections in neural networks (Elman et al., 1998; Seidenberg, 1997). Both the usage-based theory and connectionism hold similar views on the greater role of input in language acquisition in contrast to nativist theories.

1.1.1.1 The minor role of language input from the nativist perspective.

Language acquisition is a complex learning process (Gómez & Gerken, 2000; Saffran & Thiessen, 2007), and one of its most impressive achievements is that young children not only learn grammatical regularities they are exposed to, but are able to generalise these to novel situations. In language acquisition, generalisation refers to the ability to go beyond specific exemplars to form abstract representations which can be used to recognise and produce an infinite number of novel grammatical forms (Goldberg, 1999; Seidenberg, 1997). This is a core concept when considering the role of input in language acquisition as generalisation is considered a hallmark of language proficiency. The importance of the role of language input can be judged on this generalisation ability (Chomsky, 1965). For the nativist, this ability to generalise is evidence of underlying innate grammar knowledge, most famously put forward as 'Universal Grammar' by Chomsky (1965). A key argument here relates to the proposed impoverished nature of language input. When the nativist theory was first put forward, the poverty of stimulus was seen as a major obstacle in language acquisition due to the lack of explicit corrective feedback from caregivers, the small number of input examples to generalise from, as well as input containing many errors. Thus, nativists concluded that generalisation could only happen with the support of an innate knowledge module (Chomsky, 1965; Gathercole & Hoff, 2007; Gerken, 2007; Saffran & Thiessen, 2007). So, how does this innate knowledge module, or universal grammar, drive the generalisation process in the face of the proposed impoverished input?

According to (Chomsky, 1965), universal grammar is the idea that infants are born with all the complex grammar parameters (rules) needed for language acquisition in any language. Additionally, this knowledge module is associated with a domain-specific language learning process or device. This innate language learning device then uses the language input to set the needed parameters within the knowledge module. As can be seen, this approach to grammar generalisation is a modular, domain-specific process, with language input only being needed to help set the parameters on knowledge that is already innately present (Chomsky, 1965; R. Frost et al., 2015; Gathercole & Hoff, 2007; Saffran & Thiessen, 2007). From the nativist perspective, this explanation helps to solve the 'logical problem of language acquisition' as it addresses the problem of impoverished input by reducing the importance and need for it. This theory of language acquisition was historically more dominant in the early understanding of language development. However, views started to shift around 40 years ago to the current more contemporary view which reinstated the importance of input (for a review of this paradigm shift see Fernald & Marchman, 2006).

1.1.1.2 The importance of language input from the connectionist perspective – what is innate?

As discussed, the argument for a smaller role for language input in grammar learning is due to the proposed impoverished nature of the language environment for the learning child. This assumption however has been challenged, with domain-general, connectionist theories arguing that the information needed to construct grammatical knowledge *is* present and abundant in the language environment (e.g., Monaghan et al., 2005). Furthermore, the theories propose that infants possess the necessary mechanisms to detect, assimilate and generalise this information without the need for this knowledge to already be innately present and that this mechanism is not domain-specific (Frost et al., 2015; R. Gomez & Gerken, 2000; Gomez & Gerken, 1999; Hsu & Bishop, 2010). Whilst a simplistic view of these theories could posit them as an argument against innateness in language acquisition, a more accurate description is that they question what exactly is innate in the developmental process (Elman et al., 1998; Frost et al., 2015; Gathercole & Hoff, 2007; Gerken, 2007; Gomez & Gerken, 2000; Saffran & Thiessen, 2007; Seidenberg, 1997).

A connectionist view of development is one in which developmental learning occurs from the interaction between nature and nurture. This view holds that the information needed is available in the environment and it is this information that interacts with learning mechanisms that drive development, not just in language but across different learning domains (Elman et al., 1998; R. Frost et al., 2015; Saffran & Thiessen, 2007). Thus, it is the learning mechanism/memory system (or its composite cognitive networks/neural correlates) which are more strongly influenced by innate factors rather than grammatical knowledge. From this perspective, language input has a much greater role in the development of grammar. A large body of empirical research which more soundly supports this view comes from studies into statistical learning. Statistical learning is being put forward as a domain-general learning mechanism in this connectionist learning process and is the focus of this review.

1.1.2 What makes natural language input rich in detectable information within a connectionist framework?

If connectionists are arguing for a lack of input poverty, it must in turn provide evidence of information ‘abundance’ in natural languages. The nativist argument of impoverished input (Chomsky, 1965; Gathercole & Hoff, 2007; Gerken, 2007; Saffran & Thiessen, 2007) led to extensive investigations of the input to explore this claim. Early research considering this focused on language modifications and corrective feedback made by the caregiver. Language modifications refers to caregivers use of a slower speaking speed and higher pitch when talking to children, a behaviour termed ‘parentese’. These modifications could be seen as cues or aids that highlight aspects of language to an infant language learner, helping to make the language input more informative. Corrective feedback provided by caregivers can be explicit or implicit in nature. Parents may either directly correct an ungrammatical utterance by a child as an exchange in and of itself or indirectly correct by recasting the ungrammatical utterance as part of the conversational exchange. These were areas of early interest as if language modifications and implicit and explicit corrective feedback are reliably present in an infant’s language input this could counter the ‘poverty of stimulus’ argument (Gathercole & Hoff, 2007).

However, research suggested these tools were not reliably present to overcome the 'poverty of stimulus' argument. The lack of cross-cultural use of parentese meant it was not a reliable form of input that could support grammar knowledge development from the input. In terms of intentional, explicit corrective feedback, it was found that this was not reliably provided by adults to infants, with the number of incidences that do occur not reaching the required number needed for it to account for the eventual language proficiency that children achieve. Research did find examples of implicit corrective feedback to be more common and could then increase the language information available to infants in the input, however it's likely that this provides more of a supportive rather than instructive role (see Gathercole & Hoff, 2007 for a review). These initial investigations provided some support for the nativist poverty of stimulus view, but it did not consider patterns within natural languages. When this is considered, it can be argued that there is a wealth of information available.

1.1.2.1 Statistical patterns in word boundaries - transitional probabilities.

Statistical learning is the ability to extract statistical patterns (e.g., Batterink et al., 2015; Frost et al., 2015; Gomez, 2017) and is a learning mechanism that has been observed across different input domains and in other species (see e.g., Frost et al., 2015; Krogh et al., 2013 for a review). Statistical learning then may be an innate, domain-general learning mechanism that supports language acquisition through the detection of the patterns available within language input. If this is the case, firstly the presence of statistical patterns within natural languages that could be detected by statistical learning needs to be established.

Much of the statistical learning research in this regard has focused specifically on transitional probabilities in the context of word boundaries. Before the grammar of a language can be learnt, a language learner needs to be able to identify words within speech, particularly as the speech signal does not include pauses between words. As such, the task of learning where the boundaries lie between words is considered to be one of the first challenges a young infant learning their native language must overcome (Christiansen et al., 1998; Jusczyk & Aslin, 1995; Saffran, Newport, et al., 1996). A number of suggestions for this process have been made (for reviews see Christiansen et al., 1998; Jusczyk & Aslin, 1995; Saffran, Newport, et al., 1996) including word segmentation as a result of lexical information, word isolation use by caregivers and the use of prosodic cues by the learner.

While the use of lexical information may be utilised by those who have prior linguistic knowledge of a language (Cutler, 1994) it is unlikely that infants have this prior knowledge to utilise. The use of isolation techniques is also unlikely as firstly, there are many grammatical words (e.g. *a, an, the, in, with,* etc.) which would not be used in isolation and secondly, caregivers rarely use isolation in practise, even when explicitly asked to teach vocabulary to their children (Woodward & Aslin, 1990). Prosodic cues (i.e., patterns of stress and intonation in a language) may be of some help as these cues can be found in some natural languages (e.g., French has a final word stress). Studies have shown infant sensitivity to these cues (e.g., Cutler, 1994). Despite this, it is unknown as to whether all languages have these cues and at least some pre-existing knowledge of word boundaries is needed to learn these patterns. Thus, prosodic cues may have more of a supportive role in this process (for a review see Saffran et al., 1996).

A statistical cue for word boundaries has been proposed in the form of transitional probabilities. Transitional probabilities are distributional patterns and are based on the idea that words are a sequence of speech sounds, staying the same wherever it occurs in a speech stream. Although learners do not have prior knowledge of these fixed sequences of speech sounds, they do have access to statistical information present across many exemplars in the speech stream. This information can be operationalised as strong correlations between sound sequences within the same word, and weak correlations across words (Saffran, Newport, et al., 1996). Thus, transitional probabilities are the likelihood of one sound following another within speech, with transitional probabilities usually being higher within words compared to transitions across word boundaries. For example, in the sound sequence '*pretty#baby'* the transitional probabilities between *pre* and *ty* are lower than between '*ty'* and '*ba'* (Saffran, Newport, et al., 1996).

A computational model study conducted by (Christiansen et al., 1998) supported the idea of transitional probabilities as a reliable statistical cue to word boundaries. The authors used a simple recurrent network trained on phonemic prediction (with explicitly provided information on phonemes, lexical stress, and boundaries between utterances). After exposure to a large corpus of English child-directed speech, the model was able to utilise these cues to correctly detect and learn word boundaries. This suggests the supportive role of prosodic (lexical stress) cues in word boundary learning as discussed earlier (Saffran, Newport, et al., 1996), but also that transitional probabilities are present and a reliable cue to word boundaries in natural languages. As such, the presence of transitional probabilities suggests one important element of language input where statistical information is abundant and a possible cue for statistical learning.

1.1.2.2 Statistical patterns in grammatical categories.

Beyond cues to word boundaries, natural languages are also abundant in terms of cues to other important aspects of linguistic structure. Grammatical categories (e.g., nouns and verbs) are a core part of language and knowledge of these are essential for successful grammar use as the grammatical category of a word has an impact on how it is used and interpreted (Lany & Saffran, 2010; Mintz, 2003; Monaghan et al., 2005; Redington et al., 1998). Grammatical categories also provide an example of the presence of statistical patterns within a language which cue to language knowledge beyond word boundaries.

Two important statistical cues to grammatical categories are phonological and distributional cues. Phonological cues consist of speech sounds which are associated with a word class; for example, phonemes are generally longer for nouns than verbs, and consonant clusters are more likely to occur at the start of an open class words (e.g., nouns and verbs like 'brother' and 'drive') than closed class words (e.g., prepositions like 'in', 'on' and 'at'). Phonological cues can occur at the word, syllable, or phoneme level. Distributional cues are cues taken from the linguistic context within which a word usually sits, for example by the use of frames where two co-occurring words (or morphemes) frame an interleaving word (e.g., '...*is* walk*ing'* where the co-occurring distributed '*is'* and *'ing'* frame and cue the interchangeable verb 'walk'; Mintz, 2002, 2003; Monaghan et al., 2005; Redington et al., 1998).

Monaghan et al. (2005) conducted a corpus study using an English-language database of child-directed speech (CHILDES) to consider whether phonological and distributional cues are present in a natural language and reliable indicators of grammatical categories. To investigate the presence of phonological cues in the corpora, sixteen types of phonological cues were identified (e.g., phoneme and syllable length, consonant clusters, presence of a syllable stress, syllable stress position within a word etc.) and then searched for in the corpora. Monaghan et al. (2005) found that these cues were present within the corpora and could reliably indicate grammatical categories. The authors also investigated the presence of distributional cues in the form of a bigram. These bigrams represent a distributional regularity through a target grammatical category word (e.g., a noun like 'gnome') being preceded by a context word (e.g., *the* gnome, *a* gnome, *this* gnome, *your* gnome etc.). Using bigrams of this kind, Monaghan et al. (2005) also found the presence of distributional cues in the corpora which also reliably indicated grammatical categories. Further to these findings the authors also found that the strength in indication for a grammatical category increased when phonological and distributional cues were combined.

Mintz (2003) also conducted a corpus study using the CHILDES database, although only focused on distributional cues. However instead of using bigrams like Monaghan et al. (2005), Mintz (2003) used 'frequency frames'. Here the distributional regularity is formed through non-adjacent, co-occurring words or morphemes that 'frame' the grammatical category word, as described earlier (e.g., *is* walk*ing*). They were termed 'frequency frames' as the frames used to assess whether they were reliable indicators for grammatical categories, were the frames that reached a given level of frequency in the corpus as well as having a variety of interleaving grammatical category words. When assessing these 'frequency frames' Mintz (2003) found that they were an effective, reliable and robust cue to grammar categories. These findings taken together with Monaghan et al. (2005) suggest that different types of distributional cues are not only present in natural languages but also effective at indicating grammatical categories.

Overall, these studies show that there is a wealth of informative statistical information present in natural languages, which can reliably indicate different grammatical categories. These results considered with the presence of transitional probabilities for word boundaries suggest that language input is not impoverished. Rather language input seems to be abundant in statistical structure that can tell the language learner useful information about the language being learnt.

1.1.3 Statistical learning: the use of statistical information by the language learner.

While demonstrating the presence of statistical information in natural languages is an important step in underlining the importance of the language input, it also needs to be shown that these statistical cues can be utilised by individuals when learning a language. Statistical learning is a strong candidate as a domain-general mechanism used by learners to detect and utilise this information (e.g., Batterink et al., 2015; Frost et al., 2015; Gomez, 2017).

1.1.3.1 In the beginning was the… word boundary.

One of the first studies to look at the role of statistical learning in linguistic development was conducted by (Saffran, Aslin, et al., 1996). The authors focused on how infants solve the problems of speech segmentation, as discussed earlier, and put forward evidence of 8-month-old infants using statistical learning to detect transitional probabilities to identify word boundaries. This study consisted of a two-minute familiarisation task, where infants listened to a continuous stream of four tri-syllabic nonsense words. This was followed by two test conditions where the infants listened to either ‘words’ present or ‘non-words’ which were not present in the familiarisation task. The stimuli were designed to contain high transitional probabilities within ‘words’ and low transitional probabilities between ‘words’ (word boundaries). The authors found a significant difference between the two test conditions, with infants showing a novelty preference for the ‘non-words’ that had not been presented in the familiarisation task.

While this showed the infant's ability to extract serial-order information, serial-order is not the only indication of word boundaries. Thus, a further experiment replicated and extended these findings through the added use of ‘part-words’ in the testing phase. After listening to the same familiarisation task, infants were tested using two ‘words’ from the familiarisation task and two ‘part-words’. ‘Part-words’ consisted of syllables used in the familiarisation task but did not use the same transitional probabilities as the familiarisation ‘words’. The authors found infants showed a novelty preference for these ‘part-words’, indicating the infants had used transitional probabilities to detect word boundaries. A possible confound in these studies is that the high within word transitional probabilities were more frequent, meaning that the results could be demonstrating a frequency effect. However, subsequent studies helped to cement this finding by controlling for transitional probability frequency, supporting the interpretation that the infants are using conditional probability information to detect word boundaries (Aslin et al., 1998).

These landmark studies provided a powerful demonstration that infants can use statistical cues (e.g., transitional probabilities) and thus may be able to take advantage of these cues in natural language learning. Due to the artificial nature of the ‘language’ used in these experiments, however, it was necessary to also demonstrate that infants can use statistical learning to detect word boundaries in natural languages. Later research using a similar methodology but with fluent speech streams of an unfamiliar natural language (Italian for infants from an English language environment) has replicated these results, including controlling for frequency effects (Pelucchi et al., 2009). This provides stronger support for the use of statistical learning in detecting the statistical information present (e.g., transitional probabilities) to identify word boundaries in natural languages. These studies show that young infants can utilise a powerful statistical learning mechanism to help them learn a key aspect of language - word boundaries.

1.1.3.2 Grammatical categories and statistical learning with phonological and distributional cues.

As discussed, statistical cues in the form of phonological and distributional cues, are present in natural languages and reliably indicate grammatical categories (e.g., nouns and verbs; Mintz, 2003; Monaghan et al., 2005). Again, as for the use of transitional probabilities in word boundary detection, it needs to be established as to whether learners can utilise these statistical cues. For example, in the same series of studies discussed earlier, Monaghan et al. (2005) started to test whether English-speaking adult participants were able to detect these English statistical cues and use them to build knowledge of word categories in an artificial language. The authors built an artificial language which included distributional and phonological cues similar to those found present in the CHILDES database. English-speaking adults were able to use these cues to detect word categories within this artificial language. The results showed that phonological cues were more reliable, particularly in low frequency words but that participants were able to use distributional cues in the high frequency words.

These results are promising in showing the ability to detect and use phonological and distributional cues in grammatical category learning; however, the participants were explicitly asked to look for patterns in the artificial language. This is pertinent as statistical learning is proposed to be an implicit process which takes place without the need for conscious awareness or overt pattern finding behaviour, particularly when considering language acquisition in infants (Aslin, 2017; Aslin & Newport, 2012; Batterink et al., 2015; Gómez, 2002; Hall et al., 2018; Saffran, Aslin, et al., 1996). In line with this, there is a large body of research that has considered the use of phonological and distributional cues by infants and children, where participants are not directed to look for patterns (see Romberg & Saffran, 2013; Saffran & Kirkham, 2018 for reviews).

As an example which bridges the gap between adults and children here, Gómez (2002) showed that both adults and infants were able to use non-adjacent dependencies (distributional cues) to distinguish between two artificial grammatical categories. Non-adjacent dependencies are two co-occurring words or units that have one or more interleaving words or units in between, like the frequent frames used by Mintz (2003). The language consisted of an *a*X*b* and *c*X*d* type structure, with the *a, b, c* and *d* elements providing the non-adjacent dependencies constructed to cue two grammatical categories, and the X element providing an interchangeable arbitrary stem (e.g., *a*X*b* - <u>pel</u> wadim <u>*vot*</u>). Gómez (2002) manipulated the variability of the trained stimuli by creating conditions with varied training sets by changing the number of interchangeable stems

(X). This effectively changed the number of items participants experienced in each grammatical category for the different training sets.

Through this manipulation Gómez (2002) found that adults endorsed training items to a high level in all set size conditions but were more able to reject incorrect test items in the highest set size condition. A similar finding was found in 18-month-old infants in that a novelty preference between trained correct and incorrect items (measured through a head-turn preference procedure) was only found in the largest set size condition. The role of variability during exposure will be discussed in more detail later in section 1.2.2. For the present, this result shows the use of a distributional cue for learning grammatical categories in both adults and children. More recently, Hall et al. (2018) demonstrated that typically developing children were also able to extract grammatical categories from distributional cues alone without the need for other cues. While the authors of this study suggest that other statistical cues are not necessarily needed to aid an individual's detection and generalisation of grammatical categories, other research does demonstrate the use of phonological cues in this process as well.

In a series of experiments Lany and Saffran (2010, 2011) demonstrated the use of both phonological and distributional cues in the acquisition of grammatical categories. Lany and Saffran (2010) constructed two-word grammatical phrases (*a*X or *b*Y construction) with either distributional or distributional and phonological cues denoting two lexical (grammatical) categories. To create a phonological cue two word-types were created which differed in the number of syllables they contained, with disyllabic *X* words (e.g., *coomo*) and monosyllabic *Y* words (e.g., *deech*). To create the distributional cue, these disyllabic and monosyllabic words were preceded by one of two determiner types. The determiner was either an '*a'* word (*ong* or *erd*) or a '*b'* word (*alt* or *ush*). Each created phrase was paired with a picture, with one of the grammatical categories using animal pictures and the other vehicle pictures. Twenty-two-month-old infants were able to demonstrate learning and generalisation of the lexical categories when only the distribution cue was present, however learning was enhanced when both cues were present within the language. From these results, Lany and Saffran (2010) concluded that infants can integrate distributional and phonological cues to detect, learn and generalise grammatical categories. These findings suggest, while infants can use distributional cues on their own, phonological cues reinforce and strengthen this learning, demonstrating the use of both cues in grammatical category learning.

1.1.3.3 Grammatical Categories, statistical learning and the possibility of semantic cues.

Lany and Saffran's (2011) next study reported results which replicated the use of phonological and distributional cues found in their previous study discussed above. However, in this study the authors also considered the interaction of these two statistical cues with the semantic information provided by the referent pictures introduced during training: the semantic categories of animals and vehicles. Lany and Saffran's (2011) examined generalisation performance to novel items after training as they did in their earlier study, however this time they had an addition task which either presented the phonological or distributional cue with a novel referent picture. Here the author's found that infants were able to use just phonological and just distributional cues when detecting and generalising the lexical category/semantic mappings of words. Together with the previous study, this evidence demonstrates the infants' ability to use both cues jointly and in isolation.

The authors also investigated the relationship between cue use and standardised measures of vocabulary. Interestingly, the authors found that infants with higher vocabulary abilities were more likely to use distributional cues and those with lower vocabulary abilities were more likely to use phonological cues. This finding also fits with Hall et al.'s (2018) finding that older children's generalisation performance using a distributional cue can be predicted from their raw score on a standardised vocabulary task. This link between statistical cue use and vocabulary ability is an interesting one which can suggest a broader link between vocabulary and grammar. This possibility will be considered further in the second part of this review in section 1.2.2.

Within the statistical learning literature, this study by Lany & Saffran (2011) was the first to demonstrate a possible involvement of semantics when building grammatical category representations. Providing evidence for a potential role for semantic cues in this statistical learning process. However, direct testing of the role of semantic cues in statistical learning for the learning of grammatical categories is sparse in the statistical learning literature. Whilst Lany and Saffran (2010, 2011) show how phonological and distributional cues help to support the learning of semantic mappings for grammatical categories, these studies do not test specifically the role of semantic cues in the learning of grammatical categories. It is possible that there exists a reciprocal relationship, such that phonological and distributional cues are used to support the learning of semantic mappings, semantic information can in turn support the detection and generalisation of grammatical categories. Support for this idea can come from outside the statistical learning literature, with research taken from linguistics and other areas of psycholinguistics that consider the idea of semantic bootstrapping and the role of gender markers, a form of grammatical category used in many natural languages.

1.1.3.4 Grammatical Categories and linguistics: Bootstrapping and grammatical gender

The role of semantic cues has, so far, not received much attention within statistical learning, although this is starting to change (e.g. Lany & Saffran, 2010; 2011 and discussed later Brown et al., 2022; Mirkovic et al., 2011). However, it has been extensively considered within linguistics and other areas of psycholinguistic research. Most work within this area stems from theories of semantic bootstrapping (e.g., Bowerman, 1973; Braine, 1987; Pinker, 1979, 1984, 1989). This is the idea that the semantic content of words is ever-present in natural languages and as such, words with similar meanings may serve as a cue to grammatical categories. For example, an infant would frequently hear a ‘the’ before a noun, the noun that follows also usually refers to an object, helping to build the semantic idea of what a noun commonly refers to. In the same way that exposure to the framing of a verb with ‘is’ and ‘ing’, will most often be experienced with an action, helping to build the semantic relation of action. Bowerman (1973) and Pinker (1979, 1984, 1989) considered semantic bootstrapping within a nativist framework, where semantic information would aid the learning of word classes through the use of a specific language learning mechanism. This mechanism allows for some form of self-editing ability which enables the child to consider what is common across all nouns or verbs within their long-term memory. Braine (1987) argues against this form of bootstrapping and proposes a theory more akin to linguistic, usage-based theories of language acquisition. The theory incorporates phonological regularities as well (phonological and semantic bootstrapping), and proposes the use of phonological and semantic cues is based on a child’s sensitivities to different regularities, linking this process to human pattern learning abilities and associative learning.

This theory was partially based on an artificial language study with adults, reported in the same chapter Braine (1987) proposes their alternative theory. This study considered the use of both phonological and semantic cues by creating an artificial language which contained two grammatical categories based around natural gender (male and female). Natural gender was used here as a link to gender markers found in natural languages. Gender-markers can take the form of a distributional cue (e.g., determiners) and a phonological cue (e.g. suffixes) that provide a reference to the grammatical gender class that a noun belongs to. Commonly these are feminine and masculine categories, an example of which can be found in German where the determiner ‘*der’* and a number of suffixes (e.g. ‘*ist’, ‘ich’* and *‘ling’* among others) denote masculine nouns and the determiner ‘*die’* and a number of different suffixes (e.g. *‘enz’, ‘ung’* and *‘keit’* among others) denote feminine nouns. Gender classes in natural languages are also often associated with semantic cues (Corbett, 1991; Mirković et al., 2005; Zubin & Köpcke, 1981). Drawing from this natural grammatical category, Braine (1987) created their simplified artificial language by grouping their two categories using the semantic concept of natural gender.

To denote the categories, artificial nouns were arbitrary paired with pictures of people with either traditional female professions or male professions, thus providing the semantic cue of natural gender. Additional pictures of inanimate objects arbitrarily were also placed into either category, to mimic arbitrary pairing of this nature from natural languages (e.g. rug is classed as masculine in German). The phonological cue was provided in the form of a plural indicating suffix, like the plural ‘*s*’ in English, but with a number of different suffixes created to indicate more specific amounts (e.g. one suffix indicated ‘singular, another ‘dual’ and a third ‘plural’). Each gender category contained different suffixes for the three amounts, again mimicking the multiple different suffixes that can be found for each grammatical gender class in natural languages (see the German example above). An additional control language was then created where the phonological regularities were present but not the semantic. After being trained on the language through a paired-associative learning task, participants were tested on item-based knowledge using a production task and then the same production task was used to assess grammatical generalisation of the grammatical regularities on unencountered items.

Using this artificial language paradigm, Braine (1987) found that more adults were able to generalise the grammatical regularities when semantic cues were present compared to when only phonological cues were present in the control language. This provides early support for a role of semantic cues within adult grammatical generalisation and for its use within a statistical learning paradigm. Many studies have subsequently built on this proposed theory and finding, with work particularly considering Braine's (1987) usage-based ideas and linking phonological and semantic bootstrapping to general cognition (Brown et al., 2022; Culbertson et al., 2017; Culbertson et al., 2019; Ferman et al., 2009; Ferman & Karni, 2010; Schwab et al., 2018; Williams, 2005). To help make this link, this body of work uses variations on a semi-artificial language, which includes novel affixes paired with real nouns and/or verbs taken from the the participants' native language, rather than using novel words to create new nouns. This general approach was taken in an attempt to replicate within a linguistic paradigm, similar studies into non-linguistic skill learning which considers novel skill learning based on existing knowledge (see Ferman et al. 2009; Ferman & Karni, 2010). By doing this, these studies aimed to link and allow for comparison between general skills learning and linguistic grammatical category learning and generalisation, building on usage-based, domain-general theories of language acquisition. In these studies, the nouns and/or verbs usually fall within one of two semantic categories (often animate or inanimate) of which the novel affix also provides a phonological or sometimes distributional cue to. For example, in Experiment 1 of Williams (2005) the prefixes '*gi*' and '*ul*' are paired with the animate category which contains nouns such as '*dog', 'cat' and 'pig'*; the prefixes '*ro*' and '*ne*' are paired with the inanimate category which contains nouns such as *'sofa', 'cup' and 'television'*.

In support of Braine's (1987) theory and findings, a number of studies using variations on this basic semi-artificial language have shown a role for semantic cues in the generalisation of grammatical categories in adults (Culbertson et al., 2017; Culbertson et al., 2019; Ferman et al., 2009; Ferman & Karni, 2010; Schwab et al., 2018; Williams, 2005). However, there are mixed findings for children with some finding they can (Brown et al., 2022; Culbertson et al., 2019) and others that they cannot (Ferman & Karni, 2010; Schwab et al., 2018). These findings suggest that certain conditions are needed to enable generalisation from semantic cues in children. In particular semantic cues seem to need to be deterministic (Brown et al., 2022; Culbertson et al., 2019) or need to be introduced before other cues (phonological cues in the case of Culbertson et al., 2019). Using grammatical categories that do not already exist in participants' native language may also aid the use of semantic cues for generalisation in children. This is because using known grammatical categories may allow prior language knowledge to interfere with the acquisition and generalisation of new grammatical category knowledge (Ferman & Karni, 2010; see also analysis and discussion of this result in Brown et al., 2022).

The ease at which adults generalise using semantic cues compared to children is an interesting finding, particularly when considering language acquisition. When multiple cues are present, often phonological and semantic cues in the studies under discussion here, adults seem to be able to generalise when cues are deterministic or probabilistic (e.g. Brown et al., 2022; Schwab et al., 2018). They also seem to show a preference for what is most salient within the language (Culbertson et al., 2017) and if all of the cues available are equally salient the preference seems to fall to semantic cues (Culbertson et al., 2019). Whereas children, as discussed above, seem to need specific conditions in order to use semantic cues and if the multiple cues presented are deterministic, equally salient and simultaneously, their preference seems to fall to the phonological cues (Culbertson et al., 2019, Experiment 2). This preference is also mirrored within in natural language, where semantic cues are often more reliable than phonological cues, however children still show a preference for phonological cues over semantic (see Culbertson et al., 2017 & Culbertson et al., 2019 for a review and discussion). This could suggest either a general processing preference for phonological cues or that a preference has evolved from being able to access the phonological features of a language before the semantic (Culbertson et al., 2017; Culbertson et al., 2019; Brown et al., 2022). Findings from Culbertson et al. (2019) suggests it could be latter although this is not conclusive.

Overall though, this body of work suggests a role for semantic cues in grammatical category generalisation in adults and under certain conditions in children. Patterns of findings are similar to those within non-linguistic skill learning which supports domain-general theories of semantic bootstrapping and the role of semantic cues in grammatical generalisation. As such it also supports a role for semantic cues in building grammatical category knowledge within statistical learning paradigms, building on findings from Lany and Saffran (2010, 2011). The findings discussed here and their contribution to hypotheses regarding grammatical category learning within statistical learning frameworks, is also supported by research within natural languages, considering the online processing of grammatical gender markers in Spanish.

Lew-Williams and Fernald (2007, 2010) found that young children learning Spanish as well as native adult speakers utilise these distributional-semantic cues in online sentence processing, when learning novel Spanish nouns. The authors also showed that when non-native adult learners of Spanish were learning the same novel nouns, they were not able to take advantage of these gender-marker cues in on-line processing. This is despite the adult second-language learners being able to report the grammatical 'rule' here. The difference in real-time processing use of gender-markers between native and second-language learners here is noteworthy and is a point that will be returned to later in section 1.3.3.2. For the moment, this finding shows that semantic categories/cues provide useful information to native speakers during on-line language processing. This lends support to the idea that semantic cues are firstly, present in natural languages and secondly, can be utilised to aid online processing of grammatical categories. The semi-artificial language findings discussed here, along with the Lany and Saffran's (2010, 2011) studies in turn suggest statistical learning mechanisms could utilise them when learning grammatical categories as well.

However, while this body of work may support the use of semantic cues for generalisation within statistical learning, there are aspects of the methodologies that differ from statistical learning paradigms that may influence the findings and also move further away from reflecting real-world language learning. An important consideration for the statistical learning literature which aim to consider language acquisition and learning under natural conditions as closely as possible. Firstly, much of the work described here uses a semi-artificial language, using real words from participants' native language for the nouns and verbs that are incorporated. This approach does have positive implications in that it has helped to link linguistic to non-linguistic skills learning as well as also showing patterns of behaviour that are found within natural language acquisition (e.g. the phonological cue preference in children). It is also argued that this methodology helps to reduce cognitive demands on participants, particularly children, as participants can use prior lexical knowledge to support them within the task. This does allow for the investigation of areas of language processing that might otherwise be inaccessible within completely novel stimuli (Brown et al., 2022). Despite this, it does mean the methodology moves further away from mimicking natural language learning, infants acquiring their first language do face and meet the challenge of learning completely novel 'stimuli' without being able to bring prior lexical knowledge to this process. Adults and children also navigate learning second languages where the phonological lexical knowledge will be completely or mostly novel to them. Secondly, all of the studies referenced also use direct, item-level feedback of varying types as part of the training protocol (e.g. Culbertson et al., 2019; Schwab et al., 2018; Ferman & Karni, 2010; Brown et al., 2022). As discussed in section 1.1.2, in real-world language acquisition feedback is not this prevalent or consistent.

Given these diversions from real-world language learning and acquisition, it is important to assess whether the same pattern of findings occurs when participants cannot bring prior language knowledge to the task and do not receive feedback. Lany and Saffran's (2011) findings suggest that a similar role for semantic cues within statistical learning paradigms might be the case (at least in children) as their work uses novel words for their lexical items and did not include feedback in their training. However direct examination using similar statistical learning paradigms is needed to confirm this. While using a statistical learning paradigm will still not perfectly mirror the natural language acquisition process, it can help to take the next step towards mimicking and thus understanding, natural language learning.

1.1.3.5 Grammatical Categories, statistical learning and grammatical gender

While research considering semantic cues within statistical learning is lacking, Mirkovic et al. (2011) is a notable exception along with Lany and Saffran (2010; 2011). Mirkovic et al. (2011) provides more direct evidence for the possible role of semantic cues in adults (building on Lany & Saffran, 2010; 2011) within a statistical learning framework. This study utilised grammatical gender from natural languages by creating an artificial language consisting of two gender classes (feminine and masculine) by incorporating distributional, phonological, and semantic cues. Semantic cues were incorporated using word-picture pairings with pictures denoting human occupations/characters or different types of animals. "Feminine" words consisted of birds, insects, domestic animals and human occupations or characters depicted as female. Distributional cues were incorporated using the determiner *tib* and phonological cues with the suffixes *esh* or *eem.* "Masculine" words consisted of fish, reptiles, wild animals, and occupations/characters depicted as male, and used the determiner *ked* and the suffixes *aff* or *ool*. Each picture was then also paired with an interleaving, arbitrary stem word which was created using English pronounceable non-words. This created a similar artificial language to Brain (1987). Participants were trained using a large vocabulary set built through the arbitrary stems, within an implicit training paradigm (Breitenstein et al., 2007; Dobel et al., 2010). Thus, like Lany and Saffran (2010; 2011) real words were not used within the stimuli and participants did not recieve any feedback during training or testing.

Mirkovic et al. (2011) found that in their adult participants, the role of phonological and distributional cue use in grammatical category learning was replicated, as these cues were found to aid the learning of the two grammatical gender classes. Additionally, this investigation considered the role of semantic cues and found these cues influenced the use of phonological and distributional cues, despite the lack of explicit semantic cue training. In fact, these implicit semantic cues, embedded in the stimuli, were not intrinsically useful for the training and testing tasks. Despite this, participants still utilised the semantic cues and demonstrated greater accuracy in generalisation tasks when phonological-semantic and distributional-semantic mappings were consistent with the regularities that had been present in the training stimuli. In addition, the implicit semantic cues also significantly hindered participant performance when distributional-semantic mappings were inconsistent.

Together with the findings presented within section 1.1.3.4, from Lany and Saffran (2010, 2011) and Lew-Williams and Fernald (2007, 2010) this evidence suggests that semantic cues have a role in the development of grammatical categories. Mirkovic et al. (2011) provides more direct evidence that statistical learning mechanisms may be able to utilise these cues (as it does for phonological and distributional cues) when grammatical category learning takes place; as suggested by the studies using semi-artificial language methodologies (Brown et al., 2022; Williams, 2005; Ferman et al. 2009; Ferman & Karni, 2010; Schwab et al., 2018; Culberston et al., 2017; Culbertson et al., 2019). However, as studies that examine multiple cues for grammatical category learning are only a minority in the statistical learning literature, more research is needed to fully explore the role of semantic cues here. As well as see whether the same findings and pattern of results from the semi-artificial language studies are found within statistical learning paradigms. There have also been recent calls for statistical learning research to expand and investigate more complex and naturalistic language processing (Frost et al., 2015; Frost et al., 2019). Further investigation of the role of semantic cues within the context of multiple cues would also help to answer this call.

1.1.4 Part 1: Summary and Conclusions

1.1.4.1 Nativist/connectionist perspectives and further statistical cue considerations.

The growing body of evidence for the presence of reliable statistical information in a child's language input and the ability of adults and children to detect these statistical cues through a domain-general learning mechanism, support the connectionist idea that grammatical knowledge is not innate. The discussed research helps to solve the 'logical problem of language acquisition' in two ways. Firstly, by demonstrating the lack of poverty in an infant's language environment, a key argument for nativist accounts of grammar acquisition (Chomsky, 1965). This is shown through corpus research detailing the presence of reliable statistical cues (e.g., Mintz, 2003; Monaghan et al., 2005; Redington et al., 1998) including phonological, distributional and semantic cues to grammatical structure in natural languages (e.g., Lew-Williams & Fernald, 2007, 2010).

Secondly, statistical learning, the proposed cognitive mechanism that enables infants to capitalise on this rich linguistic environment, is thought to be a domain-general learning mechanism (e.g., Batterink et al., 2015; Frost et al., 2015; Gomez, 2017), and as such crucially different from the domain-specific 'language learning device' (Chomsky, 1965). The domain-general nature of the statistical learning mechanism is shown through the ability of adults, children, infants, and other species to use it (R. Frost et al., 2015; Krogh et al., 2013). It has also been demonstrated by the use statistical cues across a range of linguistic domains, including transitional probabilities to detect word boundaries (Aslin et al., 1998; Saffran, Aslin, et al., 1996), and the detection and use of phonological and distributional cues when learning grammatical categories in adult and children (Gómez, 2002; Hall et al., 2018; Lany & Saffran, 2010, 2011; Mintz, 2002; Mirkovic et al., 2011).

Semantic cues have also been found to interact with the learning and use of phonological and distributional cues within statistical learning studies, suggesting that they also play a role in forming grammatical category knowledge (Lany & Saffran, 2011; Lew-Williams & Fernald, 2007, 2010; Mirković et al., 2011). Research from linguistics and other areas of psycholinguistics also supports this idea (e.g., Brown et al., 2022; Culbertson et al., 2017; Culbertson et al., 2019; Ferman et al., 2009; Ferman & Karni, 2010; Schwab et al., 2018; Williams, 2005). However, evidence for the use of semantic cues is sparser within the statistical learning literature thus further research considering the role of semantic cues is needed which would also support calls for more complex and naturalistic statistical learning studies (Frost et al., 2015; Frost et al., 2019).

1.1.4.2 Statistical cues and grammatical generalisation.

The ability to generalise grammatical regularities is a hallmark of language proficiency. Berko's (1958) now famous *'wug'* test was one of the first to show empirically that adults and children were able to generalise grammatical regularities within a natural language. The literature presented so far supports Berko's (1958) findings, with evidence of adults and children utilising statistical cues for grammatical categories to learn *and* generalise grammatical regularities (e.g., Gómez, 2002; Hall et al., 2018; Lany & Saffran, 2010, 2011; Mintz, 2002; Mirkovic et al., 2011).

The more recent statistical learning literature however can go further than Berko (1958) and suggest *what* is being generalised - statistical regularities learnt from the environment through a domain-general learning mechanism, rather than generalising from grammatical knowledge present from birth. However, Berko's (1958) *'wug'* tests and the statistical learning literature are limited in what they can demonstrate regarding generalisation. They show language users *can* generalise and suggest the nature of *what* they are generalising, but not *how* they are generalising. Exploring the *how* is important as beyond providing an understanding of a key cognitive mechanism supporting language learning and use, it would have relevant, applied implications in several areas, such as second-language learning and in understanding developmental language disorder, which is characterised by particular difficulties in grammar learning (Hsu & Bishop, 2010; Lammertink et al., 2017). In the next section of this review, I will consider how a statistical learning framework can be applied to consider this question.

1.2 Part Two: Potential influencing factors on grammatical generalisation – starting to explore the *'how'.*

1.2.1 Grammatical generalisation and the role of representations.

The connectionist approach offers a framework to consider the how of linguistic generalisation. In this approach, knowledge is operationalised as complex networks of neural connections that support behaviour with these complex neural networks being referred to as 'representations' (Elman et al., 1998).

Considering the process of how these representations are formed may be useful when examining the *how* of grammatical generalisations. Representations are unlikely to be innate, particularly when referring to higher-level knowledge. Rather, they are argued to be the product of innate learning mechanisms interacting with experiences of the environment as well as prior knowledge (Elman et al., 1998). In this view, grammatical knowledge is the result of building a network of representations that supports the use of language (language behaviour), and this would occur through the interaction of learning mechanisms (such as statistical learning), language experience and prior language knowledge. This process suggests a domain-general view, rather than domain-specific one. Representations are formed from what experience shows to be important to support the behaviour of interest, in this case language behaviour (Elman et al., 1998). In other words, the learning process does not depend on whether the necessary information is phonological, semantic, lexical, or grammatical, but rather it creates the connections needed to form the representations that will support the desired behaviour. It is this domain-general view of representation formation that is key to identifying factors that could influence grammatical generalisation.

Statistical learning fits well into this framework, supporting its use for investigating the *how* of grammatical generalisation. In terms of language behaviour, the statistical learning literature suggests this as an important, domain-general learning mechanism for building language representations through language experience. Research has shown that the domain of the input statistical learning is accessing may present different constraints on how it is used, which could be seen as statistical learning being domain specific. However, it's use across different domains in both linguistic and non-linguistic stimuli (e.g., Conway, 2020; Conway & Christiansen, 2005; Frost et al., 2015; Saffran & Thiessen, 2008) as well as across written and spoken language (e.g., Arciuli, 2018; Gómez, 2002; Saffran, Aslin, et al., 1996; Seidenberg & MacDonald, 2018) support the idea of statistical learning being domain-general. In terms of the resulting language representations, it suggests the knowledge gained from these different domains are stored and processed in similar ways. Thus, if information from two domains can inform a behaviour, like for Lany and Saffran (2010, 2011) and Mirkovic et al. (2011), they could be stored and processed together. This aligns with constructivist language theories which suggest different aspects of language knowledge are not processed separately, but in conjunction with one another, and in a similar way to non-linguistic information, for example with general-purpose memory systems (Bates & Goodman, 1997; Batterink et al., 2019; Christiansen & Chater, 2015; Isbilen et al., 2018; MacDonald et al., 1994).

So, what does this mean for the *how* of grammatical generalisation? This section of the chapter will focus on the implications of two aspects of this proposed view of representation formation in terms of grammatical generalisation processing. The first consideration is the idea that different aspects of language use similar underlying storage and processing methods and as such representations can be formed using knowledge from these different aspects of language. This proposes the hypothesis that another aspect of language may influence grammatical generalisation. The second consideration is based on the idea that if language representations are processed in similar ways to other areas of knowledge, it is also reasonable to assume that factors that affect other areas of knowledge may also impact grammatical generalisation. To explore these two considerations as potential influences on grammatical generalisation, this review will focus on a connectionist, memory-based statistical learning theory referred to as 'Chunk and Pass' (Christiansen & Chater, 2015; Isbilen & Christiansen, 2020).

1.2.1.1 *Language knowledge as part of general systems of memory – Chunk and Pass.*

Chunk and Pass (CaP) is a memory-based theory of the mechanisms underpinning language processing. It specifically considers how the context in which language is processed shapes its development and use. While the authors of this theory describe it in the context of language, they argue that the same mechanisms also operate in other areas of cognition (Christiansen & Chater, 2015). In the CaP model, language is just another 'perception and action' process, no different from other perceptuo-motor processes.

The CaP theory as proposed by Christiansen & Chater (2015) starts from the premise that how language is processed is a result of overcoming a general constraint on processing – the 'Now or Never Bottleneck'. Speech or other sensory input comes in rapidly and needs to be processed rapidly to overcome memory constraints such as interference from new input and forgetting, as well as the limits of the human sensory systems. So, it needs to be processed 'now' (or at least very rapidly) or it never will be. The theory's authors propose a three-stage processing strategy to overcome the 'Now or Never Bottleneck': eager processing, multiple levels of representations, and anticipation.

CaP processing *eagerly* recodes the perceptual input into more manageable, compressed representations or 'chunks' to work within the limits of memory capacity. To avoid interference from new input, these initial chunks need to be increasingly abstracted into larger chunks for processing to take place over longer time periods. This creates *multiple levels of linguistic representation*, with chunks being 'passed' through these multiple levels to avoid interference and allow for successful processing within the 'Now or Never Bottleneck' constraint. For example, when comprehending spoken language, the language system may initially encode speech into phoneme chunks, which then pass to the next level to form a larger word chunk, which then passes to the next level of a phrase chunk and so on.

Language production follows the same process in reverse, starting at high levels of linguistic representations and passing down to lower levels. Information is passed through the levels as quickly as possible following a 'just in time' principle. What is pertinent to note here and will be of importance later in this review, is that within this system, grammatical or syntactic knowledge is not contained within chunks. Rather, grammatical knowledge is part of how language knowledge is chunked. Using statistical learning examples to help explain this in terms of language comprehension, at a lower level of linguistic representation the transitional probabilities that help to detect and build word boundary knowledge are not stored in a chunk, they form the chunk that phonemes are compressed and abstracted into for further processing. At a higher level of linguistic representation, distributional regularities indicating grammatical categories are not stored in a chunk, they are part of how words and morphemes are compressed and abstracted into chunks for later processing.

The third key element in the CaP model is *anticipation,* which incorporates experience and how statistical learning uses this experience into the model. As the language system experiences more language processing, it strengthens and develops representations/chunks at multiple levels. These well-used, easily accessible representations can then influence and constrain the initial, eager re-coding of the perceptual input. As well as influencing and constraining how this input is chunked and passed through the multiple linguistic levels of processing, whether that's passing from lower to high levels for comprehending or passing from higher to lower levels for production. The influence and constraints the well-used representations from previous language processing have on current language processing, allows the language system to predict or *anticipate* what will come next, helping to speed up CaP processing. For example, if a person encounters the word '*is*' they can anticipate a verb will follow with an '*ing*' suffix from previous encounters with verbs and can start to process this or be ready to process this information before it is encountered.

Thus, statistical learning allows for past language processing, or in other words language experience, to build up language representations by using regularities within a language, to re-encode and abstract language input into chunks for processing. Whether that be transitional probabilities to form 'word' chunks from phonemes or distributional and phonological regularities to form grammatical category chunks for processing meaning. The regularities could be seen as the connections within a representation to help form the chunks here. In turn these regularities, learnt through the statistical learning mechanism, can then support *anticipation* or prediction during language processing to help speed up the CaP process (Christiansen & Chater, 2015). The regularities learnt through statistical learning may not be the only knowledge at play here in forming chunks, with other knowledge learnt through other mechanisms also potentially having an input as well. However, the knowledge learnt through statistical learning is proposed as being an important part in aiding prediction within the language system.

In line with the connectionist approach it grew out of, the CaP model proposes that language knowledge is built through the interactions between the neural learning systems (e.g., memory, sensory processing, statistical learning) with prior and current language experience (processing experience). As the system gains experience, well-used network connections are strengthened and in turn influence the formation and use of new representations/chunks. Language knowledge is thus equated with language processing itself, enabling the language user to process language input and output in real time, under the 'Now or Never Bottleneck' constraint (Christiansen & Chater, 2015). This 'processing in support of action/behaviour' idea aligns well with a connectionist perspective where knowledge is networks of neural connections forming representations that support behaviour (Elman et al., 1998). The comprehension and production of grammatical structure are therefore better conceptualised as a manifestation of processing history, rather than as separate representations.

As the authors state 'it is a processing trace of the operations used to create or interpret a sentence' (Christiansen & Chater, 2015).

This current review is using the CaP model as a connectionist framework from which to explore potential influencing factors on grammatical generalisation and is why it has been described in detail here. However, it does need to be acknowledged that there is debate as to whether statistical learning is the result of chunking or statistical computations (Perruchet & Pacton, 2006) with a recent review suggesting stronger support for chunking (Perruchet, 2019). Whilst alternative views need to be acknowledged here, for the purposes of this review the CaP theory can provide an adequate model from which to understand connectivist principles for exploring potential factors contributing to grammatical generalisation.

1.2.1.2 CaP and grammatical generalisation.

How grammatical generalisation occurs within the CaP model of language processing can be considered in conjunction with Goldberg's (2005, 2009) proposals for linguistic generalisation. Goldberg (2005, 2009) proposes that all learning is item-based, so in a language context a phoneme or word could be an item and generalisations occur as a language learner experiences more instances of these items being used. Generalisation occurs as the experience of items increases, as this repetition allows for similarities between items to become more salient. This in turn allows for abstraction of those similarities which can then be generalised and used for newly encountered but similar items.

Within the CaP model, this would translate to past processing experiences building connections between items (e.g., phonemes or words) based on similarities between these items that produce stronger, well-used connections as these similarities are experienced more often. We can use the novel noun '*wug*' and noun plurals as an example to help explain this. When a language learner encounters more and more examples of plural nouns (e.g., dogs, cats, gnomes) the similarity between all these nouns, the '*s*', will be the element here that is repeated and experienced the most. The higher repetition of the '*s*' helps to build well-used, stronger connection(s) for this grammatical regularity. These stronger connections would then allow for this similarity to become more salient giving it the chance to be abstracted and then used as part of the re-encoding and abstraction process of 'chunking and passing' through the multiple levels of linguistic knowledge. Thus, once a similarity of this nature has been processed sufficiently to form a strong enough connection, it can be generalised to support the processing of a novel item as it can now be used to form chunks for this process. So, if a language user encounters the novel noun '*wug*' and then sees several of them, they can generalise the plural '*s*' from previous noun encounters to successfully processes this novel noun to produce '*wugs*'.

Considered in this way, the CaP model has implications for how grammatical generalisation may occur. The CaP theory supports findings within the statistical learning literature of the patterned nature of grammar being utilised by statistical learning to build language knowledge. Here statistical learning is used to create representations from the connections built through the higher exposure similarities between items provide. However, what is important here is that while grammatical knowledge is part of language knowledge, its main role is in supporting the use of other linguistic item knowledge, such as phonemes and words through how they are chunked and thus processed by the language system. From this perspective, grammatical knowledge is not represented in and of itself within the system as highlighted earlier. This element is key to the current grammatical generalisation considerations and echoes hypotheses made earlier when the connectionist perspective was considered.

This aspect of the CaP theory suggests grammatical knowledge is not separate from other areas of language knowledge; rather, they work together to form one processing system operating within the confines of sensory processing and memory capacity. Thus, it is reasonable to hypothesise that other aspects of language knowledge may influence grammatical generalisation. This hypothesis will be examined in regards to the relationship between vocabulary and grammar, drawing on existing lexicalist theories where both grammar and vocabulary are argued to be stored together in the lexicon rather than separately (Bates & Goodman, 1997; MacDonald et al., 1994).

1.2.2 Grammar knowledge as word knowledge: The potential impact of vocabulary on grammatical generalisation.

Lexicalist theories specifically consider the interplay between grammatical and vocabulary knowledge. These theories propose that both grammatical and vocabulary knowledge are part and parcel of the same memory store, the lexicon. They argue against language as a modular system, where different aspects such as grammar and vocabulary are stored and processed separately and differently (Bates & Goodman, 1997; MacDonald et al., 1994). These proposals align with the CaP model (Christiansen & Chater, 2015) as well as the domain-general principles of the connectionist approach more generally. Consistent with these ideas, Bates and Goodman (1997) presented evidence that demonstrated a relationship between vocabulary and grammatical knowledge during development. They came to this conclusion through considering evidence from language development in both typical and atypical populations, language breakdown, and real-time language processing in typically developing adults. For example, Bates and Goodman (1997) reported on a longitudinal study which observed children from the ages of 10 to 28 months, within which the concurrent development and the predictive relationship between vocabulary and grammatical development between 20 and 28 months of age was considered. Here, the results found that the best predictor of grammatical ability at 28 months; was not grammatical ability at 20 months but vocabulary size at 20 months.

Taken together with findings from the other areas of consideration, Bates and Goodman (1997) hypothesised that certain grammatical developments could not occur until a critical mass of vocabulary had been reached. This proposal is additionally consistent with findings from constructivist frameworks within linguistics discussed in section 1.3.1.3 (e.g., Goldberg, 2005, 2009) and related usage-based approaches in language acquisition (e.g., Lieven et al., 1997; Tomasello, 2000). While it has not yet been directly tested within a statistical learning framework, there is some indirect, empirical evidence that is consistent with this linguistic literature and supports the hypothesis put forward by Bates and Goodman (1997). Firstly, the relationships found between standardised vocabulary measures and statistical cue use for grammatical categories by Lany and Saffran (2011) and Hall et al. (2018) also supports Bates and Goodman's (1997) hypothesis. As detailed in section 1.2.3.3, Lany and Saffran (2011) found that higher vocabulary ability was associated with distributional cue use and lower ability with phonological cue use. Hall et al. (2018) found that older children's generalisation performance using distributional cues could be predicted from their raw vocabulary score. While not a direct test, these findings within the statistical learning literature support the proposed lexicalist link between vocabulary and grammar.

Stronger support, although still not a direct assessment of this link, comes from Gómez (2002). As introduced in section 1.2.3.2, Gómez (2002) manipulated the variability of trained stimuli by creating conditions with varied training vocabulary sizes in an artificial language containing phonological and distributional cues to two grammatical categories. The testing was focused on assessing the learning of the trained grammatical regularities. During testing, while adults were able to correctly endorse trained items to a high level in all vocabulary size conditions, it was only in the highest size where adults were more able to reject incorrect test items (a test of generalisation). A similar finding was shown in infants (18 months) in that a novelty preference between trained correct and incorrect items (measured through a head-turn preference procedure) was only found in the largest vocabulary size condition. Since frequency was kept consistent between conditions, Gómez (2002) concluded that increased variability (implemented via vocabulary size) aids in the detection and learning of statistical regularities that implement the grammatical structure. Gómez (2002) suggested that highly variable contexts in a large training vocabulary may make the stable elements/statistical cues to grammatical structure in the input more salient, and thus easier to identify. This enhancement of learning by increasing variability result has been replicated in infants of different ages (15-18 months; Gómez & Maye, 2009), children (5 years; Wonnacott et al., 2012), in typically developed adults and those with language-based learning disabilities (although results are mixed here; Grunow et al., 2006; von Koss Torkildsen et al., 2012).

Considered in the context of the CaP model, and related lexicalist connectionist models (see section 1.3.1.3), as participants are exposed to the new vocabulary items, network connections and thus representations of these items are being formed and the system learns to process them. Increasing the frequency of exposure would strengthen all connections related to the item's representation, including the regularities that are related to the grammatical structure. This provides processing experience of this more abstract structure in addition to the item-based knowledge, and thus enables the learning of this structure. Additionally, Gómez's (2002) findings suggest that, beyond exposure, variability introduced by increasing the training vocabulary size can help support and even enhance this learning process. Under large vocabulary size/high variability conditions the connections associated with less variable statistical structure are given more opportunities to strengthen. In other words, these conditions allow for connections that encode the statistical structure to become more salient, as the system is given more opportunity to experience how to process them.

However, recent findings from Brown et al. (2022) did not show improved generalisation when variability was increase, in a semi-artificial language which incorporated both phonological and semantic regularities. This was for both adult and child participants within a fully consistent language which did not contain exception words and within a partially consistent language which did contain 'exception to the rule' words. Although, generalisation was improved for exception words within the partially-consistent language in the more highly variable training conditions. Despite this, the overall lack of a variability effect, may be due to the more complex nature of the stimuli compared to the stimuli used by Gomez (2002), which does not incorporate semantic regularities. This could mean that within statistical learning studies that incorporate semantic regularities, a variability effect of this kind may not be found. Brown et al. (2022) also suggest the variability effect may be more relevant for implicit grammatical learning and generalisation as their successful grammatical generalisation findings were driven by explicit grammatical knowledge use. This suggestion leads nicely onto this theses' second, overarching considering regarding the contribution of implicit and explicit knowledge to grammatical generalisation within a statistical learning framework.

1.2.3 Contributions of implicit and explicit learning to statistical learning.

One of the core assumptions in the statistical learning literature is that statistical learning is an implicit process (e.g., Aslin & Newport, 2012; for a review also see Batterink et al., 2015; Franco et al., 2011). This has been a key assumption in the field as unlike explicit learning, implicit learning emerges and matures early in development (e.g., Finn et al., 2016). The idea that statistical learning relies on implicit processes and implicit knowledge is consistent with the CaP approach and the connectionist perspective more broadly, in that language knowledge is implicitly represented within neural networks that support language use (Christiansen & Chater, 2015). In this view, grammatical structures are not explicitly represented, but rather emerge as regularities arising through functional use of language networks. This description of grammatical structure as implicit knowledge is supported by general theories of implicit and explicit learning (Dienes & Perner, 1999; A. S. Reber, 1989, 1993), with implicit knowledge being explicitly linked to connectionist networks in these accounts (Cleeremans, 1997; Dienes & Perner, 1999; Dienes & Perner, 1996). In line with this, a direct parallel between the statistical learning and implicit learning literature has been recently made (e.g., Christiansen, 2019), and recent empirical studies have started to examine the contribution of both implicit and explicit knowledge in statistical learning paradigms (e.g., Batterink et al., 2015; Franco et al., 2011; Monaghan et al., 2019).

1.2.3.1 Linking implicit and statistical learning – questioning the implicit assumption.

Both the statistical and implicit learning literatures assume learning without conscious awareness, both involve methodologies that require passive exposure to stimuli with hidden patterns and both are assumed to represent a domain-general process. These similarities have led some to consider both statistical and implicit learning as being two sides of the same coin (Batterink et al., 2015; Christiansen, 2019; Conway, 2020; Conway & Christiansen, 2005; Franco et al., 2011; Monaghan et al., 2019; Perruchet & Pacton, 2006), with a recent review of the neural basis for implicit and statistical learning showing similar, domain-general neural correlates (Batterink et al., 2019). This link with implicit learning has strengthened the understanding of statistical learning and its domain-general underpinnings, however it has also opened statistical learning up to the same issues and limitations that the implicit learning literature has needed to grapple with.

From its inception, the implicit learning literature has focused on the question of whether subjects are consciously aware or unaware of underlying patterns, and thus whether implicit learning tasks only induce implicit representations (or knowledge) of these patterns (Batterink et al., 2015; A. S. Reber, 1989, 1993). This immediate consideration of awareness is due to this literature's use of adult participants from the beginning, participants who can bring meta-cognitive abilities to the task at hand. Given the similarities between statistical and implicit learning, these questions are now being directed towards statistical learning research, particularly when conducted with adult participants. There are only a handful of studies that directly test the implicit assumptions in statistical learning. Two of the few studies that do investigate this are by Franco et al. (2011) and Batterink et al. (2015); both studies aimed to directly test how implicit statistical learning is in adults within the traditional word-boundary paradigm.

Franco et al. (2011) used a process-dissociation procedure to assess the use of explicit knowledge of transitional probabilities to detect boundaries between artificial words in continuous speech streams. In two experiments the authors demonstrated that adult participants were able to find word boundaries from two different continuous artificial syllable streams both in isolation and then after successive exposure to both artificial languages. In a third experiment, a different group of adult participants were exposed to both languages and were then required to complete an inclusion task and exclusion task (the process-dissociation procedure). In both the inclusion and exclusion tasks, the participants were presented with trisyllabic 'word' strings from both languages and novel, untrained trisyllabic strings. In the inclusion task the participants were asked to respond with 'yes' if the syllable string was from either artificial language or 'no' if it was a new 'word'. This is an old/new recognition task that can be answered using both unconscious familiarity (implicit knowledge) and conscious recollection (explicit knowledge) judgements (Franco et al., 2011; Jacoby, 1991).

The exclusion task however, required the participants to say 'yes' if the syllable string was from the first language they heard and 'no' if it was from the second language or a novel item. This requires participants to endorse 'words' from only one of the languages and thus, feelings of familiarity may impede success in this task as conscious recollection is needed to ensure a correct judgement. It assesses a participant's ability to assert voluntary control over statistically/implicitly learned items which is judged to only be possible if there is conscious (explicit) knowledge (Franco et al., 2011; Jacoby, 1991). The authors found that only participants who had successfully learnt the words from the two artificial speech streams were successful in the exclusion task, demonstrating voluntary control. Participants who had not successfully learnt the words could not do this. This result suggests that explicit knowledge is at least partially involved in statistical learning tasks with adults.

Batterink et al. (2015) used a range of measures in a similar word boundary task to investigate the role of both implicit and explicit knowledge. After exposure to a speech stream (the same as that used by Saffran, Aslin, et al., 1996), adult participants' word boundary knowledge was tested using two tasks: a recognition task using a traditional, statistical learning two-alternative forced choice design and a speeded target detection task. Explicit and implicit knowledge use was assessed in three different ways, using these two tasks across two experiments. First, in Experiment 1 participants were asked to give an overall confidence judgement on their knowledge after they undertook the recognition task and an integrated 'remember', 'familiar' and 'guess' judgement during this task in Experiment 2. Secondly, the speeded target detection task gave reaction time measures, designed to measure implicit knowledge through capturing 'in-the-moment', online processing of the statistical information for the word boundaries. Lastly, ERP measures were taken during both the recognition and target detection tasks. Specifically, the author's used a P300 effect which can indicate how predictable target stimuli are.

Using these different measures Batterink et al. (2015) found that reaction time measures from the target detection task were a more robust measure of implicit, processing-based knowledge gained through statistical learning than performance on the traditional recognition task. This is because reaction time and not recognition performance correlated with the ERP, P300 effect. Supporting this, the authors found that higher confidence ratings in Experiment 1 and higher 'remember' ratings in Experiment 2 correlated with accuracy in the recognition task, indicating the use of explicit knowledge (Dienes & Perner, 1999; Norman & Price, 2015; Rebuschat, 2013).

Based on these findings the authors concluded that both implicit and explicit representations (knowledge) can develop during statistical learning tasks, with performance on traditional, recognition-based tasks being a better reflection of explicit than implicit knowledge. These results correspond with similar findings in the traditional implicit learning literature, discussed in detail in Chapters 3 and 4 (e.g., P. J. Reber & Squire, 1994; Sanchez & Reber, 2013), supporting the proposed connection between statistical and implicit learning (e.g., Batterink et al., 2019; Christiansen, 2019; Conway, 2020; Conway & Christiansen, 2005; Franco et al., 2011; Monaghan et al., 2019; Perruchet & Pacton, 2006; P. J. Reber et al., 2019). It is important to note that these findings do not show that statistical learning cannot be implicit. Instead, they show that explicit knowledge can develop unprompted during statistical learning in adults and highlights the possible differing roles of implicit and explicit knowledge in different aspects of adult language learning and use.

1.2.3.2 Implicit and explicit representations, parallel but separate?

The findings from the Franco et al.'s (2011) and Batterink et al.'s (2015) studies suggest that in adults, statistical learning tasks may generate not only implicit representations, but that explicit ones could emerge as well. This might suggest that there is potential for explicit grammatical knowledge to be involved in adult grammatical generalisation when investigated through a statistical learning paradigm. The emergence of explicit knowledge may not play a significant role in the early development of children's language systems, which likely fully relies on implicit learning. However, in adults, explicit, meta-cognitive, non-processing-based representations of grammar could also develop alongside this, which could be accessed when knowledge is being tested (depending on how it is being tested) or even support other aspects of language acquisition. This is what Monaghan et al. (2019) demonstrated in their study.

The study used an artificial language which incorporated nouns for different shapes and verbs for different shape movements, with nouns being indicated by the particle *'tha'* and verbs by *'noo'*. The artificial language was constructed into phrases (e.g., *'Tha trepier noo vinnoy'* translates as 'the rectangle moves in a circle') and was introduced to the participants in a cross-situational learning task. In this task participants were presented with two moving shapes along with one phrase and asked to judge which moving shape was described by the phrase. The participants were then tested for verb and noun vocabulary knowledge using two further two-alternative forced choice tasks. To test verb vocabulary, participants were presented with just a trained verb word (no particle) and two moving novel shapes to choose from. The noun task displayed two motionless trained shapes with an accompanying noun word (again, with no particle). Using the same language, training, and testing tasks across two experiments Monaghan et al. (2019) compared performance between a rule-aware and unaware group and investigated the emergence and use of explicit knowledge of the grammatical regularities across training.

In Experiment 1, two groups of participants underwent the same training and testing protocols, but one group was told the grammatical 'rule' and the other was not. The authors found that the rule-aware group outperformed the unaware across the cross-situational training task blocks. Retrospective verbal reports of explicit knowledge of the rule obtained after the study was completed demonstrated that half of participants in the unaware group became aware of the grammatical ‘rule’. The participants with ‘emergent’ awareness also outperformed the participants who remained unaware across training blocks. However, across all participants, performance in the vocabulary testing tasks was equivalent indicating that support from explicit grammatical knowledge was most important during learning.

Experiment 2 explored emergent explicit knowledge in more depth by looking at a third ‘unaware’ group which underwent the same training and testing tasks. However, this time after each trial in the cross-situational training task, participants were asked to give a grammatical rule awareness rating of the language from four options: guess, intuition, recollection, and rule knowledge. Using these judgments, the authors found a change as the learning progressed from guess/intuition responses to recollection/rule knowledge responses, indicating emergence of explicit grammatical knowledge. Of particular interest is the finding that reported explicit knowledge in earlier items during training predicted accurate performance on similar items in a later training block. This suggests a supportive role of explicit grammatical knowledge on vocabulary during acquisition.

Linguistic research considering grammatical generalisation using semi-artificial languages, first introduced in section 1.1.3.4, can also support Monaghan et al.’s (2019) findings. Although they consider the role of explicit knowledge on grammatical generalisation more directly. With the exception of using real words for nouns and/or verbs, the construction of these languages is similar to the one used by Monaghan et al. (2019), but with maybe only including phonological or distribution regularities with a semantic regularity. Williams (2005), Ferman and Karni (2010), Ferman et al., (2009) and Brown et al., (2022) all consider the role of explicit and implicit knowledge when generalising semantic and distribution/phonological regularities using retrospective verbal reports. Although, these authors mainly focus their verbal reports measure on the semantic regularity. While Williams’ (2005) found that a small number of adult participants reported explicit knowledge of the regularities, most did not report explicit knowledge and as a group the participants still demonstrated successful generalisation. Ferman and Karni (2010) found successful generalisation in adults, but with almost all adult participants (7 out of 8) reporting explicit knowledge of the regularities. Ferman et al. (2009) went further, conducting a verbal report measure at the end of every training and testing session they ran (10) and found that generalisation accuracy sharply increased after explicit knowledge of the semantic regularity was reported. Brown et al. (2022) found that almost all of their adult participants (19 out of 20) reported explicit knowledge of the semantic regularity in the their fully consistent language and about half (10 out of 20) in their partially consistent language. In both languages’ adults as a group demonstrated successful generalisation, however Brown et al. (2022) found that this finding was driven by explicitly aware participants, as when these participants were removed from the analysis, the unaware participants did not demonstrate successful generalisation.

With the exception of Williams (2005), these studies focused on explicit knowledge of the semantic regularity and conclude that when semantics are present explicit knowledge of this is a part of grammatical learning and generalisation. This supports and extends conclusions from Monaghan et al. (2019) showing the explicit knowledge can not only support vocabulary learning, but also grammatical regularity learning and generalisation. These findings suggest that grammatical knowledge can develops outside of the implicit, processing-based language system and thus, may not necessarily support real-time language processing. Considered in terms of the CaP model, implicit grammatical knowledge supports real-time language use through embedded grammatical knowledge supporting the chunking and passing of linguistic information through multiple levels of abstraction. Implicit grammatical knowledge supports fast processing here because it is part of the way information is chunked, rather than being contained within a chunk itself. If the explicit grammatical knowledge shown to develop and support accuracy in Monaghan et al. (2019) and grammatical generalisation in the semi-artificial language studies (Williams, 2005; Ferman & Karni, 2010; Ferman et al. 2009; Brown et al., 2022) is stored outside of this language system, it cannot be used by the system to support the chunk and pass process in real-time. Evidence supporting this idea has been found within the implicit learning literature when participants are explicitly told to look for an underlying pattern and this hinders their performance (e.g., A. S. Reber, 1976) and is indirectly supported by findings introduced earlier from Lew-Williams & Fernald (2010).

Lew-Williams & Fernald (2010) found that both adult and child, native Spanish speakers were able to take advantage of grammatical gender cues (distributional statistical cues, e.g., *el* – singular masculine, *la* – singular feminine) when learning novel Spanish nouns. This was measured through eye-tracking and reaction times, measures able to capture implicit, processing-based knowledge use (Kelley & Lindsay, 1996; Rebuschat, 2013; Timmermans & Cleeremans, 2015; see also Chapter 4 of this thesis). However, as discussed earlier, when native Spanish speakers were compared with second-language speakers their reaction times were significantly faster (Lew-Williams & Fernald, 2010). This suggests that the second-language speakers were less able to take advantage of the grammatical gender cues in real-time, despite showing a similar level of accuracy to native speakers as well as being able to explicitly explain the grammatical gender cue. Considering this result with the CaP model, it may be due to second-language learners having less exposure to these cues than native speakers as well as exposure starting later in their childhood/teenage years. These participants may not have experienced yet, the exposure needed to form the connections that help embed implicit grammatical knowledge for efficient use within a language system. Thus, while they have explicit knowledge that can be accessed and used when knowledge is being tested and so can produce high accuracy performance, it is not as effective at supporting fast, real-time processing.

This is pertinent for findings for statistical learning, as while native speakers in this study are adults, the adult participants in statistical learning studies are more akin to the second language learners here. This is because the majority of statistical learning studies use artificial languages and as such the adult participants in these studies are encountering and learning a new language like when encountering a second language, albeit an artificial one. These findings from Lew-Williams and Fernald (2010) support the idea of separate, meta-cognitive explicit grammatical knowledge and implicit, processing-based grammatical knowledge, with only the latter used in real-time language processing (see also Tamminen et al., 2012, 2015 for related findings).

Despite these overall findings and their support for the involvement of explicit knowledge within grammatical generalisation, some interpretation caution is needed. The semi-artificial language studies use real nouns and/or verb which could potentially prompt a more explicit learning approach by participants. Prompting adults to use a more problem-solving based techniques to complete the task (Moroshkina et al., 2019; Brown et al., 2022) and increasing the likelihood of explicit knowledge emerging. The explicit knowledge prompt here may have more to do with the presence of semantic information rather than from the use of real words though, as Monaghan et al. (2019) used a fully artificial language incorporating semantic regularities and also found a role for explicit knowledge. Furthermore, Williams (2005) did find that adults can generalise semantic regularities when they did not report explicit knowledge. Suggesting that generalising semantic regularities may also be able to occur implicitly, a point acknowledged by Brown et al. (2022) when considering natural language acquisition and learning. These findings suggest that firm conclusions in this regard are currently hard to make, which means potential confounding influences on current findings needs to be considered.

Additionally, the semi-artificial language studies depend on retrospective verbal reports as their measure of explicit knowledge. While this measure can be informative, it assumes that explicit knowledge is always verbalizable which may not be the case (see Dienes & Perner, 1999; Karmiloff-Smith, 1986 and Chapter 3, section 3.31 of this thesis). This is particularly an issue when the measure is used with children as will be discussed further in the next section (1.2.3.3). Thus, the assumption, particularly made by Williams (2005), that a lack of verbalisation of the grammatical regularity means knowledge is implicit is problematic. Implicit knowledge is thus being assumed in the absence of explicit knowledge and not being directly tested. Monaghan et al.'s (2019) also use verbal reports across both of their reported experiments, however they do use a more sensitive, confidence-based measure of explicit knowledge within their second experiment; the guess, intuition, recollection, and rule knowledge option given after each item presentation (see Chapter 3, section 3.2.2 and 3.2.3 of this thesis for further details). The findings from this measure do support findings from Ferman and Karni (2010), Ferman et al. (2009) and Brown et al. (2022) of a role of explicit knowledge, strengthening these results and conclusions. However, the confidence rating measure used by Monaghan et al. (2019) can still be argued to aid the emergence of explicit knowledge due to its reference to 'rule knowledge' in the response option. This could potential prompt participants to the presence of a rule potentially causing them to adopt a more explicit, rule-search approach to the task (Moroshkina et al., 2019 and see Chapter 3, section 3.2.3 of this thesis).

These limitations still do not stop these studies from providing relevant and informative finding for language learning and processing in adults, particularly when it comes to second language learning. They can also still inform hypotheses within statistical learning studies, as they can be similar to second language learning contexts, as discussed earlier in relation to Lew-Williams and Fernald (2010) findings. However, acknowledging and recognising these details/limitations is of interest to statistical learning research, given the traditional implicit assumption and its focus and historical beginnings within first language acquisition and its attempts to understand how this works within the real-world (e.g., Aslin & Newport, 2012; R. Frost et al., 2019). In language acquisition, the 'rules' or grammatical regularities are rarely directly referred to, or at least not consistently enough or in the amounts needed (e.g., the inconsistency of feedback as discussed in section 1.1.2). It's therefore of interest to the statistical learning literature to consider the role of implicit knowledge and potentially explicit knowledge by using sensitive and direct measures for each which also interfere as little as possible with the development and use of these two knowledge types. This could help to inform theories of the theoretical underpinnings of statistical learning along with aiding the consideration of potential developmental differences here (see the next section 1.2.3.3 for more details here). For instance, whether implicit or explicit knowledge emerges and is used for grammatical generalisation could influence how vocabulary may support or not support grammatical generalisation. This speculation is based on Brown et al.'s (2022) suggestion that the explicit grammatical knowledge driving the successful generalisation found in their study, may have prevented high vocabulary variability conditions during training from support grammatical generalisation. The authors suggest this variability effect may be more relevant to implicit grammatical learning and generalisation.

Despite the discussed links to findings within implicit learning and natural language processing, along with findings from similar paradigms within linguistics, the statistical learning literature so far has only considered the role of explicit knowledge within a word boundary paradigm and explicit grammatical knowledge in vocabulary learning. As such, while Batterink et al. (2015) or Monaghan et al. (2019) findings support predictions for a role of explicit knowledge in grammatical generalisation, they do not provide direct evidence for it. Thus, research directly looking at this would be beneficial.

1.2.3.3 What about children?

So far, the discussed research on the role of explicit knowledge in statistical learning has focused on adults, including the recent reviews in this area (e.g., Batterink et al., 2019; Conway, 2020). Whilst arguably the earliest stages of language development are underpinned by implicit learning, language learning changes across development to adulthood (e.g., Newport, 1990). Statistical learning has traditionally been seen as invariant across age (Raviv & Arnon, 2017; Saffran et al., 1997), but as demonstrated by the studies described above explicit knowledge may emerge in the course of learning and influence performance on language tasks in adults. Thus, it is important to examine whether and how explicit knowledge might emerge in language learning over the course of development. One study within the implicit learning literature has explicitly addressed this question using a standard implicit learning task: Hebb learning.

Hebb learning paradigms involve participants implicitly learning repeating sequences, which are interleaved with non-repeated stimuli (Hebb, 1961). The Hebb paradigm originally used lists of digits, but it has subsequently been tested and used as a model of word learning (e.g., Page & Norris, 2009) and as a test of developmental differences in word learning (Smalle et al., 2016). Smalle et al. (2018) used a Hebb sequence-learning paradigm to examine the relative contribution of explicit and implicit processing in children as compared to adults when learning transitional probabilities for detecting word boundaries.

To compare explicit and implicit learning, Smalle et al. (2018) created two syllable Hebb sequences and before training explicitly told participants the sequence for one and withheld all information about the other (so it needed to be implicitly learnt during training). After each testing session (across three delayed time points; 4 hours, 1 week and 1 year), participants were given a verbal awareness questionnaire to ascertain emergent explicit knowledge of the implicitly learnt Hebb sequence. Both adults and 8-year-old children were tested using the same paradigm. Children displayed better long-term retention for both explicit and implicit Hebb sequences than adults, but the effect was stronger for the implicit Hebb sequence. The verbal awareness questionnaire showed that both adults and children became explicitly aware of the implicit Hebb sequence, but adults became aware earlier than children. This explicit awareness was also only associated with adult performance on the implicit Hebb sequence task, while child explicit awareness was not associated with performance on this task.

This suggests that adults were drawing on both explicit and implicit learning mechanisms during this task (supporting Franco et al., 2011 and Batterink et al., 2015), while children relied on implicit learning (Smalle et al., 2018). While this does support potential developmental differences in the use of implicit and explicit knowledge here, it again is focusing on word-boundary detection. Research from linguistics, considering grammatical generalisation using semi-artificial languages can further support potential development difference in the use of explicit and implicit knowledge. As discussed in section 1.2.3.2, the role of explicit knowledge in grammatical generalisation has been considered in this paradigm with some studies considering children and well as adults in this regard (Ferman & Karni, 2010; Brown et al., 2022). Ferman & Karni (2010) found that eight-year olds, twelve-year olds and adults all showed increased generalisation accuracy across ten sessions of training and testing. However, only the twelve-year old and adult groups reported explicit knowledge of the semantic regularity (and showed greater gains across the sessions). Brown et al. (2022) found that fewer children reported explicit knowledge compared to adults. This was for both the fully consistent semi-artificial language groups (19/20 adults compared to 13/30 children) and the partially consistent language groups (10/20 adults compared to 2/30 children). However, the authors also found that in the fully consistent language child group, their successful generalisation findings were driven by aware participants. When aware participants were removed, group performance was at chance.

Overall, these studies support Smalle et al. (2018) in that adults seem to be better able to build and draw upon explicit knowledge when generalising grammatical regularities, or more specifically here semantic regularities. The absence of reported explicit knowledge in children in both Ferman and Karni (2010) could suggest the use or reliance on implicit knowledge. Particularly where Ferman and Karni (2010) show that eight-year olds do show increases in generalisation accuracy across sessions. A speculation that would be supported by Smalle et al.'s (2018) results. However, as discussed in section 1.2.3.2. these studies do not directly assess the use of implicit knowledge, it's assumed from a lack of verbalisation of explicit knowledge. When verbalising explicit knowledge may not always be possible (see Dienes & Perner, 1999; Karmiloff-Smith, 1986 and Chapter 3, section 3.31 of this thesis) particularly where children are concerned, conclusions around both implicit and explicit knowledge use need to be made with this in mind. Brown et al.'s (2022) finding also suggests that generalisation when semantic regularities are present may be dependent on developing explicit knowledge of the regularity, in both adults and children. This finding suggests that when semantic information is involved, there may not be developmental differences in implicit and explicit knowledge use. Although this finding and prediction needs to be further investigated and replicated (see also discussion in section 1.2.3.2).

Despite these considerations, taken together with Smalle et al.'s (2018) findings it supports the hypotheses of a potential development difference in the use of explicit and implicit knowledge for grammatical generalisation within statistical learning paradigms. So far, all the studies that consider the role of explicit and/or implicit knowledge within statistical learning only consider adult participants. The majority of which focus on word-boundary detection (e.g., Batterink et al., 2015; Franco et al., 2011) or on how explicit grammatical knowledge supports vocabulary learning (e.g., Monaghan et al., 2019), not grammatical generalisation. This highlights a clear gap in the statistical learning literature which would be worthwhile investigating. Particularly given the discussed potential implications of the role for implicit and explicit knowledge on understanding both grammatical generalisation and statistical learning mechanisms.

1.3 Summary and Conclusions

The statistical learning literature to date has demonstrated that both children and adults are able to learn grammar from statistical regularities that exist in the linguistic environment. Additionally, it has demonstrated *what* was being generalised when this same knowledge is used in new situations: the same statistical regularities. By considering the connectionist perspective (Elman et al., 1998) and by operationalising it with the CaP model (Christiansen & Chater, 2015) this review has identified two potential areas to start investigating the *how* of grammatical generalisation within a statistical learning framework. The domain-general hypothesis of connectionist theories suggests that language is processed in the same way as other areas of cognition but also that different aspects of language knowledge are not stored separately. This is operationalised by the CaP theory demonstrating domain-general mechanisms for processing language as well as showing how different aspects of language knowledge are bound together to support language processing.

This has helped to form two hypotheses. The first hypothesis which directly stems from the CaP model, is that other aspects of language knowledge may interact with grammatical knowledge and specifically for this review, influence grammatical generalisation. A second hypothesis is formed as a result of the implicit assumptions inherent within connectionist theories such as CaP, which do not consider the contribution of explicit knowledge, which again for this review is considered for grammatical generalisation. The first hypothesis was refined to consider the role of vocabulary on grammatical generalisation, based on the connectionist-based, lexicalist theory. This theory proposes that grammar and vocabulary are stored together in the lexicon rather than separately, with Bates and Goodman (1997) proposing a critical mass of vocabulary knowledge needed before grammatical knowledge can develop during language acquisition (see also related theoretical ideas by Goldberg, 2005, 2009). There is currently a lack of direct evidence within a statistical learning framework to support a relationship between vocabulary and grammatical generalisation. However, findings related to word variability in grammatical category learning (Gómez, 2002) and explicit grammatical knowledge aiding vocabulary learning (Monaghan et al., 2019) support the hypothesis that vocabulary may influence grammatical generalisation within a statistical learning framework.

There are two main aspects of vocabulary which could interact with grammatical generalisation, namely vocabulary breadth and vocabulary depth (Schmitt, 2014). Vocabulary size or breadth considers the number of words in the lexicon, so considers surface level knowledge of words where recognition may or may not be coupled with meaning knowledge (Qian & Schedl, 2004; Schmitt, 2014). Vocabulary depth on the other hand, considers detailed knowledge of vocabulary meaning, semantic context, pronunciation and other lexical properties (Qian & Schedl, 2004; Schmitt, 2014). Research on vocabulary's role in grammatical generalisation within a statistical learning framework is in its infancy (e.g. Gomez, 2002; Brown et al., 2022), which means there is a limit to our current understanding of its role here, if there indeed is a role. Thus, to build from current theoretical and empirical understanding, this thesis will focus on exploring the role of vocabulary size (breadth). The lexicalist, ‘critical mass’ hypothesis (Bates & Goodman, 1997), and constructionist explanations for the variability effect in Gomez (2002; Goldberg, 2005, 2009) are currently mainly considered on concepts of vocabulary size rather than depth. With investigations looking into the variability effect specifically focusing on the size of vocabulary available to participants within the learning context (e.g., Brown et al., 2022; Gomez, 2002). Although it should be acknowledged that the item-level approach to grammatical generalisation put forward by constructionists and constructivists (e.g., Christiansen & Chater, 2015; Elman et al., 1998; Goldberg, 2005, 2009) could easily incorporate vocabulary depth as well as breadth. It is hoped then that the work presented in this thesis could help to support potential future work that considers this.

The second hypothesis was prompted by the recent parallels that have been drawn between statistical and implicit learning (Batterink et al., 2015; Conway, 2020; Conway & Christiansen, 2005; Franco et al., 2011; Monaghan et al., 2019; Perruchet & Pacton, 2006). This has opened statistical learning up to issues that the implicit learning literature have investigated from the beginning, that is questioning the implicit nature of what has been learnt. Connectionist theories such as the CaP model currently only consider the contribution of implicit language processing, but findings from the implicit learning and linguistics literature could suggest an additional contribution from explicit knowledge when considered within a statistical learning framework, at least where adults are concerned. Studies within statistical learning have so far only considered this within a word boundary detection paradigm or have only considered the role of explicit knowledge within grammatical learning and not grammatical generalisation. However, these studies have found the use of explicit knowledge within this paradigm in adults (Batterink et al., 2015; Franco et al., 2011) along with the neural basis for statistical and implicit learning found to activate neural correlates for both implicit and explicit processing (Batterink et al., 2019). While again this is not direct evidence for explicit knowledge use in grammatical generalisation, it supports the hypothesis that explicit knowledge as well as implicit, may influence grammatical generalisation in adults.

A secondary hypothesis related to this, is that this may not be the case in children. Evidence for this comes from a Hebb learning paradigm assessing word boundary detection (Smalle et al., 2018). This study found that children and adults did not utilise implicit and explicit knowledge in the same way, with measures of implicit learning but not explicit processing correlating with grammar processing performance in children. Further support for this idea also comes from studies using semi-artificial language studies investigating grammatical generalisation using different but similar methodologies to statistical learning (Ferman & Karni, 2010; Brown et al., 2022). Overall these studies suggest that adults are more able to develop and use explicit grammatical regularity knowledge (specifically semantic regularities) when learning to generalise. Again, direct studies need to be done within a statistical learning paradigm that specifically assess grammatical generalisation which helps to takes the next steps towards mimicking natural language learning. However, these findings do propose a potential difference between adults and children in terms of the factors that may influence grammatical generalisation within a statistical learning framework.

1.3.1 Thesis overview

This thesis aims to address these hypotheses by considering two overarching questions:

1) Does vocabulary knowledge influence grammatical generalisation?
2) Does explicit as well as implicit grammatical knowledge influence grammatical generalisation?

Following this chapter, is the first empirical chapter, Chapter 2, which focuses on experiments that consider the first research question. To then address the second question, a literature review was conducted to develop a measure of explicit and implicit knowledge that was compatible with the paradigm used in this thesis. This review can be found in Chapter 3. Chapter 4 is the next empirical chapter which uses this developed measure along with a further measure of implicit knowledge, to directly consider the second question of this thesis. The final chapter, Chapter 5, provides an overall discussion of the findings from this thesis related to current theories within statistical learning and recommendations for future research.

Chapter 2. Grammatical generalisation in statistical learning: The role of variability in the learning context and variability in knowledge in adult and child learners.

2.1 Introduction

New words are constantly entering our collective vocabulary. Recent world events have highlighted examples of this with the verb ‘to skype’, which refers to conducting online video phone calls. When new words enter our vocabularies, we are instantly able to incorporate them into daily language use e.g., *‘I’m just skyping with a friend’* or *‘I skyped them yesterday’*. This ability to recognise the properties of new words and use them in grammatically correct ways through grammatical generalisation is a key skill of proficient language use. The use of grammatical regularities in a first language was first examined experimentally in children and adults by Berko (1958) with the now famous *‘wug’* test. Berko (1958) found that English speaking children and adults could generalise the plural *‘s’* regularity to the previously presented novel singular noun *‘wug’*. Generalisation of grammatical knowledge was a key piece of evidence to demonstrate proficient use of English grammar.

Statistical learning is one of the approaches to language learning that has examined extensively the question of how we learn and generalise grammatical knowledge (Aslin & Newport, 2012; Gómez & Gerken, 2000; Lany & Saffran, 2013; Romberg & Saffran, 2013). According to this approach, the ability to generalise in the examples *to skype* and *wugs* is partly underpinned by our knowledge of the properties of the respective grammatical categories (nouns and verbs). Properties of grammatical categories are reflected in many statistical regularities among the words within a category which can reliably cue the type of grammatical category they are. For example, distributional regularities consider the linguistic context that a word sits in. Distributional regularities can come in the form of word co-occurrences (e.g., *<u>the</u> wug* or *<u>to</u> skype*) or when two co-occurring linguistic units frame the word (e.g., *<u>is</u> skyp<u>ing</u>*). Another type of regularity is phonological, and it considers associations with speech sounds that can be placed at the word, syllable, or phoneme level. An example is syllable stress position in English disyllabic words, with nouns tending to stress the first syllable (e.g., a *re*cord) and verbs the second (e.g., to re*cord*).

Corpus analyses of child-directed speech confirm the presence of these and other similar statistical regularities in natural language where they reliably indicate grammatical categories (Kelly, 1992; Mintz, 2003; Monaghan et al., 2005; Redington et al., 1998). It is also well documented that these statistical regularities can be learnt by both children and adults via implicit statistical learning mechanisms to build grammatical category knowledge in both natural and artificial languages (Farmer et al., 2006; R. L. A. Frost et al., 2019; Gómez, 2002; Hall et al., 2018; Lany, 2014; Lany & Saffran, 2010, 2011; Lew-Williams & Fernald, 2007, 2010; Mintz, 2002; Mirkovic et al., 2011). As an example, Gómez (2002) demonstrated that both 18-month-old infants and adults were able to use distributional regularities to distinguish between two artificial grammatical categories. Hall et al. (2018) found that 6–9-year-old children were also able to use distributional regularities in an artificial language to learn novel grammatical categories.

Semantic regularities are less well researched in the context of statistical learning but present a third regularity for grammatical categories. Semantic regularities are based on what words in particular grammatical categories refer to, for example nouns often refer to entities and verbs to actions (e.g., Gentner, 1982, 2006). Our examples illustrate this idea well with the verb 'to *skype*' referring to a physical action (talking over a video call), and the noun *'wug'* referring to a sociable blue creature. Even apparently arbitrary grammatical categories such as grammatical gender (masculine, feminine) contain semantic regularities beyond that of the associated biological sex (Corbett, 1991, 2013; Mirković et al., 2005; Zubin & Köpcke, 1981). The learning of semantic regularities to build grammatical category knowledge in the context of statistical learning is less well examined. The few studies that do examine it demonstrate a supportive relationship between semantic, phonological, and distributional regularities in both infants and adults (R. L. A. Frost et al., 2019; Lany, 2014; Lany & Saffran, 2010, 2011). Adults have also been found to use semantic regularities, along with distributional and phonological regularities to learn grammatical gender-like categories (e.g., Mirkovic et al., 2011; Mirkovic & Gaskell, 2016).

The statistical learning framework is closely related to connectionist theories of language (e.g., Elman et al., 1998; Seidenberg, 1997; Seidenberg & MacDonald, 2018). One of the commonalities between the two approaches is their domain-general view of language learning and use. This domain-general view proposes that different aspects of language are not separately processed, utilising differing mechanisms. Rather, all aspects of language are learnt and processed using common mechanisms, such as statistical learning. One implication of this view is that other areas of language might impact the learning and use of grammatical knowledge. Supporting this idea are related lexicalist theories, which specifically consider the interplay between grammatical and vocabulary knowledge (e.g., Bates & Goodman, 1997; MacDonald et al., 1994; Wonnacott, 2011; Wonnacott et al., 2008; Wonnacott et al., 2012). These theories propose that both grammatical and vocabulary knowledge are part and parcel of the same system, the lexicon.

Bates and Goodman (1997) presented evidence to support a unified representation of lexical and grammatical knowledge within the lexicon, rather than separate modular based representations. They considered evidence from language development in both typical and atypical populations, language breakdown in aphasia, and real-time language processing in typically developed adults. Evidence from these different areas showed strong links between lexical and grammatical knowledge. For example, Bates and Goodman (1997) reported on a longitudinal study which observed children from the ages of 10 to 28 months. The concurrent development and the predictive relationship between vocabulary and grammatical development between 20 and 28 months of age was considered. The results showed that the best estimate of grammatical ability at 28 months was not grammatical ability at 20 months but vocabulary size at 20 months. Taken together with other similar findings from the literature, Bates and Goodman (1997) hypothesised that certain grammatical developments could not occur until a critical mass of vocabulary had been reached. For example, in the acquisition of the English past tense, a well-documented U-shaped curve with initially correct use of irregular past tense forms is followed by a prolonged phase of over regularisations followed by a final phase of correct regular and irregular use. Marchman and Bates (1994) demonstrated that this developmental pattern can be explained by the size and composition of the early vocabulary, starting with an initial period dominated by a small number of irregular forms, followed by a phase where the majority of the verb vocabulary were regular forms. Marchman and Bates (1994) argued that this change in the composition of the early vocabulary contributed to the pattern of the past tense use and the development of this morphological system.

The hypothesised link between vocabulary knowledge and grammar, and specifically a critical mass of vocabulary needed for grammatical development, has not been directly tested within the statistical learning context. However, there is some empirical evidence from the statistical literature that might support it. In an artificial language containing phonological and distributional cues to two grammatical categories, Gómez (2002) manipulated the variability of trained stimuli by creating conditions with varied training set sizes. The language consisted of an *a*X*b* and *c*X*d* type structure, with the *a, b, c* and *d* elements providing the non-adjacent dependencies constructed to cue two grammatical categories, and the X element providing an interchangeable arbitrary stem. It was the number of the arbitrary stem elements during training that Gómez (2002) manipulated, creating different set size conditions that effectively changed the number of vocabulary items the participant experienced in each grammatical category.

During testing, while adults were able to correctly endorse trained items to a high level in all set size conditions, it was only in the highest set size where adults were more able to reject incorrect test items. A similar finding was found in 18-month-old infants in that a novelty preference between trained correct and incorrect items (measured through a head-turn preference procedure) was only found in the largest set size condition. Since frequency was kept consistent between conditions, Gómez (2002) concluded that the increased number of stems leads to a greater variability in the training set which aids in the detection and learning of the regularities (*a...b, c...d*). Gómez (2002) suggested that the condition with the greatest number of stems produces a highly variable context that makes the stable elements in the input more salient, and thus easier to identify (see also Goldberg, 2005).

This result has been replicated in infants of different ages (15-18 months; Gomez & Maye, 2009), children (5 years; Wonnacott et al., 2012), and in typically developed adults and those with language-based learning disabilities (although results are more mixed here; Grunow et al., 2006; von Koss Torkildsen et al., 2012). While these findings do not directly test the role of vocabulary in grammatical learning or generalisation, they are consistent with Bates and Goodman's (1997) lexicalist proposal. Gómez (2002) considers the role of vocabulary in terms of the variability within the learning context or the learner's learning environment. Bates and Goodman's (1997) critical mass hypothesis is consistent with this view in that the learners need to reach a critical mass of vocabulary knowledge before the right amount of variability is present in the lexicon to support the extraction of grammatical regularities.

Research on grammatical generalisation within a constructivist framework in linguistics (e.g., Goldberg, 2005, 2009) and related usage based approaches in language acquisition (e.g., Lieven et al., 1997; Tomasello, 2000) also support the idea of a relationship between vocabulary and grammatical generalisation. For instance, Goldberg (2005, 2009) proposes that language knowledge consists of both item-specific knowledge (e.g., words/vocabulary) and generalisations (e.g., grammatical constructions) and that both need to be considered for a coherent account of language. Based on a constructivist perspective, Goldberg (2005, 2009) proposes that all learning is item-based, and generalisation occurs as learners experience more instances of item usage. This is due to item instances allowing for similarities between items to become more salient until they are salient enough for them to be abstracted and generalised. Thus, Goldberg's (2005, 2009) proposal is also consistent with the idea of a supportive relationship between vocabulary and grammatical generalisation within a statistical learning framework.

Based on the literature reviewed above, the current study aims to directly assess the role of vocabulary in grammatical learning and generalisation within a statistical learning framework. Vocabulary knowledge can involve both vocabulary size/breadth (the number of words in the lexicon) and vocabulary depth (detailed knowledge of meaning, pronunciation and other lexical details; Qian & Schedl, 2004; Schmitt, 2014). While both aspects of vocabulary may potentially play a role in grammatical generalisation, this study focuses on Vocabulary size (sometimes referred to as item-level knowledge). This is so the current study can be grounded in current theoretical models, findings and understanding within this area. For instance, the lexicalist, 'critical mass' hypothesis (Bates & Goodman, 1997), and constructionist explanations for the variability effect in Gomez (2002; Goldberg, 2005, 2009) are based on concepts of vocabulary size rather than depth. It is hoped that this work can help to build a foundation upon which further research can better consider other aspects of vocabulary such as depth.

To explore the role of vocabulary size, we use a paradigm that incorporates regularities in the distributional, phonological, and semantic domains simultaneously (Mirkovic et al., 2011; Mirkovic et al., 2021; Mirkovic & Gaskell, 2016). This paradigm has advantages over typical statistical paradigms, not only in that it uses a more realistic learning environment (R. Frost et al., 2019), but it also allows us to examine the interplay between vocabulary and grammar learning for different types of regularities. We examined the role of vocabulary in grammar learning from two perspectives. In Experiment 1, similar to the Gómez (2002) study with distributional regularities, we examined the role of vocabulary by manipulating the size of the training set (variability in learning context), but we used a more complex artificial language comprising regularities in multiple domains. Here we tested child and adult learners. In Experiment 2, we used the same artificial language as in Experiment 1 and we kept the size of the vocabulary constant, but we manipulated the level of vocabulary knowledge achieved by using a criterion learning paradigm (variability in knowledge). Here we only tested adult learners.

We used an artificial language that mimics grammatical gender systems found in natural languages. Grammatical genders are typically associated with regularities in the semantic, phonological, and distributional domain (Corbett, 1991, 2013; Mirković et al., 2005; Zubin & Köpcke, 1981). To create semantic regularities, two semantic categories were used in our artificial language: animals and artefacts. Phonological regularities were incorporated using a "suffix" that was attached to a "stem" (e.g., *-eem:* mof<u>eem</u>). Finally, distributional regularities were incorporated as a co-occurring "determiner" and "suffix" (e.g., <u>*tib*</u> mof<u>*eem*</u>; see Table 1 for examples). Each co-occurring determiner and suffix were paired with a semantic category (animals or artefacts) through the use of a picture referent. This provided an *a*X*b* structure for animals and *c*Y*d* structure for artefacts, with X and Y denoting the interleaving arbitrary stem, *a* and *c* the determiners and the *b* and *d* the suffixes.

Table 1. Design of noun classes

	Determiner	Suffix	Examples
animal	tib	eem	*tib* mof*eem* = dog
			tib zeap*eem* = duck
artefact	ked	ool	*ked* larsh*ool* = table
			ked snar*ool* = TV

The current study builds on the findings from Mirkovic et al. (2021) who used the artificial language described above with child and adult learners, focusing on the question of cognitive mechanisms underpinning statistical learning. Over three experiments and using different types of training, they showed that child learners consistently achieved both lower levels of vocabulary learning, and lower levels of grammatical generalisations, relative to adult learners. Children only showed successful grammatical generalisation when language production was included at training, and only in the generalisation test that included all three types of regularities (distributional, phonological, semantic). Adults' generalisation performance was also influenced by the changes in the type of training, which was mainly reflected in the changes in generalisation patterns across different types of regularities. However, Mirkovic et al. (2021) did not systematically examine the role of vocabulary size. The current study directly examines this question.

In Experiment 1 in the current study, we directly assessed the role of vocabulary in grammatical generalisation by manipulating the size of the training set (variability in the learning context) while utilising the production effect (e.g., Hopman & MacDonald, 2018; Icht & Mama, 2015; Zamuner et al., 2016). We tested both child and adult learners. We replicated the training and testing procedures from Mirkovic et al. (2021) but increased the number of "stems" and thus the variability of the training items. This enabled direct comparison between results found within this current experiment and those found within Experiment 3 of Mirkovic et al. (2021) to assess the role of training variability. We examined the effect of vocabulary size on generalisation for regularities involving only phonological and distributional cues, and regularities that also included semantic cues. Based on Gómez (2002) it was hypothesised that increasing variability would enhance generalisation performance in both adults and children.

In Experiment 2, we explored the role of vocabulary in grammatical generalisation by manipulating the level of word knowledge achieved at training (variability in word knowledge), and we only tested adult participants here. The same training and testing procedures were used as in Experiment 1, but the training protocol was manipulated to use a criterion learning procedure. This enabled training to stop once participants reached a given level of word knowledge, set to be lower than that of adults and equivalent to children in Experiment 1. Based on the literature reviewed above (e.g., Bates & Goodman, 1997; Goldberg, 2005, 2009; Gómez, 2002) it was hypothesised that reducing word-level knowledge would reduce generalisation performance.

Statistical learning is traditionally considered an implicit learning mechanism (Aslin & Newport, 2012; Batterink et al., 2015, 2019; Franco et al., 2011; P. J. Reber et al., 2019), but recent research using a variety of statistical learning paradigms has demonstrated the role of explicit knowledge in statistical learning in both adults and children (e.g., Batterink et al., 2015; Mirkovic et al., 2021; Smalle et al., 2018). Thus, in both experiments we assessed the extent to which explicit knowledge of the grammatical regularities might emerge in the course of training and influence generalisation.

2.2 Experiment 1

The focus of Experiment 1 was to address the first aim of this study and assess the role of vocabulary size (i.e., variability in the learning context) in grammatical generalisation in a complex statistical learning paradigm. We used the same artificial language (see Table 1) and study design as Experiment 3 in Mirkovic et al. (2021). The only difference between the two studies was in the size of the training set, which was increased from the original 24 items to 32, allowing for direct comparison to assess the role of training variability. Experiment 1 included two groups: adults and children. Both adults and children were trained using a language production component at training, thus capitalizing on the production effect (e.g., Hopman & MacDonald, 2018).

All participants were tested on three tests of grammatical generalisation, assessing three types of regularities. In the Determiner and Suffix test, participants were tested on the regularity that involved the mapping between the determiner and suffix and the relevant semantic category (e.g., *tib...eem* + animal). In the Suffix Only test, participants were tested on the regularity that involved the suffix only and its co-occurrence with the determiner (distributional cue), as well as the relevant semantic category. In the Phonological Form Only test, participants were tested on the ability to generalise the determiner and suffix distributional regularity (e.g., *tib...eem*), with no reference to the semantic regularity. This last test is the most similar to the test of distributional learning in Gómez (2002). In Mirkovic et al. (2021) Experiment 3, the artificial language comprised 24 items and both adults and children learned it to a good level (children: 76%, adults: 86% accuracy on a vocabulary test), but the children's vocabulary learning was significantly lower than the adults'. Children were only able to generalise in the Determiner and Suffix test, adults in both the Determiner and Suffix and the Suffix Only test, and neither group was able to generalise in the Phonological Form only test. In Experiment 1 we assessed the changes in performance between the two groups, and across the tests assessing different regularities, after learning the same artificial language, but with an increased vocabulary size (32 items). Based on the literature reviewed above, we hypothesised that by increasing the training set generalisation performance in both adults and children should be enhanced.

As a secondary aim, we also assessed the emergence of explicit awareness of the grammatical regularities. Explicit grammatical knowledge was measured using a debriefing questionnaire taken from Mirkovic et al. (2021). To avoid this measure potentially influencing the emergence of explicit knowledge during the experiment as well as influencing the experimental tasks themselves (e.g., Monaghan et al., 2019), the questionnaire was administered at the end of the experiment. We first present the results from the vocabulary learning and grammatical generalisation tasks in both experiments. The results from the explicit awareness test are presented together with Experiment 2 results.

2.2.1 Method

2.2.1.1 Participants

Sixty-one participants took part: 31 adults with a mean age of 19.70 years (19.08-20.67 years; 1 male) and 30 children with a mean age of 10.21 years (9.67-10.82 years; 13 males). The average age in both groups was comparable to Mirkovic et al. (2021; adults: 28.03; children: 10.03). The adult sample was drawn from the undergraduate population at the University of York and received course credits for their participation. The child sample was drawn from primary schools in North Yorkshire, UK. The study protocol was approved by the ethics committee at the Department of Psychology, University of York.

2.2.1.2 Stimuli

All non-proprietary materials are provided on OSF, see also Appendix A (A1-A4, A6-A7) for full stimuli lists.

Colour pictures of familiar objects drawn from Rossion and Pourtois (2001) object database (281x173ppi), and artificial words created from the English database of pronounceable nonwords (Rastle et al., 2002) were used to create the artificial language. The artificial words were constructed using the three elements described earlier (e.g., *a*X*b*) and were digitally recorded by a native English speaker and only presented to participants in audio form. This process was based on the stimuli created by Mirkovic et al. (2011) and Mirkovic et al. (2021). All stem (X) elements consisted of one syllable with a CVC, CCVC or CVCC (C= Consonant, V = Vowel; 'cat' = CVC) structure. The same training and generalization word sets were used across both experiments.

Training Set

Thirty-two word-picture pairs were created, and 16 were used in each semantic category. As illustrated in Table 1, each semantic category was associated with a specific determiner and suffix combination. For the stem (X), an overall balance of CVC, CCVC and CVCC words between the animal and artefact categories was controlled for. Each word was paired with a picture, which denoted the assigned meaning of the word. The stem onset did not match the English onset for the paired animal/artefact.

Thirty-two additional words were created using the same procedure, to be used as foils in the Phonological old and new task, a measure of vocabulary learning (described in more detail below). The stems were similar to the trained word set in the overall phonological structure (e.g., trained words: *tib mofeem*, *ked roivool*; foil words: *tib lupeem*, *ked bisool*).

Generalisation Sets

Three sets of items were designed to test generalisation. None of the words or pictures in these sets had been presented at training. Half of the items in each set were consistent with the trained regularities, and the other half were inconsistent. By changing the construction of the inconsistent items, the three sets aimed to test a specific grammatical regularity: determiner/suffix-semantic mapping, distributional and semantic mappings involving the suffix only, and distributional-only determiner-suffix co-occurrence mapping (Table 2). Each test set is described in more detail below.

For all three sets endorsement rates for consistent vs. inconsistent items were used to derive an A' metric (Pallier, 2002) for analysis. Endorsement of all consistent trials ('hits') and no inconsistent trials would yield an A' of 1.0, while the reverse would yield an A' of 0. The same rate of endorsement for consistent and inconsistent trials would yield an A' of 0.5. Thus, A' scores above 0.5 were taken as an indication of reliably endorsing consistent items more often than inconsistent items and evidence of generalising the regularities present in the training items. The non-parametric A' metric was used as it doesn't make distributional assumptions (Donaldson, 1992; Pallier, 2002; Pollack & Norman, 1964; Zhang & Mueller, 2005).

Determiner and Suffix Generalisation Set

This set of 8 words and pictures was used to assess the mapping between the determiner/suffix and the associated semantic category. Half of the items were consistent, i.e. they conformed to the regularities present in the training set (e.g., <u>*tib*</u> lek<u>*eem*</u> + rabbit). For the inconsistent items, the determiner-suffix co-occurrence conformed to the training set (e.g., *tib_eem*), but it was presented with a picture from the semantic category it had not been paired with during training (e.g., tib darleem was paired with the artefact, bowl, instead of an animal, see also Table 2).

Table 2. Inconsistent item construction for the word-picture matching generalisation tasks.

Task	Determiner	Suffix	Picture Category	Example
Determiner & Suffix	***tib***	***eem***	artefact	<u>*tib*</u> darl<u>*eem*</u> - bowl
	ked	***ool***	animal	<u>*ked*</u> soid<u>*ool*</u> - wolf
Suffix Only	*tib*	***ool***	animal	<u>*tib*</u> sen<u>*ool*</u> - goat
	ked	***eem***	artefact	<u>*ked*</u> dorg<u>*eem*</u> - bell
Phonological Form	*tib*	*ool*	-	<u>*tib*</u> jit<u>*ool*</u>
	ked	*eem*	-	<u>*ked*</u> narp<u>*eem*</u>

Items in **bold** indicate the morphemes which are inconsistent with the semantic category shown in the picture pair (as compared to the training items).

Suffix Only Generalisation Set

This set of 8 items was used to test the co-occurrence between the semantic category and the suffix, as well as the distributional co-occurrence with the determiner. As in the other sets, half of the items contained a word-picture pairing with the consistent mapping conforming to the regularities in the training set. For the inconsistent items, the determiner matched the semantic category of the picture, but the suffix did not match either the determiner or the semantic category of the picture. For example, *tib* sen*ool* was paired with a picture of a goat; here the co-occurrence of tib with the picture of an animal conformed to the trained regularities, but the suffix ool was inconsistent with both the determiner tib and the semantic category of animal (see also Table 2).

Phonological Form Generalisation Set

This set of 8 words tested the co-occurrence between the determiner and suffix, that is, the distributional cue, and without any reference to semantic cues, so no pictures were used in this set. The consistent items conformed to the regularities used in the training set, whereas the inconsistent items had a mismatch between the determiner and the suffix (e.g., *tib* spom*ool* see also Table 2).

2.2.1.3 Procedure

Participants completed all experimental tasks in one session of approximately 40-60 minutes. Responses were recorded by the DMDX programme (Forster & Forster, 2003) on a PC laptop computer, unless otherwise stated. Both adult and child participants were tested individually, with adults completing the experiment in a lab and children in a quiet room or space at their school. The experimental protocol is presented in Figure 1. All participants were introduced to experimental tasks as a series of games involving ‘alien’ words introduced by a visiting extra-terrestrial.

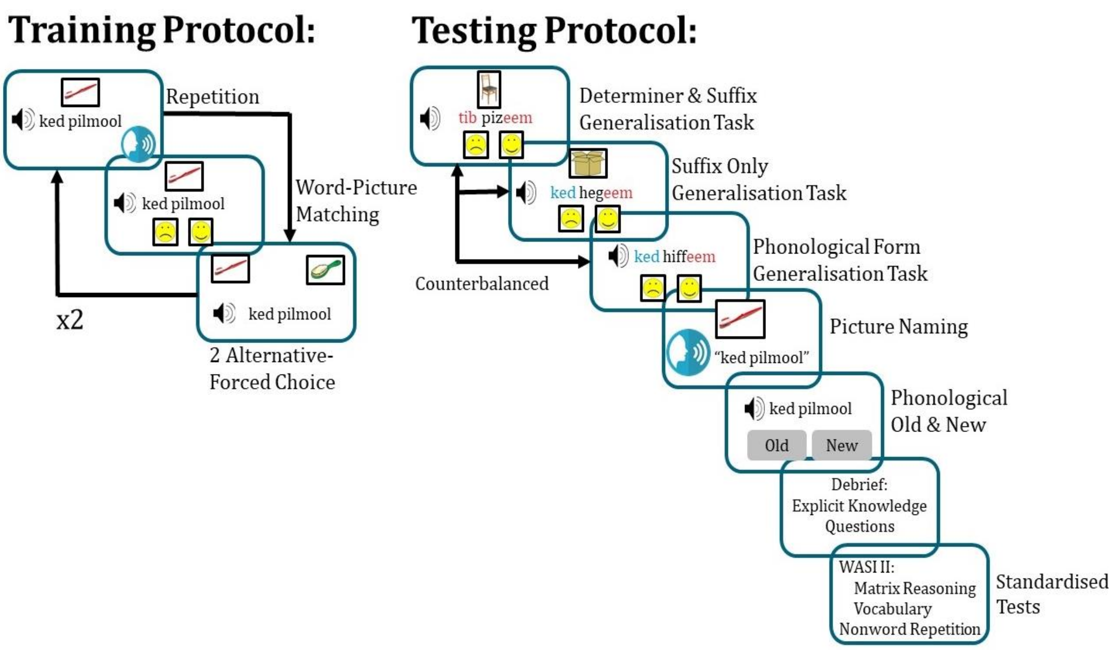

Figure 1. Experiment 1 procedure. Note that the new words were only presented in the audio form. The generalisation tasks depict an inconsistent item.

Training Tasks

Participants were trained on the artificial language using two tasks: word repetition, and word-picture matching. The training tasks were completed twice in succession to one another (Figure 1).

i) *Repetition*

This task introduced the artificial language. Participants were told that they would hear some of the words the alien uses for the different things he finds as he explores earth. They were asked to attend to the presented word-picture pair and then repeat the word aloud once. Each trial started with a fixation cross presented centrally on the screen for 500 ms. This was followed by the presentation of a picture for 500 ms, which was followed by the audio-only presentation of a new word. Participants were asked to press the spacebar to move on to the next trial after they had repeated the word. The picture stayed on screen until the spacebar was pressed, which started the next trial.

This task started with two practice items consisting of a semantic category (fruit pictures) and words (e.g., poy spimoor) which were not a part of the training set. Then, the thirty-two training stimuli were presented in random order once within a block, for three blocks. After the practice trials, and between the three blocks, participants had the opportunity to take a short break, the duration of which was decided by the participant. This task took approximately 5 minutes to complete.

ii) *Word-Picture Matching (WPM)*

In the second training task, participants were presented with word-picture pairs from the alien language and were instructed to indicate if they thought the word and picture 'went well together'. Participants were told that as it was a new language for them, at first they would be guessing, but it would get easier as they went along. For each trial, a fixation cross was displayed centrally for 500 ms, followed by the presentation of a picture for 500 ms, which was followed by an audio-only presentation of the word. Participants indicated their response from one of two options. One, that the 'word and picture **do** go well together' by pressing a happy face sticker placed on the '/' key of the keyboard, or two, that the 'word and picture **do not** go well together' by pressing a sad face sticker placed on the 'z' key. The next trial started automatically after the participant pressed one of the response keys, or after a total of 8000 ms elapsed from the beginning of the trial.

After two practise trials using the same practice items as in the repetition task, participants were exposed to all 32 trained word-picture pairs once (match trials) interspersed with 16 mismatch trials. These consisted of a trained word paired with a previously unpaired trained picture, with eight of these being from the same semantic category as at training and eight not. These mismatch trials provided stimuli for a reject response, modelling the Breitenstein et al.

(2007) paradigm. Two different sets of mismatch trials were used for the two cycles of the word-picture matching task, resulting in all trained words and pictures being used in the mismatch trials. This resulted in matched training items being presented twice and mismatched training items being presented once across the two cycles. The order of trials was randomised for each participant. This task took approximately 3 minutes to complete.

Across the two training tasks and two training cycles, the trained word-picture pairs were presented for a total of 8 times (6 times in the repetition task, 2 times in the word-picture matching task).

Testing Tasks

Three tasks were used to assess levels of word learning, and one task with three different item sets was used to assess grammatical generalisation.

i) Word Learning: Two Alternative Forced Choice (2AFC)

This task tested learning of the novel words and was conducted during and at the end of the training session (Figure 1). In each trial a fixation cross was displayed centrally for 500 ms, followed by a simultaneous presentation of two pictures on either side of the screen for 800 ms, followed by an audio-only presentation of one of the trained words. One of the presented pictures was the correct training picture and the other picture was a foil taken from the training set (with the screen side counterbalanced).

Participants were asked to indicate which of the two pictures matched the word by pressing one of two computer keys. Pressing a blue sticker placed on the '=' key indicated the picture presented on the right side of the screen, and pressing a yellow sticker placed on the '1' indicated the picture on the left side. To ensure that a correct choice was based on word knowledge rather than grammatical knowledge (i.e., knowledge of the determiners and suffixes) the foil picture was always from the same semantic category. The next trial started automatically after a response key was pressed or after 8000 ms elapsed, whichever came first. The order of trials was randomised for each participant. The experimental trials were preceded by two practice trials (using the same practice items as in the training tasks). This task took approximately 3 minutes to complete. Overall accuracy (proportion correct) was used as the outcome measure for this task.

ii) Word Learning: Picture Naming

This task tested word knowledge through the accuracy of cued recall of the items in the training set and was conducted after the generalisation tasks (Figure 1). For each trial a fixations cross was displayed centrally for 500 ms, followed by a picture. The participant was asked to try their best to say the 'alien' word for that picture out loud, and then press the spacebar to proceed to the next trial. Participants were instructed to say if they could not remember the word but were encouraged to say just part of the word if that was all they could remember. This was to enable us to assess any partial knowledge of the words. The next trial started after the spacebar was pressed. The picture remained on screen until the start of the next trial. The order of trials was randomised for each participant. This task took approximately 5 minutes to complete.

Responses were recorded manually by the experimenter and scored off-line, with each word's morphemes and stem coded as correct or incorrect (1 or 0). All word elements were marked as correct if they matched exactly the presented trained words. For determiners, productions were also marked as correct if at least the first two phonemes were correct, and the final was phonemically similar (e.g., /tid/ or 'tib' or /ket/ for 'ked'). Suffix productions were also marked as correct if the first vowel exactly matched (/i:/ for 'eem' or /u/ for 'ool') and the final consonant was phonemically similar (e.g., /i:n/). Stem productions were also marked as correct if the first two phonemes were correct or the first phoneme was correct, and the vowel was phonemically similar (e.g., /gaʃ/ or 'gatch' or /zip/ for 'zeap').

iii) Word Learning: Phonological Old and New

This test assessed recognition of the phonological form of the trained words only and was conducted after the Picture Naming task (Figure 1). In each trial a fixation cross was displayed for 500 ms, followed by an audio-only presentation of one of the words from the training set, or a foil word. Participants were asked to say out loud whether the word was 'old' (they had heard it before) or 'new' (they had not heard it before). All training words and all foils were presented once, in one of two pre-arranged random orders. Participants moved to the next trial by pressing the spacebar. Spoken responses ("old/new") were recorded manually. Word learning on this task was assessed using the accuracy (proportion correct) on the old trials.

iv) Generalisation: Determiner and Suffix and Suffix Only Generalisation

For the Determiner and Suffix and the Suffix Only generalisation sets, we used the same word-picture matching task as the one used at training. The two sets of items were presented in separate blocks. As at training, participants were instructed to attend to 'alien' word and picture pairings and to indicate if they thought they 'went well together'. This trial procedure was the same as in the word-picture matching training task.

v) Generalisation: Phonological Form

For the Phonological Form generalisation set, the trial procedure was the same as the word-picture matching training task with the exception that only the word audio was presented. Participants were instructed to listen to the presented 'alien' word and asked to indicate if the word 'went well with' the 'alien' language they had been listening to.

The generalisation tasks took place immediately after the final 2AFC task, with the order of the three generalisation tasks counterbalanced across participants (Figure 1). Each generalisation task took approximately 2 minutes to complete. A separate A' measure was calculated for each generalisation task for analysis.

vi) Explicit Knowledge Questionnaire

To assess the extent to which any explicit knowledge of the grammatical regularities might have emerged during the experimental tasks, once all tasks were completed, participants were asked 'Did you notice anything about the alien language? Did you use any kind of strategies or clues to decide whether the word and the picture matched?'. Answers were recorded manually. A score from 0-3 was given separately for determiner and suffix knowledge: 0 for no reference to the morphemes or semantic dependencies, 1 for reference of the morphemes but not the dependent semantic cue, 2 for partial knowledge of the morpheme and semantic dependency, and 3 for full knowledge (see Table 3 for examples).

Table 3. Explicit grammatical knowledge: questionnaire scoring system with examples.

Score	0	1	2	3
		Phonology: *Phonological form only explicitly mentioned.*	Partial phonology-semantic knowledge: *Phonological form and semantic association reported for one category*	Full phonology-semantic knowledge: *Phonological form and semantic association reported for both categories*
Determiner Knowledge	No reference to the determiners	*'tib'* and/or *'ked'*	*'tib'*=animals or *'ked'*=artefacts	*'tib'*=animals and *'ked'*=artefacts
Suffix Knowledge	No reference to the suffixes	*'eem'* and/or *'ool'*	*'eem'*=animals or *'ool''*=artefacts	*'eem'*=animals and *'ool''*=artefacts

2.2.1.4 Data Analysis

All analyses were conducted in R (R Core Team, 2020). All data, scripts and outputs are provided on OSF (https://osf.io/mrxwy/ & https://osf.io/2xf6v/). For each task, analysis conducted on data collected within this current experiment will be presented first, followed by comparisons with data from Experiment 3, Mirkovic et al. (2021).

Within the current Experiment, one-sample t-test against chance were used to assess word learning in the 2AFC and the phonological form old-new tasks, and an independent samples t-test to assess differences between the adult and child groups. For picture naming, a mixed-effects logistic regression was used to assess group (child and adults from this current experiment), stem and morpheme differences, using the buildmer package. Generalisation performance was assessed using a one-sample t-test against 0.5 to assess levels of generalisation, and an independent t-test to assess group differences.

To compare the current experiment's results with those from Experiment 3 of Mirkovic et al. (2021), independent t-tests were conducted to compare performance in the 2AFC, phonological form old-new tasks and generalisation tasks. These were conducted to compare performance between the two adult and two child groups from the two respective experiments. Mixed-effects logistic regressions were again run for the picture naming task using the buildmer package. These assessed group differences between the two experiments (i.e., adults vs. adults and children vs. children), stem and morpheme recall.

Effect sizes will be reported as Cohen's *d* or f^2 (Cohen, 1992). For clear and concise graphs, groups will be labelled as follows: Adults/Children HV (High variance) = adults or children from the current experiment, Adults/Children LV (Low variance) = adults or children from Mirkovic et al. (2021), Experiment 3.

2.2.2 Results

2.2.2.1 Word Learning

2AFC

Vocabulary knowledge at the end of training was examined through accuracy on the final 2AFC task. One-sample t-tests against chance (.5) showed that both adults and children learnt the novel words to a good level (Figure 2; adults: M = 0.91; t(30)=28.93, p<.0001, d=5.20; children: M = 0.76; t(29)=9.22, p<.0001, d=1.68). An independent-samples t-test showed that adults had higher levels of word knowledge than children (t(59)=4.89, p<.0001, d=1.25). These results demonstrate that both adults and children within the current experiment, showed good levels of vocabulary knowledge as reflected in this task, with adults achieving this more successfully.

2AFC performance here shows similar levels to those found in Mirkovic et al. (2021; Figure 2; adults = 0.86, children = 0.76), with independent-samples t-tests finding no evidence of a significant difference between adults (t(62)=-1.88, p=.064, d=0.47) and children (t(56)=-0.13, p=.895, d=0.03) in the two experiments.

Phonological Form Old and New

The learning of the phonological form of the novel words was examined at the end of the testing session and yielded similar results to the 2AFC task (Figure 3). One-sample t-tests showed that both adults and children performed significantly above chance (0.5; adults: M=0.88; t(30)=23.83, p<.0001, d=4.28; children: M=0.71 t(29)=5.60, p<.0001, d=1.02). An independent-samples t-test shows that adults demonstrated higher levels of phonological form knowledge than children (t(59)=4.33, p<.0001, d=1.11).

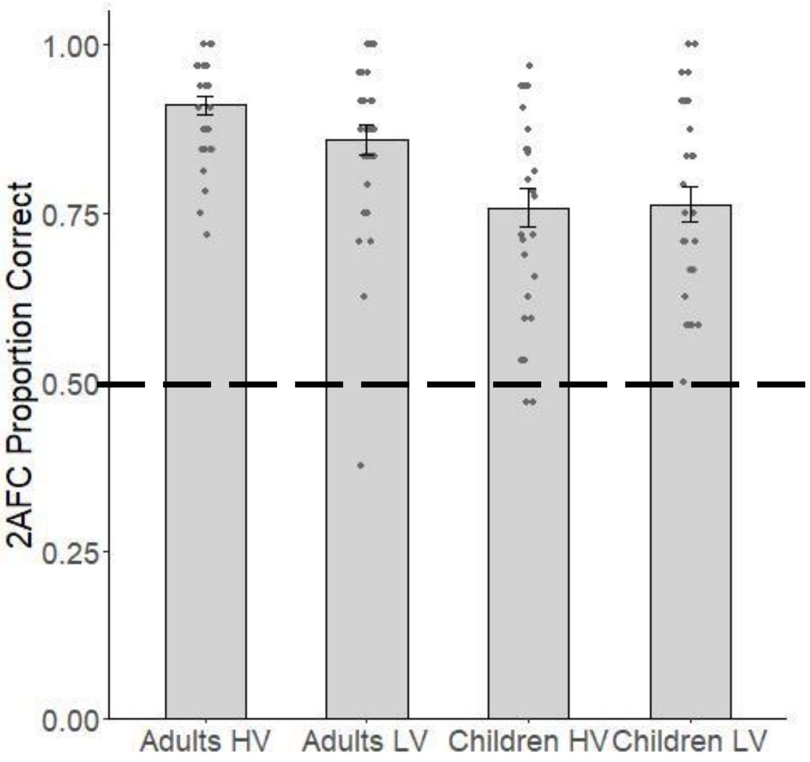

Figure 2. Word Learning: Accuracy on the 2AFC task at the end of training in Experiment 1 of the current study and Experiment 3 of Mirkovic et al. (2021).

These results demonstrate that both adults and children had a good knowledge of the phonological forms of the novel words, with the adults' stronger than the children's.

Considering the results with Mirkovic et al. (2021; Figure 3; adults = 0.81, children = 0.67), independent sample t-tests showed no evidence of a difference in performance between child groups (t(56)=-0.72, *p*<=.474, *d*=0.19). However, adults in the current experiments demonstrated higher levels of phonological form knowledge then adults from Mirkovic et al. (2021; t(62)=-2.67, *p*=.010, *d*=0.67). These results suggest that children are performing similarly here despite the difference in vocabulary variability during training, however higher variability during training seems to have enhanced learning for the phonological forms of the training items in adults.

Picture Naming

Vocabulary knowledge was also assessed using picture naming, a language production-based cued recall task. This task allowed us to separately assess the learning of the stem, and the learning of the determiners and the suffixes. As illustrated in Figure 4Figure 4, overall the performance on this task was lower than on the recognition-based tasks (2AFC and Phonological Form Old and New). Stem recall was particularly low in both groups, and lower relative to the recall of the grammatical morphemes (Figure 4). This is likely due to the differences in the frequency of exposure to the different elements of the word. Similar to the recognition-based tests, a logistical mixed effects regression with stem accuracy as the outcome variable and group as the fixed factors, showed that stem recall was higher in adults than in children within the current experiment (see Table 4).

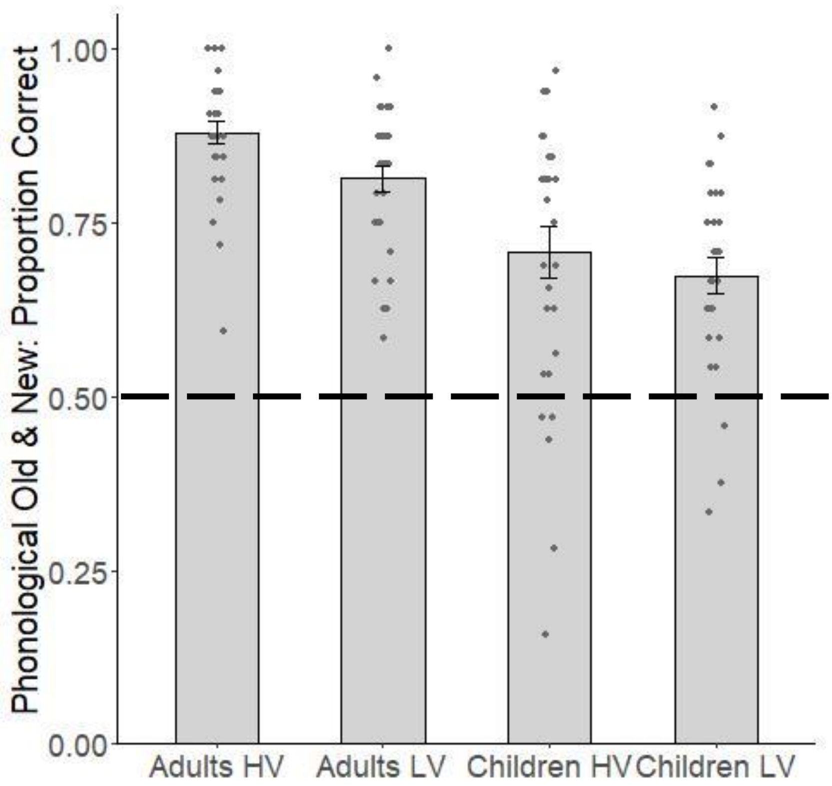

Figure 3. Word Learning: Accuracy for the Phonological Old and New task in in Experiment 1 of the current study and Experiment 3 of Mirkovic et al. (2021).

To assess performance on the grammatical morphemes within this current experiment, another mixed effects logistic regression analysis was run, with accuracy as the outcome variable, and group (adults vs. children) and morpheme (Determiner vs. Suffix) as the fixed factors. The results are summarised in Table 5 and illustrated in Figure 4. As in the other tasks, adults demonstrated higher recall accuracy than children within this current study. These results show that adults demonstrated higher overall morpheme recall accuracy then children but recall accuracy between the two grammatical morphemes was similar within both groups (Figure 4**Error! Reference source not found.**).

The same mixed effects logistic regression analyses were run again, but this time to compare performance with Mirkovic et al. (2021) for stem and morpheme recall between the respective adult and child groups. Results are summarised in Tables 6, 7 and 8 and are illustrated in Figure 4. As can be seen in Table 6, adults and children in this current experiment significantly outperformed their respective comparison groups from Mirkovic et al. (2021) in stem recall.

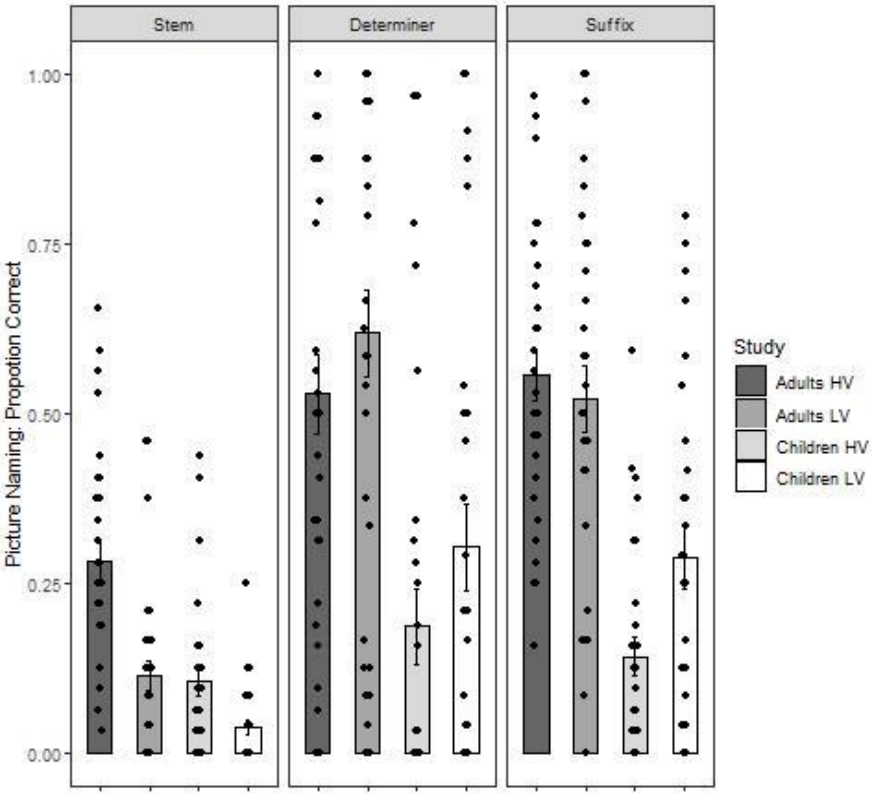

Figure 4. Word Learning: Recall accuracy for the stem, determiner, and suffix in the Picture Naming task in Experiment 1 and Experiment 3 of Mirkovic et al. (2021).

Table 4. Picture Naming: Coefficients for the fixed effects in stem recall in Experiment 1.

	Estimate	Std. Error	z	*p*	*f*²
Group (Adults x Children)	**-1.73**	**0.33**	**-522**	**<.0001**	**0.06**

Random effects included in the final model: random intercepts for group and random slopes for group|items.

Table 5. Picture Naming: Coefficients for the fixed effects in morpheme recall in Experiment 1.

	Estimate	Std. Error	z	*p*	*f*²
Group (Adults x Children)	**-4.23**	**0.92**	**-4.58**	**<.0001**	**0.11**
Morpheme (Determiner x Suffix)	0.12	0.49	0.24	.812	0.00
Group x Morpheme	1.44	0.83	1.73	.084	0.00

Random effects included in the final model: random intercepts for group and morpheme, and random slopes for morpheme|participants, and group|items.

Significant results are highlighted in bold.

An interesting numerical pattern to note here, adults from Mirkovic et al. (2021) performed numerically similarly to children in the current study.

Table 6. Picture Naming: Coefficients for the fixed effects in stem recall in Experiment 1 and Experiment 3 of Mirkovic et al. (2021).

	Estimate	Std. Error	z	*p*	f^2
Group (Adults HV x Adults LV)	**-1.19**	**0.32**	**--4.88**	**<.0001**	**0.07**
Group (Child HV x Child LV)	**-1.60**	**0.68**	**-2.37**	**.018**	**0.03**

Random effects included in the final models: random intercepts for group and random slopes for items.

Significant results are highlighted in bold.

Table 7 presents the adult comparison results for morpheme recall and demonstrates a lack of evidence for a difference in overall morpheme recall between adult groups, nor was their evidence for a difference in recall for the two morphemes both overall and within the two groups. Table 8 presents the child comparison for morpheme recall and interestingly shows a significant difference between the two child groups for overall morpheme recall, with the children from Mirkovic et al. (2021) showing enhanced recall compared to the children in the current experiment. Table 8 also shows that here is no evidence of a difference in recall between the two morphemes both overall and within the two child groups, although the interaction comparison here is marginal. This suggests a trend towards higher determiner than suffix recall potentially driven by children within the current experiment (Figure 4).

Table 7. Picture Naming: Coefficients for the fixed effects in morpheme recall for adults in Experiment 1 and Experiment 3 of Mirkovic et al. (2021).

	Estimate	Std. Error	z	*p*	f^2
Group (Adults HV x Adults LV)	0.26	0.64	0.41	.682	0.00
Morpheme (Determiner x Suffix)	-0.02	0.49	-0.04	.968	0.00
Group x Morpheme	-0.25	0.73	-0.34	.732	0.00

Random effects included in the final model: random intercepts for group and morpheme, and random slopes for morpheme|participants, and items.

Significant results are highlighted in bold.

In sum, in tests assessing levels of vocabulary knowledge both adults and children within the current experiment demonstrated good levels of word learning. Both groups showed good levels of performance in both recognition/language comprehension-based tasks, and poorer performance in the language production-based task. Across all tasks within this current experiment, at the group level adults showed higher levels of word knowledge than children, but significant individual variation was evident in both groups (Figures 2-4). In terms of comparisons with Mirkovic et al. (2021), unlike for vocabulary recognition (2AFC task) both adults and children in this current experiment outperform the equivalent groups in Mirkovic et al. (2021) in stem recall. Although this pattern is similar to adult comparisons in the phonological form old and new task, a test of phonological vocabulary recognition. This suggests that increased variability within the training language has helped to enhance vocabulary production recall for both adults and children (arguably a more demanding task than tests of vocabulary recognition), and phonological knowledge of the vocabulary in adults. In terms of morpheme production recall, there is no evidence that increased variability has enhanced recall for the morphemes in adults, however increased variability during training seems to have decreased child morpheme recall.

Table 8. Picture Naming: Coefficients for the fixed effects in morpheme recall for children in Experiment 1 and Experiment 3 of Mirkovic et al. (2021).

	Estimate	Std. Error	z	*p*	f^2
Group (Children HV x Children LV)	**3.24**	**1.31**	**2.48**	**.013**	**0.06**
Morpheme (Determiner x Suffix)	2.12	1.30	1.63	.103	0.00
Group x Morpheme	-2.26	1.36	-1.66	.098	0.01

Random effects included in the final model: random intercepts for group and morpheme, and random slopes for morpheme|participants, group and morpheme|items.
Significant results are highlighted in bold.

2.2.2.2 Grammatical Regularity Generalisation Tasks

Across all three generalisation tasks, we used the A' measure described above to: i) assess performance against 0.5 for each group with levels higher than 0.5 indicating successful generalisation, ii) for each test we compared the performance between adults and children and, iii) for each test we also compared performance between adults and then children from this current experiment and Experiment 3 of Mirkovic et al. (2021).

Determiner and Suffix

The Determiner and Suffix test assessed generalisation of the distributional regularity and its mapping to the semantic cue. As illustrated in Figure 5**Error! Reference source not found.**, w hilst there was clear individual variation, at the group level one-sample t-tests revealed that only adults performed significantly above 0.5. Children did not show evidence of generalisation (Table 9). Further, an independent t-test confirmed that as a group, adults demonstrated better generalisation in this task than children (Table 9).

These findings partially replicate the findings reported by Mirkovic et al. (2021), in that adults showed better generalisation in this task than children. In both of these experiments' adults, as a group, performed significantly above an A' of 0.5 demonstrating successful generalisation. With an independent sample t-test (Table 10, Figure 5) suggesting the two adult groups performed similar, as no evidence of a significant difference was found. Unlike what we hypothesised, and unlike what was found by Mirkovic et al. (2021) at the group level children in the current study did not show evidence of generalisation in this task. However interestingly, an independent sample t-test shows no evidence of a significant difference in performance between these two child groups, which suggests similar generalisation performance within this task. Although it is worth noting that there is a marginally significant trend towards a higher performance in children from Mirkovic et al. (2021; Table 10, Figure 5).

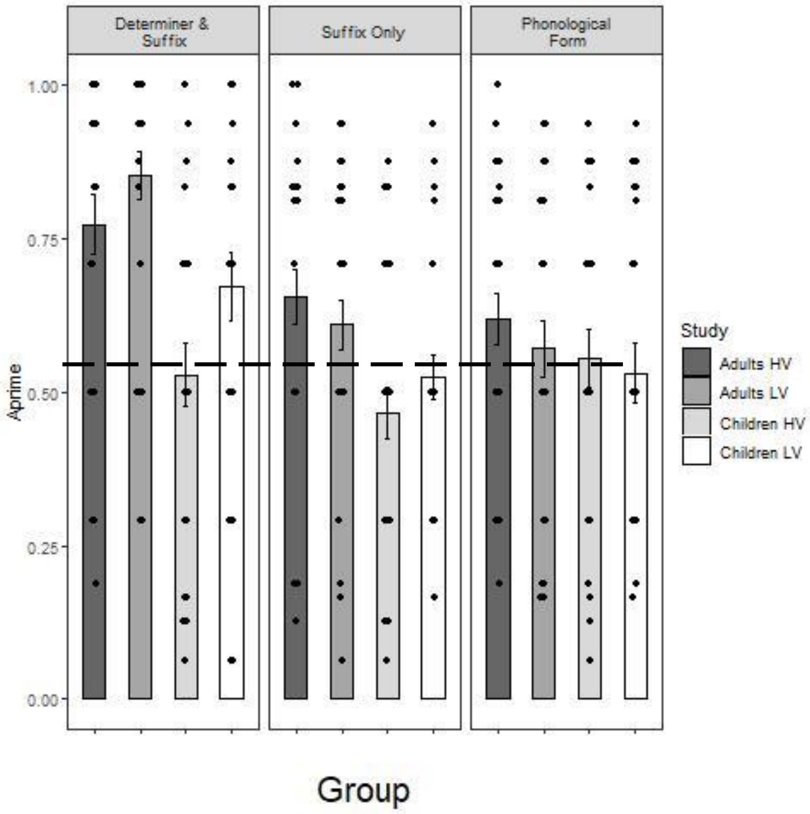

Figure 5. Generalisation: A' performance across all generalisation tasks for Experiment 1 and Experiment 3 of Mirkovic et al. (2021)

Table 9. Statistical analysis for the generalisation tasks in Experiment 1

Generalisation Task	Adults	Children	Comparison	Adults HV		Children HV	
	One-Sample t-test	One-Sample t-test	Independent-sample t-test	M	SD	M	SD
Determiner & Suffix	**t(30)=5.68, *p*<.0001, *d*=1.02**	t(29)=0.54, *p*=.297, *d*=0.10	**t(59)=3.48, *p*<.001, *d*=0.89**	0.77	0.27	0.53	0.28
Suffix Only	**t(30)=3.48, *p*<.001, *d*=0.63**	t(29)=-0.78, *p*=.780, *d*=-0.14	**t(59)=3.07, *p*=.003, *d*=0.79**	0.65	0.25	0.47	0.23
Phonological Form	**t(30)=2.79, *p*<.005, *d*=0.50**	t(29)=1.17, *p*=.126, *d*=0.21	t(59)=1.01, *p*=.316, *d*=0.26	0.62	0.24	0.55	0.26

Significant results are highlighted in bold

Suffix Only

The Suffix Only task assessed the distributional and the semantic regularity but focusing specifically on the suffix. Similar to the performance in the Determiner and Suffix task, again there was clear individual variation (Figure 5), but as before at the group level one-sample t-tests revealed that only adults performed significantly above 0.5. Children did not show evidence of generalisation (Table 9). Additionally, an independent-sample t-test showed significantly better generalisation in adults than in children (Table 9).

These results again replicate the findings by Mirkovic et al. (2021), in that adults successfully generalised in this task, and as in the current study reliably better than children. Table 10 (illustrated in Figure 5) suggests that both adults and children across both experiments performed similarly, as there is no evidence of a significant difference in performance between adults or children. As in Mirkovic et al. (2021) generalisation performance within this current experiment was greater in the Determiner and Suffix task than in the Suffix Only task, as evidenced in the average A' (Figure 5, Table 9 & 10).

Phonological Form

The Phonological Form task assessed generalisation of the distributional co-occurrence regularity with no reference to semantic regularities. One-sample t-tests revealed that adults' performance was significantly greater than 0.5, showing evidence of generalisation in this task (Figure 5 and Table 9). As a group, children did not show evidence of generalisation (Table 9). The group difference was also not reliable (Table 9), and as in the previous generalisation tasks, there was clear individual variation in both groups (Figure 5).

These results partially replicate the findings by Mirkovic et al. (2021) in that children did not show evidence of generalisation in this task with an independent t-test showing no evidence of a difference in performance between children in the current experiment and Mirkovic et al. (2021; Table 10, Figure 5).

Table 10. Statistical comparisons for the generalisation tasks in Experiment 1 and Experiment 3 of Mirkovic et al. (2021).

Generalisation Task	Comparison Adults HV vs. Adults LV	Comparison Children HV vs. Children LV	Adults LV M	Adults LV SD	Children LV M	Children LV SD
Determiner & Suffix	t(62)=-1.30, *p*=.197, *d*=0.33	t(56)=-1.91, *p*=.062, *d*=0.50	0.85	0.22	0.67	0.30
Suffix Only	t(62)=0.75, *p*=.459, *d*=0.19	t(56)=-1.04, *p*=.301, *d*=0.27	0.61	0.23	0.53	0.19
Phonological Form	t(62)=0.76, *p*=.448, *d*=0.19	t(56)=0.35, *p*=.731, *d*=0.09	0.57	0.27	0.53	0.26

Significant results are highlighted in bold

Unlike the adults in Mirkovic et al. (2021), adults in this study demonstrated evidence of successful generalisation, which could suggest that increased variability in vocabulary during training has enhanced generalisation performance. With adults in Experiment 3 of Mikovic et al. (2021) only successfully generalising in the Determiner and Suffix task and the Suffix Only task and adults in this current study now showing successful generalisation across all three generalisation tasks. However, an independent sample t-test comparing performance between the two adult groups for the phonological form generalisation task, does not show evidence of a difference in performance (Table 10, Figure 5). Thus, the evidence for successful generalisation found for this task in adults from the current experiment, needs to be interpreted with caution.

Overall, adults from this current study demonstrated successful generalisation across all word-picture matching generalisation task, with children showing no evidence of successful generalisation performance. When compared with results from Mirkovic et al. (2021), there is no evidence of a significant difference in generalisation across all tasks for both adult and child group comparisons. This suggest that increasing vocabulary variability during training has not enhanced generalisation performance in either adult or child participants.

2.2.3 Discussion

The aim of Experiment 1 was to examine the role of vocabulary size in grammatical generalisation, in terms of variability within the learning context. This was done using a statistical learning paradigm that implements several regularities, representing a more realistic learning environment than typical statistical learning studies (R. Frost et al., 2019). We used a paradigm that mimicked a grammatical gender system. We manipulated variability by increasing the number of stems and thus the training set size in comparison to Experiment 3 of Mirkovic et al. (2021), similar to the original study by Gómez (2002). Like Mirkovic et al. (2021) we tested adults and children. Based on lexicalist theories of grammar (Bates & Goodman, 1997; Goldberg, 2005; Goldberg, 2009) and previous empirical evidence using simpler statistical learning paradigms (e.g., Gómez, 2002), we hypothesised that increasing the training set size would improve generalisation performance in both groups.

Both adults and children showed good levels of new vocabulary knowledge within recognition measures, similar to the original study by Mirkovic et al. (2021) which used a smaller training set. However, vocabulary recall measures showed enhanced performance for both adults and children within this current study compared to Mirkovic et al. (2021). Although there was no evidence of a difference in morpheme recall between adult groups but evidence of higher morpheme recall for children within Mirkovic et al. (2021). This suggests that higher vocabulary variability within training enhanced vocabulary production knowledge, a more demanding element of vocabulary knowledge than recognition. However, this support was not extended to morpheme recall and even seemed to hinder morpheme recall in children.

In the current experiment, adults successfully generalised in all three tests of generalisation, suggesting an improvement relative to previous findings using the same paradigm (Mirkovic et al., 2011; Mirkovic & Gaskell, 2016). In particular, in the current study adults successfully generalised in the test assessing the distributional co-occurrence regularity and its mapping to the semantic regularity (Determiner and Suffix task), as in previous studies. In the current study adults additionally showed evidence of generalisation in the tests that assessed less salient regularities: the Suffix Only test assessed the mapping between the suffix and its mapping to the determiner and the semantic cue, and the Phonological Form Only test assessed the distributional co-occurrence between the determiner and the suffix, with no reference to the semantic cues. This performance also suggests improvements relative to the smaller set size used by Mirkovic et al. (2021). With adults in both experiments demonstrated successful generalisation in Determiner and Suffix and Suffix Only task, but only adults in the current experiment with a larger set size showing evidence of successful generalisation within the phonological form task. However, direct comparisons between these two adult groups did not show evidence of a difference in performance across all tests of generalisation. This finding suggests that increasing vocabulary variability during training has not enhanced grammatical generalisation performance in adults. Thus, the difference in finding and not finding evidence for generalisation between the two adult groups here and its implications for the role of vocabulary variability, needs to be treated with caution.

Unlike adults in the current experiment, children did not show evidence of generalisation, even with an increased training set size. Despite demonstrating similar or enhanced levels of new vocabulary knowledge to children trained using a smaller set size (Mirkovic et al., 2021), they seemed to showed reduced generalisation performance. This is due to finding no evidence of successful generalisation in the Determiner and Suffix test, the only test in which children had successfully generalised in Mirkovic et al. (2021). However, as with adults, direct comparisons here do not show evidence of a difference in generalisation performance between these two child groups, within the Determiner and Suffix test or any of the other generalisation tasks. Therefore, there seems to be no change in generalisation performance between children trained within higher or lower variability conditions.

This was an unexpected result both in terms of a lack of evidence for grammatical generalisation in children and a suggested lack of a variability effect. This is due to previous findings of infants' ability to generalise within a similar artificial language (e.g. Lany & Saffran, 2010; 2011), along with evidence of generalisation and a variability effect in infants (e.g., Gomez, 2002) and findings from Mirkovic et al. (2021). The potential reasons for difference in these findings may be methodological in nature and will be discussed further in the general discussion. Although, it is interesting to note for now that a lack of a variability effect has also been found by Brown et al. (2022), whose paradigm may more closely match the one used here than other previous research.

While increasing variability within the learning context does not seem to have supported grammatical generalisation in either adults and children, within this statistical learning paradigm at least, vocabulary could still help to explain the lack of evidence found for generalisation in children. As in the original study, children's new vocabulary knowledge was lower than the adult's, and across both the current and the series of experiments in Mirkovic et al. (2021) this pattern was strikingly consistent despite differences in training conditions and training set sizes. In the current study, mean accuracy in the final 2AFC task for adults was 91% and for children 76%. In the equivalent 2AFC tasks in Mirkovic et al. (2021) Experiments 1-3, mean adult accuracy was 98%, 84% (end of training in the first session) and 86% respectively. Mean child accuracy was 75%, 75% and 76% respectively. Thus, one possible explanation of the differences in generalisation patterns between adults and children could be that the adult's vocabulary knowledge might have reached the critical mass (variability in knowledge) necessary for at least some grammatical regularity knowledge to emerge. The children's, in contrast, level of vocabulary knowledge may not have reached the needed critical mass for successful grammatical generalisation. This would be predicted by the lexicalist theories of grammar (e.g., Bates & Goodman, 1997). If this is the case, and the critical mass level is the same for adults and children, then a *reduction* in the adult's vocabulary knowledge to the child levels should result in a decrease in generalisation performance.

Reducing vocabulary knowledge to child levels is important for testing the 'critical mass' hypothesis here. Both adult groups considered in this experiment showed equivalent generalisation performance even though one group (Mirkovic et al., 2021) demonstrated lower knowledge in some measures of vocabulary. Thus, even though one groups knowledge is lower than the other, it may still have reached the 'critical mass' needed to support grammatical generalisation. Thus, adult vocabulary knowledge needs to be lowered to what we think is potentially below the 'critical mass' needed, the level of children in the current experiment and from Mirkovic et al. (2021). This question was examined in Experiment 2.

2.3 Experiment 2

Experiment 2 continued to examine the link between vocabulary and grammatical generalisation, but this time considering the role of vocabulary knowledge (e.g., Bates & Goodman, 1997; Goldberg, 2005, 2009). Building on the findings from Experiment 1, in the current study we examined the effect of a reduction in vocabulary knowledge on generalisation but keeping the same training set size as in Experiment 1. A criterion learning method was incorporated into the training to reduce adult word knowledge at the group level to the group level of word knowledge demonstrated by children in Experiment 1. Based on lexicalist theories (e.g., Bates & Goodman, 1997; Goldberg, 2005) and previous findings using simpler statistical learning paradigms (e.g., Gómez, 2002), it was hypothesised that reducing adult word knowledge would reduce their generalisation performance. To assess this question, we directly compared Experiment 2 data with Experiment 1 data, with the more specific hypothesis that Experiment 2 adults would show reduced generalisation performance compared to adults in Experiment 1 but similar to children in Experiment 1.

We tested two groups of participants in Experiment 2 using the same criterion learning method but using different training paradigms. Experiment 2a included both a word repetition and word-picture matching training task as in Experiment 1, now combined into one task. Experiment 2b only included a word repetition task. This meant it was possible to explore the potential role the decision element of the word-picture matching task could have on generalisation performance. Apart from this difference in the training task(s), Experiments 2a and 2b were identical.

2.3.1 Method

2.3.1.1 Participants

Thirty participants took part in Experiment 2a (mean age = 20.77 years (18.17-32.58 years; 4 male), and thirty in Experiment 2b (mean age = 21.09 years (18.25-31.58 years; 5 males). These two samples were drawn from the undergraduate population at the University of York and received course credits or payment for their participation. The study protocol was approved by the ethics committee at the Department of Psychology, University of York.

2.3.1.2 Stimuli

The same training and generalization stimuli sets from Experiment 1 were used. All non-proprietary materials are provided on OSF, see also Appendix A (A1-A4, A6-A7) for full stimuli lists.

2.3.1.3 Procedure

Experiments 2a and 2b followed the same protocol and task order as Experiment 1. The only difference from Experiment 1 was in how the training tasks were conducted (Figure 6 and see below for details). The same visiting alien cover story was used to introduce the experiment.

Experiment 2a Training Tasks

Repetition and Word-Picture Matching (WPM):

For this experiment, the word repetition and WPM tasks were merged. Each trial started with a presentation of a fixation cross for 500ms, followed by the presentation of an 'alien' word and a picture, with the picture onset delayed for 500 ms relative to the auditory word onset. The participants were instructed to repeat the word aloud once and then press the spacebar. Once pressed, the participant was then asked to indicate if the word and picture 'went well together' using the same response protocol from Experiment 1. The picture stayed on the screen until the end of the trial, with the next trial commencing after a response had been made or 8000 ms had elapsed.

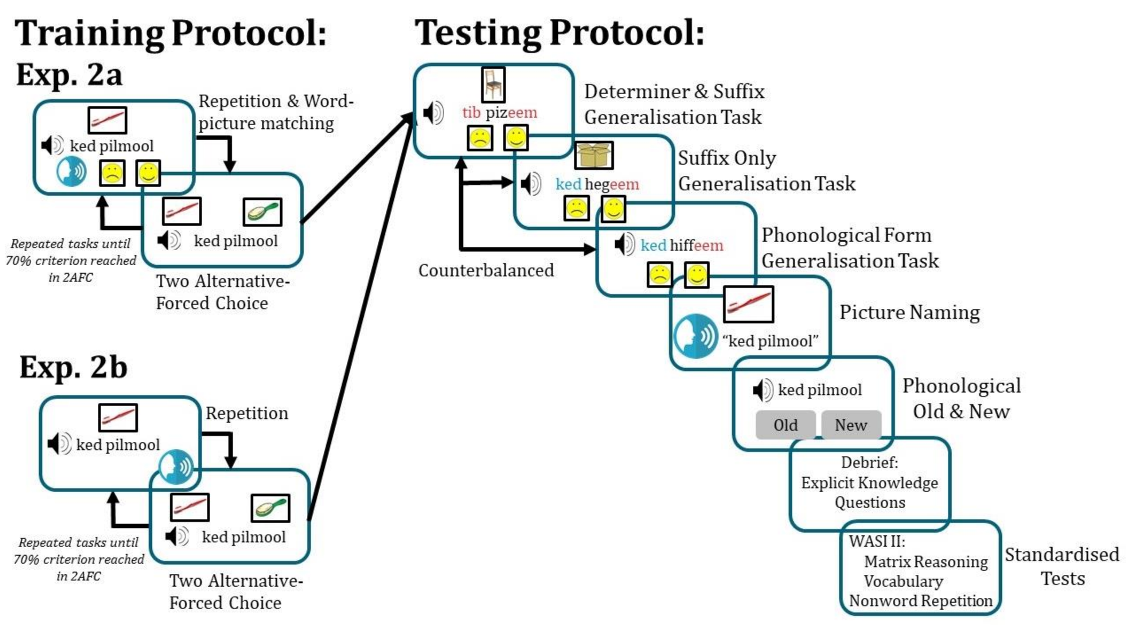

Figure 6. Procedure for Experiments 2a and 2b. The generalisation tasks depict an inconsistent item.

Participants were introduced to the task through two practice trials, using the same items as in Experiment 1. They were then presented with 32 matched trials using the word-picture pairs from the training set and an additional 8 mismatched trials, designed as in Experiment 1. The trial order was randomised. Four different 'Repetition and WPM' blocks tasks were created, each using a different set of mismatched trials. The number of blocks of this task completed by each participant differed due to the criterion learning procedure described below.

Performance on the 2AFC task following each block of the 'Repetition and WPM' training task was used to determine the amount of training: if participants achieved less than 70% accuracy on the 2AFC task, another training block and subsequent 2AFC was administered (Figure 6). Due to a programming error, the practice trial responses were included in the calculation of the 70% accuracy criterion during the task. However, this error did not change the intended reduction in adult vocabulary knowledge to the level of the children's in Experiment 1 (see Results below).

Experiment 2b Training Task

Repetition Only:

Each trial started with a presentation of a fixation cross for 500ms, followed by the simultaneous presentation of a picture and an auditory 'alien' word. As for Experiment 2a and the word repetition task in Experiment 1, participants were instructed to look at the picture and listen to the 'alien' word and repeat it aloud once. The picture stayed on the screen until the end of the trial, with the next trial commencing once the participant pressed the spacebar. Each block presented the same 32 matched and 8 mismatched trials as in Experiment 2a in a randomised order and started with the same two practice trials. As for Experiment 2a, four blocks of this task were created with a different set of mismatch items in each. .

The same criterion learning procedure as in Experiment 2a was used, with each training block being followed by a 2AFC block (Figure 6). The same programming error was found in the calculation of the target 70% accuracy rate during the task, but this did not influence the intended reduction of the adult vocabulary knowledge (see Results below). A further programming error found that two of the four 2AFC tasks used during training did not present one of the experimental items, which meant participants were only tested on 31 of the 32 training items, so the accuracy measure (proportion correct) took this into account by calculating the proportion out of 31 items.

2.3.1.4 Data Analysis

All analyses were conducted in R (R Core Team, 2020). All data, scripts and outputs are provided on OSF.

The same analyses as in Experiment 1 were run to assess levels of word learning and generalisation performance for the three different types of regularities. Additionally, as

Experiments 2a and 2b were designed to explore the generalisation performance in comparison to adults and children from Experiment 1, the results of Experiments 2a and 2b involved statistical comparisons with Experiment 1. Multiple regressions were used to compare groups in all task comparisons (using the lm function in R). The first comparison set used Helmert contrasts to compare adult participants (using the contrast.helmert function in R). Contrast 1 considered Experiment 2a adults (-1) vs. Experiment 2b adults (1), and contrast 2 considered Experiment 1 adults (2) vs. Experiment 2a adults (-1) and Experiment 2b adults (-1). The second comparison set used treatment contrasts (contrast.treatment in R) to compare adults in Experiments 2a and 2b with children in Experiment 1. Contrast 1 considered Experiment 1 children (0) vs. Experiment 2a adults (1), and then Experiment 1 children (0) vs. Experiment 2b adults (1). These treatment contrasts were slightly adapted to include a comparison with Experiment 1 adults for comparing explicit knowledge scores across groups for the exploratory aim of this study. Here, contrast 1 considered Experiment 1 children (0) vs. Experiment 1 adults (1), contrast 2 considered Experiment 1 children (0) vs. Experiment 2a adults (1), and then Experiment 1 children (0) vs. Experiment 2b adults (1). Contrasts were set using default contrast matrices within the contrast function (Schad et al., 2020).

These comparisons allow the success of the criterion learning procedure to be assessed, by comparing measures of word learning between adults in Experiments 2a and 2b and adults and children in Experiment 1. They also allow the testing of the hypothesis that adults in Experiments 2a and 2b will perform significantly lower than adults in Experiment 1, and similarly to children in Experiment 1 in measures of grammatical generalisation.

2.3.2 Results

2.3.2.1 Word Learning

2AFC

As in Experiment 1, performance on the 2AFC task conducted at the end of training was one of the measures of word learning. As illustrated in Figure 7, the criterion learning paradigm worked successfully: participants' levels of word learning in both Experiments 2a and 2b were similar to the child levels in Experiment 1 (Experiment 2a: $M = 0.78$; Experiment 2b: $M = 0.76$; Experiment 1 children: $M = 0.76$). In both cases, the group level performance was significantly above chance (.5; Experiment 2a: $t(29)=21.29$, $p<.0001$, $d=3.89$; Experiment 2b: $t(29)=21.83$, $p<.0001$, $d=3.99$).

To statistically confirm that the adult vocabulary knowledge in Experiments 2a and 2b was at the same level as the children's and lower than the adults' in Experiment 1, we ran two multiple regressions with 2AFC performance (proportion correct) as the outcome variable and group as the predictor variable, coded using the contrasts described earlier. The results of these analyses can be seen in Table 11. For the first comparison (contrast set 1), these results show that adults in Experiments 2a and 2b performed similarly, showing no evidence of a difference in performance but with both performing lower than the adults in Experiment 1. The second comparison (contrast set 2) shows that adults in both Experiments 2a and 2b performed similarly to children in Experiment 1, showing no evidence of a difference in performance. These results demonstrate that the criterion learning method used in Experiments 2a and 2b was successful at reducing the level of adult participants' vocabulary knowledge to significantly below adults and similar to that of children in Experiment 1.

Phonological Old and New

As in Experiment 1, the Phonological Old and New task was used to assess the level of learning of the phonological forms of the new words. As illustrated in Figure 8, one-sample t-tests for both Experiments 2a and 2b show above chance (0.5) performance (Experiment 2a, t(29), 15.47, *p*<.0001, *d*=2.82; Experiment 2b, t(29)15.33, *p*<.0001, *d*=2.80).

Using the same set of contrasts as in the analyses above, multiple regressions were conducted to compare performance with adults and children in Experiment 1. The results of these analysis are shown in Table 12. These results show that adults in Experiments 2a and 2b did not significantly differ in their knowledge of the phonological forms, but this knowledge was lower relative to adults in Experiment 1 (contrast set 1). Interestingly, adults in both Experiments 2a and 2b showed higher levels of phonological form knowledge than children in Experiment 1 (contrast set 2).

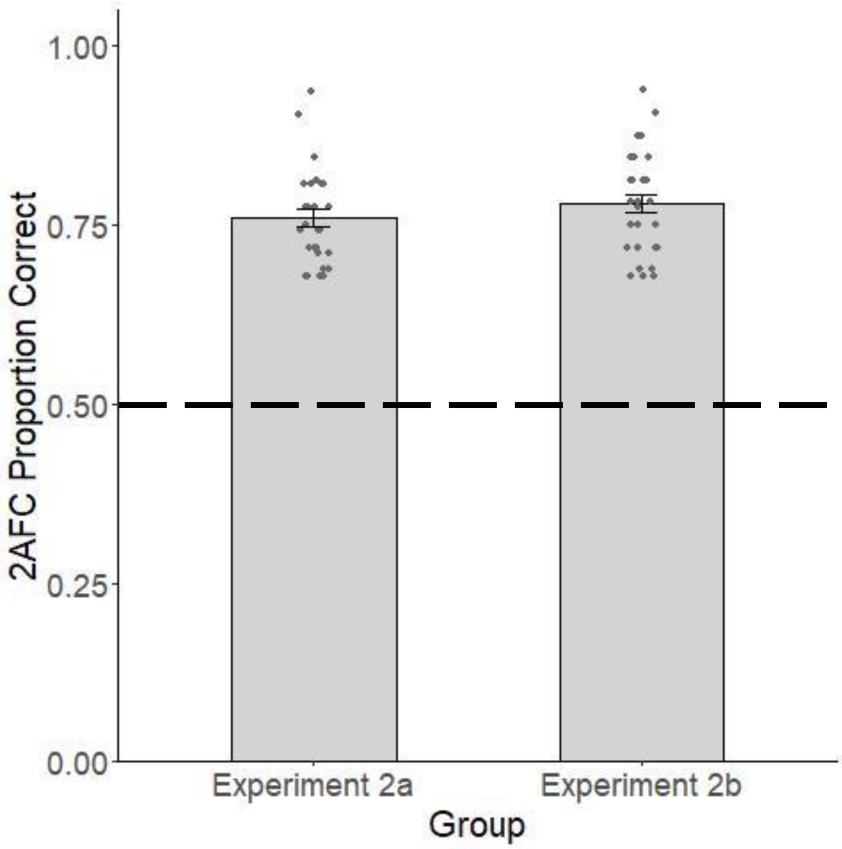

Figure 7. Word Learning: Accuracy on the 2AFC task at the end of training in Experiment 2a & 2b.

Table 11. Comparison analysis for the final 2AFC task for Experiment 1 & 2.

	Estimate	Std. Error	t	*p*	f^2
Contrast Set 1 (Helmet):					
Exp. 2a adults vs. Exp. 2b adults	-0.01	0.01	-1.10	.273	0.01
Exp. 1 adults vs Exp. 2 adults.	**0.05**	**0.01**	**8.79**	**<.0001**	**0.88**
Contrast Set 2 (Treatment):					
Exp. 1 children vs. Exp. 2a adults	0.02	0.03	0.80	.424	0.01
Exp. 1 children vs. Exp. 2b adults.	0.00	0.03	0.04	.966	0.00

Significant results are highlighted in bold

Picture Naming

As in Experiment 1, Picture naming was used to separately assess the knowledge of the stems, the determiners, and the suffixes. As illustrated in Figure 9, and similar to Experiment 1, stem accuracy in both Experiments 2a and 2b was low. Regression comparisons using the aforementioned contrasts were first run with group as the predictor, and stem accuracy as the outcome variable. Results of these analysis can be seen in Table 13. As in the 2AFC task, there was no significant difference in adult performance between Experiments 2a and 2b and in both it was lower than adults in Experiment 1 (contrast set 1). Experiment 1 children did not significantly differ from Experiment 2a and 2b adults (contrast set 2).

Figure 9 also illustrates recall accuracy for the grammatical morphemes in Experiments 2a and 2b. Two mixed effects logistic regressions were run to compare performance across groups and morphemes. The outcome variable was set as morpheme accuracy, with morpheme (determiner, suffix), group (coded as per the contrasts described above) and their interaction, as fixed factors. As shown in Table 1414, and similar to the other measures of vocabulary knowledge, adults in Experiments 2a and 2b did not differ from one another, and they had significantly poorer morpheme accuracy than adults in Experiment 1. Across all adult groups performance between morphemes were similar with no significant interactions between group and morpheme recall (Figure 4, Figure 9, Table 14). When compared to children in Experiment 1 (Table 15), morpheme accuracy overall was similar to adults in both Experiment 2a and 2b. Across groups, there was significantly higher recall for determiners than for suffixes however there were no significant interactions between groups and morphemes.

Overall, the findings from the tasks assessing new vocabulary knowledge demonstrate that the criterion learning method worked well and it resulted in a reliably reduced level of vocabulary knowledge relative to adult participants in Experiment 1 across all three measures of word learning. Relative to children, in the two measures assessing the form-to-meaning mapping aspect of vocabulary knowledge (2AFC, picture naming) the adults' knowledge in Experiments 2a and 2b was reduced to the level of children in Experiment 1. In the test that assessed the knowledge of the phonological form only, adults in Experiments 2a and 2b had reliably higher knowledge than children in Experiment 1. The exclusion of the word-picture matching task in Experiment 2b did not result in a change of levels of vocabulary knowledge relative to Experiment 2a.

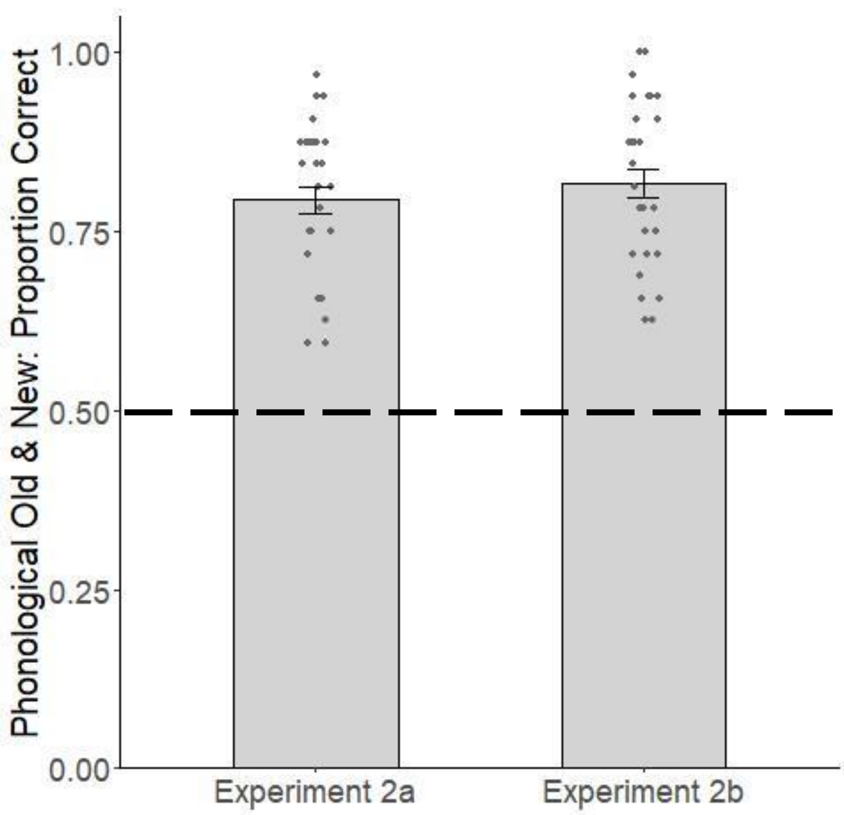

Figure 8. Word Learning: Accuracy for the Phonological Old and New task in Experiment 2.

Table 12. Comparison analysis for the Phonological Form Old and New Task for Experiment 1 & 2.

	Estimate	Std. Error	t	*p*	f^2
Contrast Set 1 (Helmet):					
Exp. 2a adults vs. Exp. 2b adults	-0.01	0.01	-0.87	.388	0.12
Exp. 1 adults vs Exp. 2 adults.	**0.02**	**0.01**	**3.31**	**.001**	**0.01**
Contrast Set 2 (Treatment):					
Exp. 1 children vs. Exp. 2a adults	**0.11**	**0.04**	**2.88**	**.005**	**0.10**
Exp. 1 children vs. Exp. 2b adults.	**0.09**	**0.04**	**2.28**	**.025**	**0.06**

Significant results are highlighted in bold

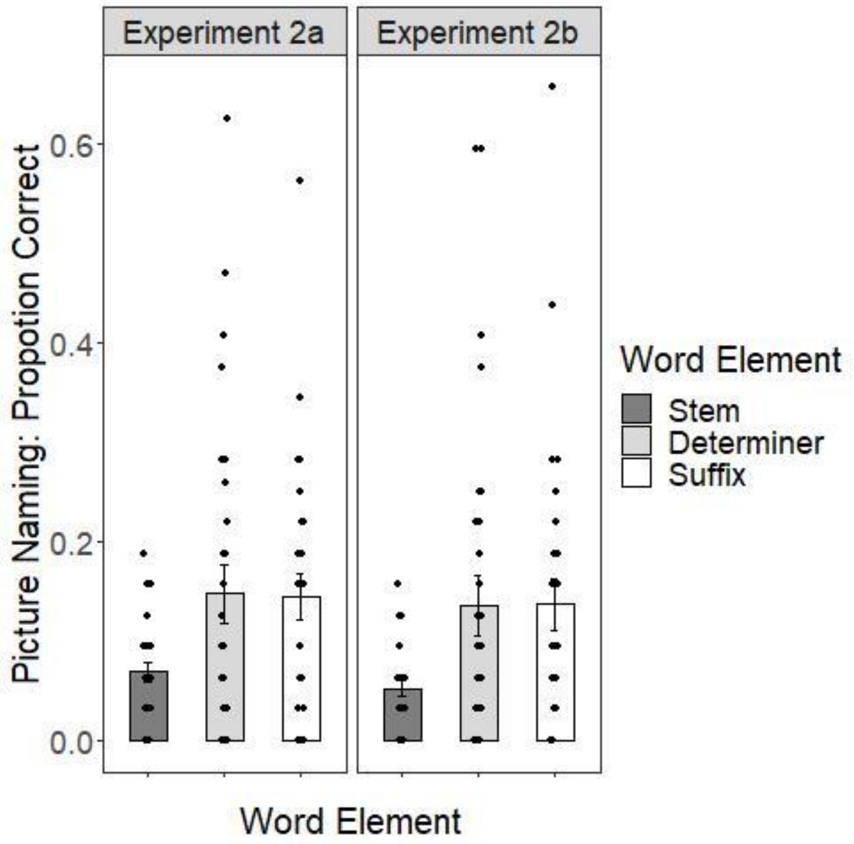

Figure 9. Word Learning: Recall accuracy for the stem, determiner and suffix from the Picture Naming task in Experiments 2a and 2b.

Table 13. Mixed effects regression coefficients for stem accuracy in the Picture Naming task for Experiment 1 & 2.

	Estimate	Std. Error	z	*p*	f^2
Contrast Set 1 (Helmet):					
Exp. 2a adults vs. Exp. 2b adults	-0.15	0.14	-1.09	.275	0.00
Exp. 1 adults vs Exp. 2 adults.	**0.66**	**0.07**	**9.56**	**<.0001**	**0.06**
Contrast Set 2 (Treatment):					
Exp. 1 children vs. Exp. 2a adults	-0.28	0.35	-0.80	.422	0.00
Exp. 1 children vs. Exp. 2b adults.	-0.48	0.38	-1.28	.199	0.00

Random effects included in the model: intercept by participants and items and slopes for group contrasts by participant.
Significant results are highlighted in bold.

Table 14. Mixed effects regression coefficients for morpheme accuracy in the Picture Naming task comparing adults from Experiment 1 & 2 using helmert contrasts.

	Estimate	Std. Error	z	*p*	f^2
Contrast 1: Exp. 2a adults vs. Exp. 2b adults	-0.07	0.25	0.27	.787	0.00
Contrast 2: Exp. 1 adults vs Exp. 2 adults.	**0.94**	**0.14**	**6.78**	**<.0001**	**0.08**
Morpheme (Determiner x Suffix)	0.31	0.19	1.59	.111	0.00
Group Contrast 1 x Morpheme	0.03	0.23	0.12	.902	0.00
Group Contrast 2 x Morpheme	-0.10	0.13	-0.77	.439	0.00

Random effects included in the model: intercept by participants and items, slopes for morpheme by participant and group contrasts by item.
Significant results are highlighted in bold.

Table 15. Mixed effects regression coefficients for morpheme recall in the Picture Naming task comparing children from Experiment 1 with adults from Experiment 2 using treatment contrasts.

	Estimate	Std. Error	z	*p*	f^2
Contrast 1: Exp. 1 children vs. Exp. 2a adults	0.59	0.68	0.86	.389	0.00
Contrast 2: Exp. 1 children vs. Exp. 2b adults.	0.51	0.68	0.75	.454	0.00
Morpheme (Determiner x Suffix)	**0.88**	**0.40**	**2.19**	**.029**	**0.00**
Group Contrast 1 x Morpheme	-0.29	0.52	-0.55	.581	0.00
Group Contrast 2 x Morpheme	-0.26	0.52	-0.49	.624	0.00

Random effects included in the model: intercept by participants and items, slopes for morpheme by participant and group contrasts by item.
Significant results are highlighted in bold.

2.3.2.2 Generalisation

As for Experiment 1, an A' metric based on endorsement rates for consistent and inconsistent trials was derived to analyse performance in the generalisation tasks. The key question in Experiments 2a and 2b was whether a reduction in vocabulary knowledge in adult participants would result in a reduction in generalisation performance.

Determiner and Suffix

The Determiner and Suffix task assessed generalisation of the determiner and suffix regularity and the mapping to the semantic cues. As illustrated in Figure 10, for participants in both Experiments 2a and 2b the A' scores were numerically higher than 0.5, but not statistically reliable (Table 16). Generalisation performance in this task was not significantly different between Experiments 2a and 2b, and it was reliably poorer than the adults' in Experiment 1 (Table 17), and it was statistically not different from the children in Experiment 1 (Table 18). Thus, the overall reduction in vocabulary knowledge in Experiments 2a and 2b did result in poorer generalisation performance in this task, and the generalisation performance was similar to children with the same level of vocabulary knowledge (Experiment 1).

Suffix Only

The Suffix Only task assessed the regularity involving the suffix only: its distributional co-occurrence with the determiner, and the mapping to the semantic cue. As shown in Figure 10 and Table 166, for participants in Experiment 2a the A' score was reliably higher than 0.5, unlike for participants in Experiment 2b. However, the difference between the two groups was small, and not statistically significant (Table 17).

Overall, generalisation performance in this task across both Experiments 2a and 2b was reliably poorer than the adults' in Experiment 1 (Table 17). Generalisation performance for participants in Experiment 2b was not statistically different from the children's in Experiment 1, but for adults in Experiment 2a it was reliably higher than the children's (Table 18).

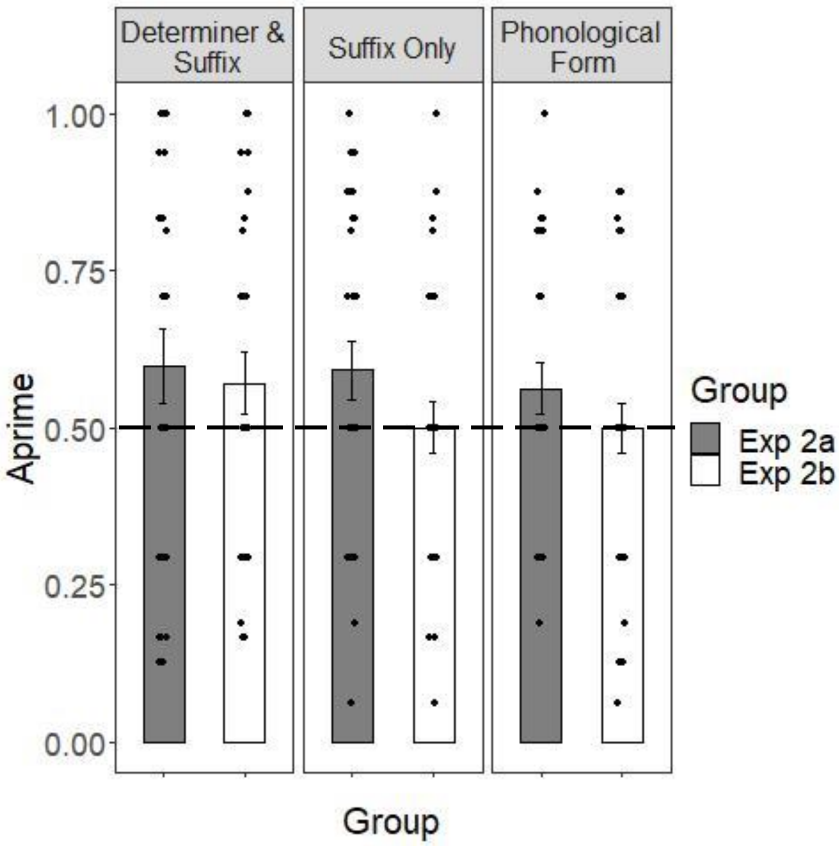

Figure 10. Generalisation: A' performance across all generalisation tasks for Experiment 2.

Phonological Form

The Phonological Form task assessed the determiner-suffix co-occurrence regularity with no reference to the semantic cues. There was no statistically reliable evidence of successful generalisation in this task in either Experiment 2a or Experiment 2b (Figure 10 and Table 16), and the performance between the two experiments did not differ statistically (Table 17). Additionally, there was no statistically reliable difference relative to adults in Experiment 1 (Table 17), who showed weak evidence of generalisation in this task. Finally, when generalisation performance was compared to the children with the same levels of vocabulary knowledge and no evidence in generalisation in this task, there were no differences in performance (Table 18).

In sum, the reduction of adult vocabulary knowledge in Experiments 2a and 2b resulted in overall poorer generalisation performance relative to adults in Experiment 1 across all three tasks, although this lower performance is numerical only in the Phonological Form task.

Generalisation in Experiments 2a and 2b was similar to the children's in Experiment 1, with the exception of the Suffix Only task when the training included a word-picture matching component, where it was significantly better than the children's.

Table 16. Descriptive statistics and one-sample t-test results for Experiment 2's generalisation tasks.

Generalisation Task	2a	2b	2a		2b	
	One-Sample t-test	One-Sample t-test	M	SD	M	SD
Determiner & Suffix	t(29)=1.68, *p*=.052, *d*=0.31	t(29)=1.41, *p*=.084, *d*=0.26	0.60	0.32	0.57	0.27
Suffix Only	**t(29)=1.97, *p*=.029, *d*=0.36**	t(29)=0.00, *p*=.500, *d*=0.00	0.59	0.25	0.50	0.22
Phonological Form	t(29)=1.51, *p*=.071, *d*=0.28	t(29)=0.00, *p*=.500, *d*=0.00	0.56	0.22	0.50	0.22

Significant results are highlighted in bold.

2.3.2.3 Adult word knowledge and grammatical generalisation

To further explore the role of word knowledge in grammatical generalisation, we used an individual differences approach by combining adult data from both experiments and assessing correlations between each generalisation task's A' scores and word knowledge. Accuracy on the final 2AFC task was used as the measure of word knowledge. A significant correlation was found between final 2AFC performance and performance in the Determiner and Suffix generalisation task ($r(89)=0.28$, $p=.008$) but not between the Suffix Only ($r(89)=0.17$, $p=.098$) or Phonological Form ($r(89)=0.16$, p=.120) tasks.

These results partially support the results from the group-level comparisons above, with a significant relationship found between 2AFC performance and Determiner and Suffix generalisation performance. The lack of a significant relationship for the Suffix Only and Phonological Form tasks may be related to overall poorer performance in these two generalisation tasks.

2.3.3 Interim summary

Overall adults in Experiments 2a and 2b demonstrated similar levels of word knowledge to Experiment 1 children, performing below that of Experiment 1 adults. This confirms that the criterion learning paradigm was successful in reducing adults' vocabulary knowledge. It was hypothesised that reducing adults' word knowledge in Experiment 2 would also reduce their grammatical generalisation performance to lower than that of adults and similar to children in Experiment 1. The generalisation analysis supported this hypothesis, with two minor exceptions.

For Suffix Only task, Experiment 2a adults performed significantly better than children and for the Phonological Form task, generalisation performance was only numerically lower than Experiment 2 adults compared to Experiment 1 adults.

Despite these exceptions, the overall findings support a proposed relationship between vocabulary and grammatical generalisation within both the lexicalist theories (e.g., Bates & Goodman, 1997; Goldberg, 2005) and in statistical learning approaches to grammatical generalisation (e.g., Gómez, 2002) in this case demonstrating that a reduced vocabulary knowledge results in poorer grammatical generalisation. Both of these approaches assume that the implicit knowledge of the regularities within the vocabulary items is key to generalisation. However, recent evidence in the statistical learning literature has shown that explicit knowledge often emerges in the course of learning, and it contributes to performance across different paradigms (Batterink et al., 2015; Mirkovic et al., 2021; Smalle et al., 2018).

Table 17. Regression analysis comparing adult generalisation performance across Experiment 1 & 2 using helmert contrasts.

	Estimate	Std. Error	t	*p*	f^2
Determiner & Suffix:					
Contrast 1: Exp. 2a adults vs. Exp. 2b adults	-0.01	0.04	-0.38	.703	0.00
Contrast 2: Exp. 1 adults vs Exp. 2 adults.	**0.06**	**0.02**	**2.97**	**.004**	**0.10**
Suffix Only:					
Contrast 1: Exp. 2a adults vs. Exp. 2b adults	-0.05	0.03	-1.46	.148	0.02
Contrast 2: Exp. 1 adults vs Exp. 2 adults.	**0.04**	**0.02**	**2.04**	**.044**	**0.05**
Phonological Form:					
Contrast 1: Exp. 2a adults vs. Exp. 2b adults	-0.03	0.03	-1.06	.294	0.01
Contrast 2: Exp. 1 adults vs Exp. 2 adults.	0.03	0.02	1.75	.083	0.04

Significant results are highlighted in bold.

Thus, as a secondary aim of the current study we assessed a possible contribution of emergent explicit knowledge to generalisation performance in both experiments. Explicit knowledge was assessed for both grammatical elements using a retrospective verbal report administered at the end of the study, and we coded determiner and suffix knowledge separately (Table 3).

Table 18. Regression analysis comparing Experiment 1 children & Experiment 2 adult's generalisation performance using treatment contrasts.

	Estimate	Std. Error	t	*p*	f^2
Determiner & Suffix:					
Contrast 1: Exp. 1 children vs. Exp. 2a adults	0.07	0.08	0.94	.352	0.01
Contrast 2: Exp. 1 children vs. Exp. 2b adults.	0.04	0.08	0.56	.577	0.00
Suffix Only:					
Contrast 1: Exp. 1 children vs. Exp. 2a adults	**0.12**	**0.06**	**2.04**	**.045**	**0.05**
Contrast 2: Exp. 1 children vs. Exp. 2b adults.	0.03	0.06	0.54	.592	0.00
Phonological Form:					
Contrast 1: Exp. 1 children vs. Exp. 2a adults	0.01	0.06	0.12	.909	0.00
Contrast 2: Exp. 1 children vs. Exp. 2b adults.	-0.05	0.06	-0.91	.366	0.01

Significant results are highlighted in bold.

2.3.3.1 Explicit knowledge and generalisation

The explicit knowledge scores for the determiners and the suffixes for both children and adults in Experiment 1 and both adult groups in Experiment 2 are presented in Table 19. Two mixed effect regressions were conducted to compare explicit knowledge scores between groups and morphemes. The outcome variable was set as explicit knowledge score, with morpheme (determiner, suffix), group (coded using similar contrasts described above, differences will be detailed) and their interaction, as fixed factors. Table 20 shows the results of the first regression which used the previously described helmert contrasts to compare Experiment 1 and 2 adults.

This shows that overall adults in Experiment 1 demonstrated higher explicit knowledge than Experiment 2 adults, but that adults in Experiment 2a and 2b reported similar levels of explicit knowledge. Overall reported explicit knowledge was higher for determiners than suffixes, but with an interaction effect showing this significant difference to be present in Experiment 1 adults and not Experiment 2 adults (see Table 20).

Table 19. Explicit grammatical knowledge scores for Experiment 1 & 2.

	Determiner		Suffix	
	M	*SD*	*M*	*SD*
Exp. 1 Adults	2.39	1.02	0.59	1.03
Exp. 1 Children	0.73	1.11	0.10	0.31
Exp. 2a Adults	1.20	1.19	0.33	0.76
Exp. 2b Adults	1.50	1.17	0.23	0.63

Table 21 shows the results of the second regression which used similar simple contrasts described before to compare Experiment 1 children with adults in Experiment 2 but adapted for this analysis to now include a comparison between Experiment 1 adults and children as well (see section 2.2.1.3). This shows that overall Experiment 1 adults and Experiment 2b adults demonstrated higher explicit knowledge than Experiment 1 children. Experiment 2a adults also show numerically higher overall explicit knowledge but this difference is only marginally significant. Again, overall explicit knowledge is higher for determiners than suffixes, but interaction effects show that this difference is only significant within Experiment 1 adults (Table 21). Although the difference within Experiment 2b adults is marginal.

The key aim of these analyses was to examine whether this explicit knowledge that emerged in the course of the experiment contributed to grammatical generalisation. Multiple regressions were carried out for each generalisation task and for each participant group separately. The outcome variable in each regression was the A' score, while the explicit knowledge scores for the relevant morphemes for the given generalisation task were the predictors: both the determiner and the suffix score for the Determiner and Suffix and the Phonological Form tasks, and the suffix score only for the Suffix Only task.

The results of these analyses are presented in Table 22. For adults in Experiment 1 the emergent explicit knowledge of the grammatical morphemes contributed significantly to the generalisation performance in the Determiner and Suffix and the Suffix Only task, but not in the Phonological Form task. Similar to adults in Experiment 1, there was clear evidence of contributions of the emergent explicit knowledge to generalisation performance in Experiments

2a and 2b: for Experiment 2a, this was the case for the Determiner and Suffix and the Phonological Form task, and for Experiment 2b for the Suffix Only and the Phonological Form task. Unlike adults, in children there was no evidence of contribution of explicit knowledge to generalisation performance for any task. Thus, while adults in Experiments 2a and 2b had similar levels of word learning and grammatical generalisation as children in Experiment 1, and only in one case they developed greater levels of explicit knowledge than Experiment 1 children (determiners in Experiment 2b), there was clear evidence that the adults' generalisation performance was influenced by the emergent explicit knowledge and that was unlike the children's.

Table 20. Mixed effects regression coefficients for explicit knowledge scores comparing morpheme and adults from Experiment 1 & 2 using helmert contrasts.

	Estimate	Std. Error	*t*	*p*	f^2
Contrast 1: Exp. 2a adults vs. Exp. 2b adults	0.15	0.13	1.18	.240	0.01
Contrast 2: Exp. 1 adults vs Exp. 2 adults.	**0.35**	**0.07**	**4.75**	**<.0001**	**0.13**
Morpheme (Determiner x Suffix)	**-1.31**	**0.15**	**-8.98**	**<.0001**	**0.46**
Group Contrast 1 x Morpheme	-0.20	0.18	-1.11	.268	0.01
Group Contrast 2 x Morpheme	**-0.25**	**0.10**	**-2.40**	**.018**	**0.03**

Random effects included in the model: intercept by participants and items, slopes for morpheme by participant and group contrasts by item.
Significant results are highlighted in bold.

Table 21. Mixed effects regression coefficients for explicit knowledge score comparing children from Experiment 1with adults from Experiment 1 & 2 using treatment contrasts.

	Estimate	Std. Error	t	*p*	f^2
Contrast 1: Exp. 1 children vs. Exp. 1 adults	**1.65**	**0.24**	**6.82**	**<.0001**	**0.20**
Contrast 2: Exp. 1 children vs. Exp. 2a adults	0.47	0.24	1.91	.058	0.02
Contrast 3: Exp. 1 children vs. Exp. 2b adults.	**0.77**	**0.24**	**3.14**	**.002**	**0.04**
Morpheme (Determiner x Suffix)	**-0.63**	**0.24**	**-2.59**	**.010**	**0.03**
Group Contrast 1 x Morpheme	**-1.17**	**0.34**	**-3.42**	**.001**	**0.05**
Group Contrast 2 x Morpheme	-0.23	0.35	-0.68	.500	0.00
Group Contrast 3 x Morpheme	-0.63	0.35	-1.83	.068	0.01

Random effects included in the model: intercept by participants and items, slopes for morpheme by participant and group contrasts by item.
Significant results are highlighted in bold.

Table 22. Multiple regressions for the role of explicit morpheme knowledge on generalisation performance.

	Experiment 1 Adults					Experiment 1 Children				
	R^2	BSE	B	ß	*p*	R^2	BSE	B	ß	*p*
Determiner & Suffix:	**0.27**					-0.02				
Determiner		**0.04**	**0.14**	**0.54**	**.002**		0.05	0.03	0.11	.558
Suffix		0.04	0.02	0.08	.599		0.18	0.15	0.16	.403
Suffix Only:	**0.16**					0.01				
Suffix		**0.04**	**0.10**	**0.43**	**.015**		0.14	0.16	0.21	.259
Phon Form:	0.02					0.00				
Determiner		0.04	0.01	0.03	.886		0.04	-0.02	-0.08	.676
Suffix		0.04	0.07	0.29	.131		0.16	0.23	0.27	.161
	Experiment 2a Adults					Experiment 2b Adults				
Determiner & Suffix:	**0.46**					-0.03				
Determiner		**0.04**	**0.17**	**0.62**	**.000**		0.04	0.03	0.11	.553
Suffix		**0.06**	**0.13**	**0.30**	**.037**		0.08	0.08	0.19	.329
Suffix Only:	0.10					**0.11**				
Suffix		0.06	0.12	0.36	.052		**0.06**	**0.13**	**0.37**	**.044**
Phon Form:	**0.20**					**0.10**				
Determiner		0.03	0.04	0.22	.201		0.03	-0.02	-0.11	.544
Suffix		**0.05**	**0.13**	**0.45**	**.012**		**0.06**	**-0.14**	**-0.40**	**.034**

Significant results are highlighted in bold.

2.4 General Discussion

The main aim of the current study was to examine the role of vocabulary in grammatical generalisation. To explore this a statistical learning paradigm was used which included multiple regularities, creating a more naturalistic learning environment. This more naturalistic statistical learning methodology is better placed to consider the mechanisms that underlie grammatical generalisation in language learning as well as investigate the theoretical underpinnings of statistical learning (R. Frost et al., 2019). Both child and adult learners were tested in Experiment 1, and only adult learners in Experiment 2. The results from adult learners will be considered first.

Experiment 1 found ambiguous results for a role for vocabulary in grammatical generalisation in adult learners when the vocabulary available within the learning context was considered. When the number of stems, and thus the variability of vocabulary items was increased, adults' generalisation performance did seem to improved. This initial conclusion comes from considering Experiment 1 results in relation to the findings from Experiment 3 in Mirkovic et al. (2021), whose adult participants went through the same training protocol but with a smaller number of stems (less item variability). Unlike adults in Mirkovic et al. (2021) the adults in Experiment 1 in the current study demonstrated successful generalisation across all of the tested grammatical regularities. This is particularly pertinent for the regularity that involved the distributional co-occurrence between the determiner and suffix with no reference to the semantic cues (tested in the phonological form task), where adults demonstrated successful generalisation in Experiment 1 but not in Mirkovic et al. (2021). Despite this seeming improvement, direct comparison between these two groups of adults did not show evidence of enhanced generalisation performance for adults trained with higher item variability. This was for all generalisation tasks, even the phonological form task where the discussed difference in evidence for generalisation was found. This finding differs from others that have found a role for variability within statistical learning and grammatical generalisation (e.g., Gómez, 2002; Gomez & Maye, 2009; Wonnacott et al., 2012). However, it does support more recent findings from Brown et al. (2022).

The differing findings regarding variability within the literature and the results of this current study may be due to differences in the artificial languages and methodologies used. For example, Gomez (2002) and Gomez and Maye (2009) used a simpler artificial language based only within the phonological domain and as such did not include semantic regularities. The current study used a more complex artificial language incorporating regularities within both the phonological and visual domains through the use of additional semantic cues. It may be that high variability within the learning context is not enough to enhance grammatical generalisation within this more complex paradigm. This idea is also supported by recent findings from Brown et al. (2022), who also did not find a variability effect on grammatical generalisation in adults. Like the current study, the authors also used a more complex language incorporating both distributional and semantic cues, supporting the idea that increasing variability may not be enough to support grammatical generalisation when stimuli complexity is increased. Although it should be noted that Brown et al. (2022) used a semi-artificial language which incorporated real nouns which were known to participants with artificial words to create the grammatical regularities.

While vocabulary variability within the learning context does not seem to play a role in grammatical generalisation within the current paradigm, a role for vocabulary knowledge in grammatical generalisation was found in Experiment 2. Here the training set size was the same as in Experiment 1, but the vocabulary knowledge (i.e., the vocabulary learnt by the (adult) learner) was reduced using the criterion learning paradigm. By comparing performance between

Experiment 1 and 2 we found that when the amount of vocabulary learnt by the participants was reduced, this also reduced generalisation. This finding demonstrates a clearer role for vocabulary in grammatical generalisation and is consistent with the lexicalist theories and statistical learning approaches to grammatical generalisation (e.g., Bates & Goodman, 1997; Gómez, 2002; MacDonald et al., 1994; Wonnacott et al., 2012), at least in adult learners. This finding supports an item-based learning system where increased experience and usage of the item enables similarities to emerge, promoting generalisation of grammatical regularities (e.g., Goldberg, 2005; 2009). The experience and use needed to gain this knowledge enabled generalisation through reaching a 'critical mass' of vocabulary knowledge within the learner's lexicon. Variability of vocabulary within the learning context under circumstances where it does support grammatical generalisation, may be utilising similar processes and mechanisms, it's just the 'critical mass' is found within the learning context rather than within the learner's lexicon. However, the current results in conjunction with Brown et al. (2022) suggest vocabulary knowledge is the more important vocabulary driver for grammatical generalisation within more complex and naturalistic stimuli.

2.4.1 But is it just about vocabulary? Considering explicit knowledge of grammatical regularities.

Whilst vocabulary was directly manipulated in the current study, an exploratory aim was to consider the role of explicit knowledge of grammatical regularities in generalisation performance. In all adult participants, regardless of vocabulary knowledge, verbally reported explicit knowledge of grammatical regularities at the end of the experiment contributed to grammatical generalisation, although where it contributed did vary between groups. This finding replicated similar findings in the current and in other statistical learning paradigms where adults use both explicit and implicit knowledge when learning statistical regularities (Batterink et al., 2015; Batterink et al., 2019; Brown et al., 2022; Franco et al., 2011; Mirkovic et al. 2021).

This finding supports multi-componential views of statistical learning (Batterink et al., 2019; Conway, 2020; R. Frost et al., 2019) and specifically those suggesting an interacting dual system of explicit and implicit processes (e.g., Conway, 2020). While implicit grammatical knowledge has not been considered in this study, the presence of contributing explicit knowledge in a traditionally assumed implicit learning mechanism (for a review see P. J. Reber et al., 2019) is consistent with a multi-componential theory. The current study also supports Brown et al. (2022) who used a similarly complex artificial language and demonstrated that findings of successful generalisation in adults were driven by participants who verbally reported explicit grammatical knowledge. Again, showing a role for explicit knowledge in grammatical generalisation. In their study, the author's link this to the presence of semantic cues aiding the emergence of explicit grammatical knowledge within this learning context, suggesting little room for a role of implicit grammatical knowledge. It may be then, that the use of semantic cues within this current paradigm could be aiding the emergence of explicit grammatical knowledge within adults and like Brown et al. (2022) could be the main driver behind grammatical generalisation.

The presence of semantic cues in the current study may be supporting the emergence of explicit knowledge, but caution and further research is needed before negating a role for implicit knowledge here. Brown et al. (2022) used real nouns that are known to the participants as part of their language (see previous section), along with artificial grammatical words. This means prior language knowledge is available to support learning here. This greater access to prior language knowledge, could potentially be enhancing the emergence of explicit knowledge more so than the use of novel nouns within this current study. Additionally, implicit knowledge was not directly tested for within this current study or in Brown et al. (2022) and as suggested by the multi-componential view of statistical learning, both implicit and explicit knowledge could develop and be present at the same time (Batterink et al., 2019; Conway, 2020; R. Frost et al., 2019). The presence and use of both implicit and explicit knowledge was demonstrated by Batterink et al. (2015) in their word-boundary detection study, which used stimuli that did not include semantics. Thus, further research utilising this current paradigm that directly tests for both explicit and implicit knowledge use would help to better understand the mechanisms at play, both within statistical learning and grammatical generalisation.

A relevant finding in the current study was that adults in Experiment 1, who had higher levels of vocabulary knowledge than adults in Experiment 2, also had higher levels of explicit knowledge of the grammatical regularities relative to adults in Experiment 2. This finding suggests a possible relationship between vocabulary knowledge and explicit awareness of the grammatical regularities, at least in adult participants. Thus, one theoretical possibility is that vocabulary knowledge contributes to grammatical generalisation both directly and indirectly – directly, via reaching critical mass within vocabulary knowledge supporting implicit grammatical knowledge development, and indirectly via influencing explicit knowledge emergence of the grammatical regularities. This theoretical possibility requires further research but if it is the case, it further supports the argument that when adults are more able to utilise prior language like in Brown et al. (2022), this may better support the emergence of explicit grammatical knowledge and its use in grammatical generalisation.

2.4.2 What does this all mean for children?

2.4.2.1 No evidence of generalisation.

Like adults with reduced levels of vocabulary knowledge, children in the current study did not show evidence of grammatical generalisation. This finding is inconsistent with some of the previous literature demonstrating evidence of generalisation in children and infants (e.g., Brown et al., 2022; Gómez, 2002; Lany, 2014; Lany & Saffran, 2010, 2011). The lack evidence found for generalisation in the current study may be methodological in nature. The current paradigm included regularities in several domains (phonological, distributional, and semantic), and all regularities were presented simultaneously. Previous studies have only focused on either phonological and distributional regularities (e.g., Gómez, 2002; Gomez & Maye, 2009; Hall et al., 2018) or trained participants on the phonological form before introducing the semantic regularity (e.g., Lany, 2014; Lany & Saffran, 2010, 2011). It could be argued that a simultaneous presentation may have increased the complexity of the task and the relative salience of the regularities, more than could be resolved by simply increasing the training set size, as was done in Experiment 1.

While increasing the training set size may enhance grammatical generalisation in simpler stimuli (e.g., Gómez, 2002; Gomez & Maye, 2009), the additional demands of more complex stimuli may require gradual or phased introduction of the regularities in the different domains, such as, for example, introducing the phonological and distributional regularities before semantic cues (as in Lany & Saffran, 2010, 2011). This is supported by work that suggests a preference for phonological-based cues in children either due to a general processing preference or a preference that's evolved from experiencing the phonological features of language before the semantic (Culbertson et al., 2017; Culbertson et al., 2019; Brown et al., 2022). Thus, future research using the current paradigm that considers the introduction of the phonological aspects of the artificial language before introducing the semantic information would be of interest here. With the prediction that it may aid grammatical generalisation within child participants.

This idea of complexity hindering grammatical generalisation is also supported by findings from Brown et al. (2022) who unlike this current study, found evidence for successful grammatical generalisation in children (5-6 years old) despite using a similarly deterministic and complex language with all cues presented simultaneously. As discussed in section 2.4 and 2.4.1, the semi-artificial language paradigm used by Brown et al. (2022) incorporated real English nouns of which the native English-speaking participants were familiar with. As acknowledged by the authors, this means the participants were able to use prior language knowledge to support learning within this paradigm, potentially reducing the cognitive demands of the task and allowing for grammatical generalisation of a deterministic language to occur within children. Child participants within this current study could indirectly use prior language knowledge due to the use of known semantic concepts (e.g., using nouns participants would have words for in their native language e.g. dog and table). However, the use of novel noun words to label these concepts may have increased the cognitive demands of the task to the degree that it now interferes with grammatical generalisation.

This is an interesting consideration, as infants younger than those considered both in this current study and in Brown et al. (2022) are able to learn and generalise within learning context that involve completely novel words and semantic information, when it comes to real-world language acquistion. How are infants doing this, if these studies suggest the task becomes too complex for children when known semantic concepts are incorporated with novel words and grammatical regularities? In real-world language acquisition, learning under these conditions does usually take place over longer periods of time than the one session used in this current study. Brown et al. (2022) did train their participants over four session on four consecutive days, suggesting that this increase in learning time could be contributing to their successful generalisation findings. As such, future research using this current paradigm but incorporating longer periods of learning time would be of interest to see if Brown et al.'s (2022) successful child generalisation finding could be replicated in a paradigm that is a step closure to real-world language learning/acquisition.

2.4.2.2 Adults behaving like children: The role of vocabulary knowledge.

Whilst increased complexity may be inhibiting grammatical generalisation in children, the findings from Experiment 2 suggest that vocabulary knowledge may also be playing a role here. Experiment 2 demonstrated reduced grammatical generalisation performance in adult learners when their vocabulary knowledge was reduced to child levels. This suggests that the lower vocabulary knowledge found in children might be a contributing factor to the lack evidence found for generalisation. Considered within lexicalist theories (e.g., Bates & Goodman, 1997), this finding supports the idea of a critical mass of vocabulary knowledge (variability in the lexicon) for grammatical abstraction and generalisation to occur. A clear prediction thus emerges from this, suggesting that increasing the child's vocabulary knowledge (rather than just the training set size, i.e., the variability in the learning context, as in Gómez, 2002) should result in improved generalisation. This prediction can be tested in future studies.

2.4.2.3 Children not behaving like adults: The absence of the role for explicit knowledge of grammatical regularities in generalisation.

As a secondary aim in this study, we explored the contribution of explicit knowledge to grammatical generalisation. There was clear evidence of the contribution of explicit knowledge of the regularities that emerged in the course of the experiment in adult learners, irrespective of their vocabulary knowledge. Despite showing similar levels of vocabulary knowledge to adults in Experiment 2, children developed significantly less explicit knowledge than the adults, and we did not find any evidence of explicit knowledge contributing to generalisation performance in children.

This finding is in contrast to Brown et al. (2022) who found that explicit knowledge was the driving factor behind children's, group level, successful grammatical generalisation performance in the deterministic version of their semi-artificial language. However, as discussed in regards to the role of explicit knowledge in adults in section 2.4.1, this difference could be due to the greater access to prior language knowledge in Brown et al. (2022), potentially better supporting the emergence of explicit grammatical knowledge in children. Particularly if there is indeed a supportive link between vocabulary knowledge and explicit knowledge emergence. The current results are, however, supported by other findings from similar learning paradigms to statistical learning where prior language knowledge is less accessible to support learning.

With the exception of Brown et al. (2022), there is currently a lack of research both within statistical learning as a whole and grammatical generalisation within statistical learning into children's use of explicit knowledge. However, other findings from related paradigms looking at word learning have been considering the use both implicit and explicit knowledge within adults and children. Using a Hebb-sequence learning paradigm, Smalle et al. (2018) found that verbal awareness (a measure of explicit knowledge) was only associated with adult performance and not children's. Together with the findings from the current study, these findings support the idea of potential developmental differences in statistical learning mechanisms. At least when learning is less support by prior knowledge and within more cognitively demanding learning situations. In particular, for more complex grammatical regularities, adults may be utilising both implicit and explicit knowledge, with the potential for vocabulary knowledge to support both the implicit learning of the regularities, and the emergence of explicit knowledge.

In context where children are less able to use prior language knowledge, vocabulary knowledge might be a more crucial variable for grammatical generalisation. According to lexicalist theories, this would be due to item (vocabulary) usage enabling abstraction and generalisation (e.g., Bates & Goodman, 1997; Goldberg, 2005, 2009). This would indicate that it is the role of vocabulary knowledge within implicit grammatical knowledge development that is of importance for children. This idea is supported by further findings from Smalle et al., (2018) who found that children retained implicitly taught (compared to explicitly taught) Hebb sequences better than adults. This is potentially, unless prior knowledge can be utilised to support explicit knowledge emergence, as suggested by findings from Brown et al. (2022). However, further research within this paradigm in both children and adults, which directly measures implicit knowledge as well as explicit knowledge use is needed before firm conclusions can be made here.

2.4.3 Conclusions

As hypothesised by lexicalist theories, vocabulary matters when it comes to grammatical generalisation. The vocabulary knowledge of adult learners has a direct effect on grammatical generalisation, in that when it is reduced it also reduces grammatical generalisation performance. This also suggests that reduced vocabulary knowledge in children results in the lack of evidence found for successful generalisation. In all adult learners, irrespective of vocabulary knowledge, explicit knowledge of the grammatical regularities contributes to grammatical generalisation. This is not the case in children. Together, these findings provide support for multi-componential models of statistical learning and suggest potential developmental differences in the contributions from different components.

Chapter 3. Adapting Subjective Measures of Explicit Knowledge for Grammatical Generalisation within Statistical Learning: A Review

Statistical learning approaches have traditionally assumed that learning the regularities and underlying structure of language proceeds implicitly, without conscious awareness (P. J. Reber et al., 2019). Given that one of the central interests of the field is how young infants learn their first language, this is a reasonable assumption. Nonetheless, many statistical learning studies are carried out with adult participants, often as proxies for infant learners; as adults have highly developed metacognitive skills which they bring to many learning situations (e.g., Flavell, 1979, 1999; Lachman et al., 1979) it is important to scrutinise the assumption of implicitness in statistical learning paradigms. A handful of recent studies have highlighted this issue and provided evidence that adults do indeed have some explicit awareness of what they are learning (Batterink et al., 2015; Mirkovic et al., 2021; Monaghan et al., 2019 and Chapter 2 of this thesis). This awareness has the potential to impact the learning process, and in turn the conclusions that can be drawn from statistical learning studies for theories of language learning. As such, it is important to develop sensitive measures of explicit awareness that can be used within statistical learning paradigms.

Retrospective verbal reports are a straightforward and efficient way to ascertain whether participants have developed explicit awareness of what they have learned. This was the approach used in the first two experiments reported in this thesis (as well as in Mirkovic et al., 2021; Monaghan et al., 2019), and it yielded valuable insights: not only were (some) adults aware of what they had learnt, and able to articulate it at the end of the experiment, but this explicit knowledge was also associated with the level of learning. However, retrospective verbal reports also have some serious limitations, which will be discussed further in section 3.3.1. This review summarises the history and literature surrounding subjective measures of conscious or explicit knowledge, with the aim of identifying key elements to incorporate into the design of a more sensitive measure of explicit knowledge, that can be used in statistical learning paradigms - and specifically, in the grammatical learning paradigm used in this thesis. To do so, it will draw on methodologies and paradigms used in diverse fields of enquiry, including second-language learning, implicit learning, linguistics and sensory perception. These multiple fields of inquiry often use different terms for similar concepts such as the terms ‘conscious’ and ‘unconscious’ which can be synonymous with ‘explicit’ and ‘implicit’, along with ‘awareness’ to mean ‘explicit knowledge’. Unless the original terminology from a literature is needed to aid understanding, the terms ‘explicit’, ‘implicit’ and ‘explicit knowledge’ will be used within this review to avoid confusion.

3.1 Explicit knowledge in the statistical learning and implicit learning literatures.

3.1.1 A tale of two literatures: The historical background to statistical and implicit learning.

The field of Statistical Learning started as a response to Chomsky's (1965) model of formal linguistics, but it was not the only or first literature to do this. In a parallel line of investigation, the implicit learning literature also responded to this, with A. S. Reber's (1967) artificial grammar learning paradigm. Chomsky’s (1965) formal models of linguistics proposed that language is composed of formal structures which are innately specified, and which unfold with maturation. The statistical and implicit learning literatures actively challenge the presence of innate formal structures, and instead seek to understand how the patterns of language can be learnt through experience with the linguistic environment. Although these lines of research proceeded separately, using different paradigms and focusing on different though overlapping questions, they share some important elements: Both investigate learning that happens without conscious awareness, and both involve the passive presentation of stimuli which are constructed with regularities and patterns that learners are not informed about (Batterink et al., 2015; Conway, 2020; Conway & Christiansen, 2005; Franco et al., 2011; Monaghan et al., 2019; Perruchet & Pacton, 2006)

A recent review also shows that the paradigms used in statistical and implicit learning studies utilise similar neural systems, strongly suggesting that these two fields of investigation are in fact accessing very similar if not identical learning processes (Batterink et al., 2019). However, until recently, the two fields differed with respect to whether they directly investigated the possibility of explicit knowledge emerging in the learner: while statistical learning studies have generally not considered this possibility, the implicit learning literature considered this from the outset. This is most likely a consequence of the populations of interest in these fields of research: statistical learning first developed to explore language acquisition in preverbal infants (Franco et al., 2011; P. J. Reber et al., 2019; Saffran et al., 1996), whereas the implicit learning literature focused almost entirely on adults (A. S. Reber, 1967, 1989; P. J. Reber et al., 2019).

However, when statistical learning studies moved into investigating adults, the implicit assumption from infant research came with it, almost without being questioned (P. J. Reber et al., 2019). An early statistical learning study did consider how incidental learning was within a word-boundary paradigm (Saffran et al., 1997). This was considered in terms of attention, where adults and elementary aged (6–7-year-old) children were asked to complete a cover task while being exposed to an artificial language. The idea being that if participants learnt the word boundaries within the artificial language, despite not attending to the language, this would provide evidence of incidental learning and a lack of explicit knowledge use. Participants did show evidence of learning within these learning conditions supporting the idea that statistical learning tasks are a measure of incidental learning or in other words, implicit knowledge. Although, the presence of explicit (and implicit) knowledge was not directly tested for.

Findings of this kind supported assumptions of implicitness within adult participants and the use of adults as proxies for children within statistical learning. However, more recent recognition of the parallels between the implicit and statistical learning literatures means that this assumption of implicitness within statistical learning is now beginning to be revisited. This time, investigations are drawing in part on findings and conclusions from the implicit learning literature that directly tests for the presence of explicit (and implicit) knowledge (e.g., Batterink et al., 2015; Franco et al., 2011; Monaghan et al., 2019).

3.1.2 Explicit knowledge in the implicit learning literature.

Traditionally the implicit learning literature has used artificial grammar paradigms (e.g., A. S. Reber, 1967). Within these paradigms, participants are presented with letters (used as the 'vocabulary') arranged into letter strings ('sentences') constructed using a set of rules ('grammar'). Usually, participants go through a learning and testing phase. During the learning phase participants are randomly presented with a subset of examples from the 'language' and asked to learn the letter strings; recall of these letter strings is either randomly tested throughout learning or asked for after each trial. In the testing phase, participants are made aware that there is an underlying structure to the letter strings they have been learning, before then being. presented with 'grammatical' and 'ungrammatical' letter strings and asked to judge if these strings are grammatical or not.

This paradigm was designed to test the implicit learning capacity for patterns that mimicked those of grammar, to ascertain whether this type of learning was possible for natural language learning (A. S. Reber, 1989; P. J. Reber et al., 2019). However, for this to be the case, what was learnt within this paradigm needed to be the result of implicit learning rather than explicit rule learning (A. S. Reber, 1976), a particular concern when using adult participants. As such, from the start efforts were made to measure explicit knowledge to help ensure that this was not impacting learning. These early efforts suggested that explicit knowledge was not present, implying that what was learnt was implicit.

There were several methods which led to these early conclusions. A. S. Reber (1967) used a form of verbal report to assess the presence or use of explicit knowledge. Participants did not report knowledge of the underlying grammar in an open question and could only verbalise knowledge for more specific questions when provided with successive hints. Acknowledging the limitations of verbal reports, in a second experiment, A. S. Reber (1967) analysed the response patterns of participants, specifically in terms of the types of errors made, to see if they demonstrated the use of explicit knowledge strategies. The main analysis of interest here focussed on whether there was any consistency in what participants judged as 'grammatical' and 'ungrammatical'. If consistency was found, it would suggest the participants were using explicit knowledge or explicit templates to help them respond and if not, it would suggest explicit knowledge was not being used. A.S. Reber (1967) did not find patterns of this nature across participant's incorrect responses and concluded that evidence of learning was based on implicit knowledge.

A. S. Reber, (1969) further investigated the potential presence of explicit knowledge by considering whether underlying pattern knowledge could transfer from one set of letters to another. That is, participants were trained and tested using one stimulus set which incorporated an underlying grammar and then also tested on another, different stimulus set (different letters) but which had the same underlying grammar. Within this study, A.S. Reber (1969) found that participants could transfer underlying pattern knowledge from one set of letters to another, demonstrating the abstraction and use of implicit pattern knowledge rather than explicit item knowledge. Another method to investigate the presence of explicit knowledge, was to compare performance between implicitly and explicitly trained participants (A.S. Reber, 1976). The explicitly trained group were told that the stimuli had an underlying pattern before training, whereas the implicitly trained group were not told this and underwent the traditional incidental learning procedure. Here A. S. Reber (1976) found that implicitly trained participants outperformed the explicitly trained participants both in terms of remembering specific items and when learning about the underlying structure. A. S. Reber (1976) took this as further evidence of the absence of explicit knowledge within artificial grammar tasks that use an incidental learning procedure.

However, these early measures of explicit knowledge were criticised for being insufficiently sensitive (Merikle, 1994; A. S. Reber, 1989; P. J. Reber et al., 2019). For instance, the knowledge transfer procedure (A. S. Reber, 1969) seems to focus on explicit item knowledge rather than explicit pattern knowledge, an important distinction that will be discussed in more depth later in this review. Verbal reports have been highly criticised for several reasons (discussed further in section 1.3.1), an important one being a lack of sensitivity as what participants considered to be key information, may not actually be key information (Merikle & Reingold, 1992). This argued lack of sensitivity is particularly important, as some of these early findings of a lack of explicit knowledge were proving difficult to replicate (e.g., A. S. Reber et al., 1980). To address these criticisms, more sophisticated measures of explicit knowledge were developed, particularly in terms of subjective measures that will be discussed in detail within this review (e.g., Dienes & Berry, 1997). New implicit learning tasks were also developed, including the 'serial reaction time' task, in which participants were presented with sequences built on underlying patterns and asked to respond to these patterns as quickly as possible (Nissen & Bullemer, 1987; P. J. Reber et al., 2019).

Studies utilising these new measures found that explicit knowledge did contribute to performance in the artificial grammar and other implicit learning paradigms, contrary to earlier conclusions in the field (Kelley & Jacoby, 2000; A. S. Reber, 1989; P. J. Reber et al., 2019).

These newer findings were interpreted in the context of a memory systems view of implicit learning, which suggests that both implicit and explicit memory systems contribute to implicit learning, and that interaction between these two memory systems are possible (P. J. Reber, 2013; P. J. Reber et al., 2019; P. J. Reber & Squire, 1994).

3.1.3 Explicit knowledge in the statistical learning literature.

While recognition of the parallels with the implicit learning literature have prompted investigation into explicit knowledge within the statistical learning field, the number of studies examining this is still small. Nonetheless, their findings echo what has been found within the implicit learning literature. Franco et al. (2011) directly tested for the presence of explicit knowledge within a word-detection paradigm using a process-dissociation task drawn from the memory literature (Jacoby, 1991). This task found that adult participants could exert voluntary control over the transitional probability knowledge they had gained. Their ability to exert voluntary control suggests the presence of explicit knowledge (Dienes et al., 1995; Dienes, 2007; Dienes & Perner, 1996).

Batterink et al. (2015) built on these findings within the word-boundary detection paradigm, assessing the presence of both explicit and implicit knowledge using a more comprehensive range of measures. These were based on techniques from the implicit and second language learning literature. They included the use of subjective measures within a two-alternative forced choice testing task, the use of a speeded target detection task designed to capture reaction times, a measure more sensitive to implicit knowledge use (Kelley & Lindsay, 1996; Rebuschat, 2013; Timmermans & Cleeremans, 2015) and the use of ERP throughout both of these tasks, specifically using the P300 effect thought to capture the processing of predictive stimuli (see section 1.3.3.1 in Chapter 1 for more details). Using these measures Batterink et al. (2015) found that both implicit and explicit knowledge played a role within this statistical learning paradigm.

Similar findings have recently been reported in studies focusing on grammatical learning from a statistical learning perspective. Monaghan et al. (2019), Mirkovic et al. (2021) and chapter 2 of this thesis all used a cross-situational learning design to teach an artificial language which incorporated both vocabulary and grammatical regularities. Monaghan et al. (2019) tested the learning of trained grammatical regularities and Mirkovic et al. (2021) and Chapter 2 of this thesis considered the generalisation of regularities. These studies all used verbal reports to examine the role of explicit knowledge of grammatical regularities, similar to A. S. Reber (1967).

Using this measure, Monaghan et al. (2019) considered learning in both implicit and explicit learning conditions and found that even when participants underwent implicit learning conditions, some of the participants were still able to report explicit knowledge of the underlying grammatical regularities. Mirkovic et al. (2021) and Chapter 2 of this thesis only used implicit learning conditions and found that across experiments, most participants reported at least some knowledge of the grammatical regularities. Reports of explicit knowledge also partially contributed to the generalisation of grammatical regularities (except for child participants, see Chapter 2). These findings support those found within the word-boundary paradigm literature (Batterink et al., 2015; Franco et al., 2011). As mentioned in reference to the implicit learning literature (e.g., A. S. Reber, 1967), there are limitations to verbal report measures of explicit knowledge, an issue which will be explored in more detail in section 3.3.1. Despite this, these studies represent a start in exploring the role of explicit knowledge within grammatical paradigms of statistical learning.

These recent findings, along with the recognised parallels with implicit learning have prompted a recent theoretical shift within the statistical learning literature. A new multi-componential view of statistical learning has been proposed (Conway, 2020; R. Frost et al., 2019), similar to that suggested within the implicit learning literature (e.g., P. J. Reber, 2013; P. J. Reber & Squire, 1994). Specifically, it has been suggested that statistical learning draws on an interacting dual-system that includes both implicit and explicit learning mechanisms. These recent reviews and theoretical proposals have incorporated findings from the implicit learning literature as well the statistical learning literature. However, there is still a need for empirical studies using more complex paradigms to test these theories and move the literature forward (Conway, 2020; Frost et al., 2019; P. J. Reber et al., 2019; Rebuschat, 2013). Therefore, it is important to develop sensitive and robust measures of explicit knowledge for use with grammatical paradigms within statistical learning, as this represents a more complex paradigm for investigation which can better represent natural language processing.

3.2 Explicit knowledge and the attribution model.

3.2.1 Subjective measures of consciousness: How do they measure explicit knowledge?

Subjective measures of consciousness aim to assess the ‘subjective threshold’ of consciousness and follow from the concepts of first- and second-order mental states (Cheesman & Merikle, 1984; Dienes et al., 1995; Dienes, 2007; Dienes & Berry, 1997; Moroshkina et al., 2019). A first order state is described as a ‘mental state about the world’ and is measured using ‘objective thresholds’ (Cheesman & Merikle, 1984; Dienes et al., 1995; Dienes, 2007; Dienes & Berry, 1997; Moroshkina et al., 2019; Norman & Price, 2015). In the case of grammatical knowledge, a first order state would be a representation of a grammatical regularity that can inform behaviour in tasks where this knowledge is assessed. When a participant’s knowledge is assessed and they perform above chance they are said to have reached the objective threshold (Cheesman & Merikle, 1984; Moroshkina et al., 2019; Norman & Price, 2015). Reaching this objective threshold can only tell us that a first order mental state has been formed and is influencing behaviour. It cannot tell us anything about a person’s awareness of this mental state, or in other words, whether participants have drawn on explicit or implicit knowledge when giving responses (Cheesman & Merikle, 1984; Dienes, 2007; Moroshkina et al., 2019; Norman & Price, 2015).

This is where second-order states and the subjective threshold come into play. Second-order states are 'mental states about mental states' (Cheesman & Merikle, 1984; Dienes, 2007; Dienes et al., 1995; Norman & Price, 2015). They are the awareness a person has about their first order state or knowledge they are using to successfully respond to a task. Subjective measures are designed to probe participants about their mental states or awareness of first order mental states. If these measures indicate that there is awareness, the participants are said to have met the subjective threshold (Cheesman & Merikle, 1984; Dienes, 2007; Dienes et al., 1995; Norman & Price, 2015). Using grammatical regularity knowledge as an example again, if a participant has developed a second-order mental state for a grammatical regularity representation they have formed, they would be aware of this first-order state and how it guides their behaviour. This participant has then met the subjective threshold and this awareness would be detected by subjective measures of consciousness (Cheesman & Merikle, 1984; Dienes, 2007; Dienes et al., 1995; Norman & Price, 2015).

Thus, subjective measures aim to detect the presence of second-order states, or in other words, detect the presence of explicit knowledge. Confidence rating measures are a form of subjective measure. They aim to assess the subjective threshold by asking participants to report the confidence they have in their response to the objective threshold measure. For example, how confident a person is in their judgement of an item being 'grammatical' or 'ungrammatical'. Two criteria have been proposed to help assess whether the subjective threshold has been reached when measured by confidence ratings: the 'guessing criterion' and the 'zero-correlation criterion' (Dienes et al., 1995).

These two criteria are built around the idea that there are two ways in which knowledge could be implicit. Firstly, participants may be performing above chance and have reached the objective threshold, but they judge that they are guessing so have not reached the subjective threshold (Cheesman & Merikle, 1984). This has been named the 'guessing criterion' by Dienes et al., (1995). Secondly, if a person has developed second-order states about first-order states, then when they are asked about their objective threshold judgements they will know when they 'know' and know when they are 'guessing'. Thus, if explicit knowledge is present and being used it should result in a relationship between confidence and accuracy, with higher accuracy being associated with higher confidence. Then by contrast, if there is 'zero' relationship between confidence and accuracy it suggests the use of implicit knowledge. This constitutes the 'zero-correlation criterion' (Chan, 1992, but see also Dienes, 2007; Dienes et al., 1995; Dienes & Berry, 1997; Moroshkina et al., 2019; Norman & Price, 2015). Implications of carrying out analysis for these two criteria will be considered in section 3.4.5 and 3.4.6.

3.2.2 Confidence ratings within the memory literature: Recollection and feelings of familiarity.

Confidence measures within memory research have developed around the concepts of 'recollection' and 'feelings of familiarity'. These concepts are part of the 'attribution' model of knowing (Jacoby et al., 1989; Lindsay & Kelley, 1996) which proposed that memory recall or recognition is not necessarily based on memory 'traces' or in other words specifically remembering an event or item (recollection). Recall could also be based on the result of unconscious processing, with prior processing of an event or item implicitly facilitating the current recall or recognition process. This unconscious processing results in 'feelings of familiarity' for the recalled or recognised event or item (Jacoby, 1991; Jacoby et al., 1989; Kelley & Jacoby, 2000; Kelley & Lindsay, 1996; Lindsay & Kelley, 1996). The terms 'remember' and 'know' are also used to distinguish between these two concepts, with 'remember' referring to recollection and 'know' to feelings of familiarity (Jacoby et al., 1989; Lindsay & Kelley, 1996).

Within this model a confidence-based system has been developed that aims to assess the subjective threshold through distinguishing between explicit (recollection/remember) and implicit (feelings of familiarity/knowing) knowledge (Dienes & Perner, 1999; Dienes, 2007; Dienes et al., 1995; Jacoby et al., 1989; Kelley & Jacoby, 2000; Lindsay & Kelley, 1996). This is also known as the 'remember/know' paradigm (Kelley & Lindsay, 1996; Lindsay & Kelley, 1996). The different confidence rating measures built on this attribution model are illustrated in Table 23. An example of this measure can be found in a study investigating whether 'illusions of familiarity' could be created in cued-recall tasks by Lindsay and Kelley (1996). Here the authors use two forms of confidence measures which utilise the 'remember/know' paradigm. After training participants on a list of English words they were tested for knowledge of these words using a cued-recall task. After each item in this testing task, participants were asked to describe their memory of the word using three options. In two experiments, these options were: clear memory (linked to remember), feels familiar (linked to know) and no memory. In a third experiment the authors used: remember, know and no memory. These options were taken to indicate the type of knowledge participants were using with 'remember' responses indicating explicit knowledge use and 'know' responses indicating implicit knowledge use.

This subjective measure of consciousness has been developed to examine explicit knowledge of a specific event or item knowledge rather than knowledge of patterns/regularities. However, the concepts behind this measure do relate to proposed theoretical models of statistical learning in terms of implicit processing (e.g., 'chunk and pass', Christiansen & Chater, 2015). This can be highlighted when considering the 'feelings of familiarity' concept, which proposes that prior processing of an item can implicitly facilitate recognition. This echoes the idea of past processing supporting the building of connections to form representations, as proposed by the 'chunk and pass' model (Christian & Charter, 2015). This model suggests that this past processing can in turn be used to anticipate or predict incoming information within a current

Table 23. Empirical studies from the memory literature.

Reference	Stimuli Type	Key participant information	Subjective Awareness Measure	Combined Scale	Scale Measures:
Lindsay & Kelley (1996)	Native Natural Language (English words)	Undergraduate students Exp. 1: N=32 Exp. 2: N=48 Exp. 3: N=64	**Exp. 1&2 - Remember/No Word scale:** *Clear memory (remember)* *Feels Familiar (know)* *No Memory* **Exp. 3 – Remember/No Word Scale:** *Remember* *Know* *No memory*	No	Awareness
Lockl & Schneider (2002)	Native Natural Language (German words)	N=35, *Mage*=7;5 (SD=4 months) N=32, *Mage*=8;5 (SD=4 months) N=33, *Mage*=9;7 (SD=5 months) N=32, *Mage*=10;7 (SD=5 months)	**Confidence Likert Scale using happy/sad faces:** *1 = big frown (I will surely not select the correct picture)* *2 = frown* *3 = smile* *4 =Big smile (I will surely select the correct picture)*	No	Awareness
Visser, Krahmer & Swerts (Exp. 2 &3; 2014)	General Knowledge	**Exp. 1:** N=42; 8-year-olds N=48; 11-year-olds **Exp. 2:** N=38; range=18-27 years olds	**Confidence Likert Scale using happy/sad faces:** *1 = big frown (very uncertain)* *2 = frown* *3= Neutral* *4 = smile* *5 =Big smile (very certain)*	No	Awareness

processing task. Similar to the use of it for recognition or recall suggested within the 'feelings of familiarity' concept. The 'remember/know' paradigm has also been used successfully to assess explicit and implicit knowledge use within statistical learning. In Batterink et al. (2015), the authors used this measure within their word-boundary paradigm. When testing for transitional probability knowledge within a two-alternative forced choice task, participants were also asked to judge what they based their decision after each trial using 'remember', 'familiar' and 'guess' options. This measure successfully indicated explicit knowledge use.

The use of the 'remember/know' paradigm within regularity knowledge does have limitations. Although it can detect use of explicit knowledge, it cannot specify what type of explicit knowledge is being used. Tasks such as A. S. Reber's (1967) artificial grammar learning, or Monaghan at al.'s (2019) cross-situational learning paradigm, trains and tests participants using the same items. Thus, if explicit knowledge is found by the 'remember/know' paradigm, it cannot differentiate between explicit knowledge of a specific item or explicit knowledge of the underlying pattern within the stimuli. In response to this issue, the 'decision strategy attribution' test was developed (Dienes & Scott, 2005; Moroshkina et al., 2019). The choice options it provides to participants incorporate concepts of 'recollection' and 'feelings of familiarity' and also allows for the distinction between item and pattern/regularity knowledge. The 'guess' and 'intuition' options reflect feelings of familiarity, the 'recollection' option reflects the recollection of a past item and 'rule knowledge' the recollection of regularities.

3.2.3 The 'decision strategy attribution' in use.

For their grammatical paradigm study, Monaghan et al. (2019) created 'sentences' which included a noun for a pictured shape and a verb for the movement of the pictured shape. Both the noun and the verb had a marker word which would indicate if the following word was a noun or verb. For example, in *'Tha makkot noo oakrid', 'tha'* and *'noo'* served as the marker words and *'makkot'* and *'oakrid'* referred to the object or the movement of that object. For the cross-situational learning task, two scenes (one foil and one correct) were presented along with one artificial sentence. The participant was asked to decide which scene went with the sentence. This task was used throughout the experiment for both training and testing trials.

Monaghan et al. (2019) used a 'decision strategy attribution' test to assess explicit knowledge within experiment 2 of their study. This experiment only considered an incidental learning condition and after each trial in the cross-situational task, participants were asked to judge the basis of their two-alternative forced choice decision. To judge this, participants were given the four options that form part of this attribution test: 'guess', 'intuition', 'recollection' of a past item or 'rule knowledge'. Monaghan et al. (2019) found that this judgement scale showed a shift from implicit knowledge use ('guess'/'intuition') to explicit ('recollection'/'rule knowledge') through training. Explicit knowledge-based judgements could predict later performance on similar items presented subsequently. This finding provides more sensitive and robust evidence of explicit knowledge use within a complex statistical/implicit learning paradigm.

Despite this finding using a more sensitive and robust measure of explicit knowledge, there is still a potential problem that could be confounding Monaghan et al.'s (2019) findings. This issue is inherent within the 'decision strategy attribution' test in the use of the option 'rule knowledge'. This option could be highlighting to participants that there is a rule they could be looking out for. This could then be influencing and aiding the emergence of explicit knowledge within these incidental learning conditions. As such, this version of a confidence rating measure could be reducing the ecological validity of the design. This is an important consideration for the statistical learning literature which aims to better understand real-world language acquisition and processing (R. Frost et al., 2019). It is also a concern for a subjective measure, as for them to successfully measure explicit knowledge they need to meet specific criteria. This includes not having an impact on the behaviour being measured within the objective threshold measure (Moroshkina et al., 2019; Newell & Shanks, 2014), an issue which will be discussed in more depth in the next section.

The use of the 'decision strategy attribution' test by Monaghan et al. (2019) has started to improve the measure of explicit knowledge use within a grammatical paradigm of statistical learning. Whilst the use of 'rule knowledge' within this test may have encouraged the emergence of explicit knowledge, not all participants reported explicit knowledge within the verbal reports measure conducted at the end of the experiment (Monaghan et al., 2019). However, given the limitations of verbal reports (which will be detailed in section 3.3.1), this finding is unable to provide conclusive evidence that the 'decision strategy attribution' test has not encouraged the emergence of explicit knowledge. Thus, further work is needed to adapt and develop a new subjective measure for use within a grammatical paradigm of statistical learning.

3.3 Design criteria for an improved measure of explicit knowledge.

In response to the criticisms of verbal report measures of explicit knowledge, a set of criteria have been proposed for subjective measures, to ensure their validity: relevance, sensitivity, reliability, and immediacy (Moroshkina et al., 2019; Newell and Shanks, 2014). Relevance requires the explicit knowledge measure to test for knowledge that is relevant to the behaviour being tested. Sensitivity refers to the amount of explicit and implicit knowledge that can be accessed by the test. The reliability criterion requires the measure to not be biased by factors which do not also affect the behaviour being measured. Finally, the immediacy criterion stipulates that the awareness measure needs to be integrated within the behavioural measure or at least occur as soon as possible after the objective threshold measure has been conducted (Moroshkina et al., 2019; Newell & Shanks, 2014).

A further criterion, described as 'desirable' if not critical by Newell and Shanks (2014), is the criterion of non-reactivity. Non-reactivity refers to the need for subjective measures to not affect behaviour within the objective threshold measure, whether during learning or testing. As previously discussed, this is a key limitation of the 'decision strategy attribution' test. Non-reactivity is also something that is considered more seriously in areas of research that consider the verbal reporting of thinking for problem-solving (Fox et al., 2011). Moroshkina et al. (2019) directly considers the criteria of non-reactivity with the 'decision strategy attribution' test and argue that asking participants to reflect on where their knowledge is coming from (e.g., recollection or rule knowledge) may block the use of implicit knowledge.

Moroshkina et al. (2019) base this argument on the idea that reflection could aid the emergence of explicit knowledge of any underlying patterns in one of two ways. Firstly, in terms of the response options providing prompts to look for patterns. Moroshkina et al. (2019) propose that asking participants to reflect on their knowledge in this way prompts the use of internal verbalisation which can aid rule searching behaviour. Secondly, reflection and internal verbalisation could lead participants to alter strategy, utilising explicit rather than implicit memory. So, while participants may develop both explicit and implicit knowledge of underlying regularities, the reflection time may encourage reliance on explicit rather than implicit knowledge use. This change of strategy in participants may then affect their behaviour on the objective threshold measure (Moroshkina et al., 2019).

This following section will assess different subjective measures against Newell and Shanks' (2014) criteria along with the criterion of non-reactivity. This assessment aims to consider which measure would provide the most sensitive and valid measure of explicit knowledge for development within a grammatical paradigm.

3.3.1 Verbal reports.

Verbal Reports, although they can be informative, and have the benefit of being a quick and convenient measure, fail to meet many of the criteria laid out by Newell and Shanks' (2014). The open nature of verbal reports mean that relevancy is not always maintained as participants have a chance to consider and provide responses that are not relevant to the knowledge of interest. This potential for misjudgement of what is key and not key information by the participant also means it lacks reliability. Verbal reports may also lack sensitivity as verbalisation of explicit knowledge is not always possible (Dienes & Perner, 1999). This is based on the idea that there is not a strict dichotomy between implicit and explicit knowledge, rather there are different levels of explicit knowledge that a person can move through. For example, Karmiloff-Smith (1986) proposes an initial level of explicit knowledge from implicit where implicit procedural knowledge is re-coded to become accessible to different parts of the system but that this re-coding is still not conscious or accessible to verbalisation. If implicit knowledge is viewed as purely procedural based knowledge, as proposed by constructivist theories such as the 'chunk and pass' model

(Christiansen & Chater, 2005; see also section 1.2.1.1) than the ideas proposed by Karmiloff-Smith (1986; see also Dienes & Perner, 1999) suggests representations that are explicit but not accessible to verbal measures of this.

Sensitivity is also compromised by the retrospective nature of this measure, which allows participants to reflect on the experiences as a whole and reconstruct memories of knowledge based on this overview of their experience. The retrospective aspect of the task also prevents it from fulfilling the criteria of immediacy. Verbal reports however, are non-reactive as they are retrospective and objective threshold responses have already been made. These limitations of verbal report prompted suggestions for subjective measures criteria (Merikle & Reingold, 1992; Moroshkina et al., 2019; Newell & Shanks, 2014; Tunney & Shanks, 2003; Wierzchoń et al., 2012). Considering the verbal report against these criteria makes it clear that it is a suboptimal method for measuring explicit knowledge, particularly when considering the learning of grammatical patterns. The use of verbal reports within this area of statistical learning has provided a good start to investigations (Mirkovic et al., 2021; Monaghan et al., 2019 and Chapter 2 of this thesis), and some of the findings using this measure support those from word-boundary paradigms (Batterink et al., 2015; Franco et al., 2011). However, given that this method does not meet almost all the criteria for subjective measures other measures need to be considered.

3.3.2 The process-dissociation procedure.

The process-dissociation procedure is another subjective measure developed within the memory literature that also falls under the ‘attribution’ model of knowing (Jacoby et al., 1989; Lindsay & Kelley, 1996). As such this procedure utilises the concepts of ‘recollection’ and ‘feelings of familiarity’ as well as also incorporating ‘voluntary control’. The procedure was first devised and used by Jacoby (1991), where the term ‘recollection’ refers to memory test judgements being based on remembering the specific event or stimuli being tested (see also Kelley & Jacoby, 2000; Lindsay & Kelley, 1996). Thus, there is explicit knowledge that can be accessed, and voluntary control can be exerted over it (Dienes & Perner, 1999; Dienes, 2007; Dienes et al., 1995; Franco et al., 2011). The term ‘feelings of familiarity’ refers to memory test judgements being based on an intuitive feeling that the event/item has been encountered, i.e., the result of implicit processing. This intuitive feeling is based on the prior processing of this event/item implicitly facilitating current recall rather than recall from a specific memory (see also Kelley & Jacoby, 2000; Lindsay & Kelley, 1996). As such, there is no or lower explicit knowledge that cannot be or is harder to access, making it difficult to exert voluntary control over. Voluntary control is thus posed as something that can only be exerted over knowledge that a person has conscious awareness of, or in other words, knowledge that is explicit (Dienes & Perner, 1999; Dienes, 2007; Dienes et al., 1995; Franco et al., 2011).

The ‘process-dissociation procedure’ is designed to tease apart the use of ‘recollection’ and ‘feelings of familiarity’ by testing a participant’s ability to exert voluntary control over the knowledge they have gained. Franco et al. (2011) adapted Jacoby's (1991) procedure to work within a traditional statistical learning, word boundary detection paradigm. The authors created two artificial languages, containing artificial words made up of three artificial syllables, which were presented to participants in continuous streams. Across two experiments Franco et al. (2011) first made sure that adult participants could learn the transitional probabilities in the two artificial languages being used. Then in a third experiment, a different group of adult participants took part in an inclusion and exclusion task, the 'process-dissociation procedure', after being familiarised with both languages.

In both the inclusion and exclusion tasks, the participants were presented with trisyllabic 'word' strings from both languages along with novel, untrained trisyllabic strings. In the inclusion task the participants were asked to respond with 'yes' if the syllable string was from either artificial language or 'no' if it was a new 'word'. This is an old/new recognition judgement task that can be answered using both 'feelings of familiarity' and conscious 'recollection' judgements (Franco et al., 2011; Jacoby, 1991). The exclusion task however, required the participants to say 'yes' if the syllable string was from the first language they heard and 'no' if it was from the second language or a novel item. This requires participants to recognise 'words' from only one of the languages and thus 'feelings of familiarity' may impede success in this task, as conscious 'recollection' is needed to ensure a correct judgement. It thus assesses a participant's ability to assert voluntary control over statistically learned items (Franco et al., 2011; Jacoby, 1991).

Franco et al. (2011) found that only participants who had successfully learnt the words from the two artificial speech streams were able to distinguish between the two languages for the exclusion task, demonstrating exertion of voluntary control. Participants who had not successfully learnt the words could not do this. This result suggests that explicit knowledge is at least partially involved in statistical learning tasks with adults, given the proposal that voluntary control can only be exerted over explicit knowledge (Dienes et al., 1995; Dienes & Perner, 1996; Franco et al., 2016; Jacoby, 1991). This measure meets the criteria for subjective measures, in that it is measuring explicit knowledge use within the task helping to ensure immediacy, relevancy, reliability as well as increasing sensitivity (Newell & Shanks, 2014). It also has low reactivity as the measure does not refer to the underlying regularities, nor do these need to be referred to when explaining the task. However, like the 'remember/know' paradigm, this measure cannot distinguish between explicit item knowledge (recollection) and explicit knowledge of regularities. It also requires participants to be exposed to two stimulus sets or artificial languages.

3.3.3 Confidence ratings.

Confidence ratings are a versatile measure, overcoming many of the problems with retrospective verbal reports. This can be seen when considered with the confidence rating measures developed within the memory literature: the 'remember/know' paradigm and the 'decision strategy attribution' test. These confidence-based judgements are usually integrated within the objective threshold measure, being asked for after each trial or asked for once after the objective threshold measure has been administered. This ensures relevancy by asking about the test items themselves and its closer association with the items helps to ensure sensitivity and reliability. This placement of confidence rating measures also helps to meet the immediacy criterion.

While these properties of confidence ratings make them an inviting option for use, as discussed in section 3.2, these two versions of confidence ratings have limitations when considering the learning of regularities. Firstly, sensitivity and relevancy are reduced as a distinction between item and rule knowledge cannot be made when the 'remember/know' paradigm is used. Secondly, measures which aim to overcome this limitation (e.g., the decision strategy attribution test) then potentially violate the non-reactivity criterion (Moroshkina et al., 2019). As such, other confidence rating scales need to be considered. Those developed within the implicit learning literature may well work better here, given the aim of the literature to investigate the implicit learning of underlying patterns. There have been several confidence measures developed within this literature that can be considered and compared against the criteria for subjective measures.

3.3.3.1 The different versions of confidence ratings.

The implicit learning literature has developed a number of confidence rating systems to measure explicit knowledge, often drawing upon the perception detection literature. These systems include binary scales, continuous scales, Likert scales, feelings of warmth, wagering and rule (perception) awareness scales. The different confidence rating measures detailed here from the implicit learning and perception detection literatures are illustrated in Table 34 and 25.

Binary Scales simply ask participants to judge if they have high or low confidence (see Dienes & Seth, 2010; Kunimoto et al., 2001; Tunney, 2005; Tunney & Shanks, 2003 in Table 24 & 25). Continuous scales ask participants to indicate their confidence on a continuous 50-100 scale (see Dienes et al., 1995; Dienes & Scott, 2005; Tunney, 2005; Tunney & Shanks, 2003 in Table 24). Participants may be given a free choice over the number to choose, or given detailed instructions over what the numbers mean, or provided with numerical categories (Dienes, 2007).

Likert style scales can be number- and/or word-based but both forms provide participants with four to six confidence options (see Channon et al., 2002; Evans & Azzopardi, 2007; Maniscalco & Lau, 2012; Sandberg et al., 2010; Song et al., 2011; Tunney, 2005; Tunney & Shanks, 2003; Wierzchoń et al., 2012 in Table 24 & 25). When Likert scales combine the objective and subjective threshold response, they can be considered a binary scale. For example, Tunney (2005) uses a 1-4 Likert scale, options 1 and 2 provide an endorsement, with 3 and 4 providing a rejection. Within these endorsement and rejection response options, participants can provide the same high or low confidence responses, thus providing a binary scale for the confidence rating aspect of this measure (see Table 24).

The 'Feelings of warmth' scale was originally developed to assess a person's level of intuition regarding problem solving, but it has been argued to also tap into explicit knowledge (see Wierzchoń et al., 2012 for a review). This scale uses the term 'hot' and 'cold' at the extreme ends of the scale and may use intermittent labels like 'chilly' (see Table 24 for an example). Wagering was developed as a method of assessing confidence in young children and non-human animals. It was argued to be a more suitable measure of explicit knowledge in participants who were less able or unable to understand the concept of confidence (see Dienes, 2007; Dienes & Seth, 2010 for a review). Participants are asked to wager small or large amounts of tokens on the correctness of their objective threshold answers (see Dienes & Seth, 2010; Kunimoto et al., 2001 in Table 24 & 25). Finally, the perception awareness scale, which was developed in the perception detection literature, asks participants to judge the quality of their perception of a stimulus (see Overgaard et al., 2010; Sandberg et al., 2010 in Table 25). This task has been adapted for implicit learning tasks, becoming a rule awareness scale where participants are asked about their awareness of the underlying 'rule' (see Wierzchoń et al., 2012 for an example in Table 24).

While all scales are based upon the same or similar premise of ascertaining their confidence for an objective threshold response, they have been developed to meet specific task and question demands. However, most meet the subjective measure criteria in the same way the 'remember/know' paradigm and the 'decision strategy attribution' test do. The judgements are usually integrated within the objective threshold measure, being asked for after each trial or asked for once straight after these tasks are conducted. This ensures that these measures meet the immediacy criterion as well as relevancy, by asking about the test items themselves with the closer association to items helping to also ensure both sensitivity and reliability. In terms of Newell and Shanks' (2014) four criteria, all these confidence rating measures do well. However, while all measures are more sensitive than verbal reports, it has been argued that some confidence ratings are more sensitive than others.

3.3.3.2 Comparing confidence ratings for sensitivity.

Tunney (2005) and Tunney & Shanks (2003) directly compared binary and continuous confidence rating scales using an artificial grammar learning task and found the continuous scale to be more sensitive (see also Dienes, 2007). However, Dienes (2007) compared binary scales to the different types of continuous scales and found no difference in sensitivity between the scales. While Dienes (2007) used the same stimuli as Tunney (2005) and Tunney & Shanks (2003), they analysed the data using a potentially more sensitive analysis procedure, namely using signal detection analysis rather than comparison analysis to assess the zero-correlation criterion (see section 3.4.6 for a more detailed discussion). Thus, the sensitivity of these two measures may depend on how the data is analysed rather than the measures themselves.

Table 24. Empirical studies from the implicit learning literature.

Reference	Stimuli Type	Key participant information	Subjective Awareness Measure	Combined Scale	Scale Measures:
Channon, Shanks, Johnstone, Vakili, Chin & Sinclair (2002)	Artificial Grammar Learning	N=10 (Memory Disorder) *Mage*= 51.1 (SD=10.1) N=17 (Controls) *Mage*= 51.4 (SD=14.4)	**Confidence Likert Scale 1-6:** *1= certain endorsement* *2= fairly certain endorsement* *3= guess endorsement* *4= guess rejection* *5= fairly certain rejection* *6= certain rejection*	Yes	Awareness
Dienes, Altmann, Kwan & Goode (Exp. 1-3,5; 1995)	Artificial Grammar Learning	Adult volunteers Exp. 1: n=50 Exp. 2: n=40 Exp. 3: n=40 Exp. 5: n=34	**Exp. 1 & 5 - Confidence Continuous scale 50-100:** *50= complete guessing - 100= complete certainty* **Exp. 2 & 3 - Confidence Likert scale 1-5:** *1= complete guessing - 5= complete certainty*	No	Awareness
Dienes & Scott (2005)	Artificial Grammar Learning	Exp. 1: N=25 (adults) Exp. 2: N=80; *Mage* = 23.30 (SD=3.25)	**Exp. 1 & 2 - Confidence Word scale:** *guess* *intuition* *pre-existing knowledge* *rules* *memory* **Exp. 2 - Confidence Continuous scale 50-100:** *50= complete guessing - 100= complete certainty*	No	Awareness

Table 24 Continued:

Reference	Stimuli Type	Key participant information	Subjective Awareness Measure	Combined Scale	Scale Measures
Tunney (2005)	Artificial Grammar Learning	N=42, *Mage*=19.42 (SD=2.37)	**Confidence Likert/Binary scale 1-4:** *1= yes conforms to rules – more confidence* *2= yes conforms to rules – less confident* *3= no does not conform to rules – more confident* *4= no does not conform to rules – less confident* **Confidence Continuous scale 50-100:** *50= complete guessing - 100= absolute certainty*	4-point scale: Yes Continuous Scale: No	Awareness
Tunney & Shanks (2003)	Artificial Grammar Learning	University Community **Exp. 1:** N=24 **Exp. 2:** N=20 **Exp. 3:** N=20 **Exp. 4:** N=36	**Exp. 1 & 2 - Confidence Likert/Binary scale 1-4:** *1= yes conforms to rules – more confidence* *2= yes conforms to rules – less confident* *3= no does not conform to rules – more confident* *4= no does not conform to rules – less confident* **Exp. 1 - Confidence Binary response:** *More confident or less confidence* **Exp. 3 - Confidence Continuous scale 50-100:** *50= complete guessing - 100= absolute certainty*	4-point scale: Yes Others: No	Awareness

Table 24 Continued:

Reference	Stimuli Type	Key participant information	Subjective Awareness Measure	Combined Scale	Scale Measures:
Wierzchon, Asanowicz, Paulewicz & Cleeremans (2012)	Artificial Grammar Learning	Undergraduate students **Exp. 1:** N=217	**Confidence Likert Scale 1-4:** *1 = I am guessing* *2 = I am not confident* *3= I am quite confident* *4 = I am very confident* **Post-decision Wagering scale:** *Asked to wager one of 4 polish (PLN) monetary stakes on the correctness of their response (20, 40, 60 or 80 PLN). All participants started with 200 PLN.* **Feelings of Warmth scale:** *Based on feelings of intuition* *1 = cold* *2 = chilly* *3 = warm* *4 = hot* **Rule Awareness Scale:** *Asked if aware of the rule underlying their accuracy response* *1 = I do not have any clue about the rule* *2 = I have a glimpse of the rule* *3 = I think that I know what the rule is* *4 = I know the rule*	No	Awareness

Table 25. Empirical studies from the perception detection literature.

Reference	Stimuli Type	Key participant information	Subjective Awareness Measure	Combined Scale	Scale Measures:
Kunimoto, Miller & Pashler (2001)	Detecting visual stimuli (words)	Undergraduate students **Exp. 1:** N=40 **Exp. 2**: N=18 **Exp. 3:** N=64 **Exp. 4:** N=36	**Wagering:** *Betting chips when they thought they were correct.* **Binary high/low rating:** ***Exp 1:*** *Half of the participants judged after each block and half after each trial.*	No	Awareness
Maniscalco & Lau (2012)	Visual/spatial	N=30 Undergraduate students	**Confidence Likert Scale 1-4**	No	Awareness
Sandberg, Timmermans, Overgaard & Cleeremans (2010)	Visual	N=36, *Mage*=23.9 (range=22.3-25.5)	**Confidence Likert Scale, 1-4**: *1= Not confident at all* *2= Slightly confident* *3= Quite confidence* *4=Very confidence*	No	Awareness
Song, Kanai, Fleming, Weil, Schwarzkopf & Rees (2011)	Visual	N=18, range=19-33 years	**Confidence Likert Scale, 1-6**: *1= low confidence - 6= high confidence*	No	Awareness

Dienes & Seth, (2010) also used an artificial grammar learning paradigm to compare binary and wagering methods of assessing confidence. For binary judgements, participants were asked if they were 'guessing' (akin to flipping a coin) or 'sure to some degree' after each objective threshold response. In the wagering condition, participants started the testing task with 10 sweets and after each objective threshold response they were asked if they were willing to wager 1 or 2 of their sweets. Whether a participant won or lost those sweets were not revealed until the end of the task. The authors found that these two scale types have a similar sensitivity for detecting explicit knowledge. Within the perception detection literature, Sandberg et al. (2010; also reviewed in Overgaard et al., 2010) conducted a similar investigation but used a Likert scale rating system rather than a binary one. However, they came to a similar conclusion as Dienes & Seth, (2010) in that neither scale was more sensitive than the other.

Wierzchoń et al. (2012) used an artificial grammar paradigm to compare several subjective measures: Likert scale, wagering, feelings of warmth, rule awareness scale and a continuous scale like the rule awareness scale. While the authors found that all the measures detected explicit knowledge, there were some sensitivity differences between them. They found that the Likert confidence rating measure detected the broadest range of awareness, from implicit to explicit, while also being able to capture the lowest level of implicit knowledge. Other measures showed a smaller range of detection across different levels of implicit and explicit knowledge, although these ranges usually fell within levels of explicit knowledge. Although, despite detecting a smaller range, the feelings of warmth measure and the rule awareness scale demonstrated a greater degree of sensitivity within this range. So, they were able to detect more levels of awareness within these smaller ranges than measures which captured a broader range but with less sensitivity within that broader range (see also Norman & Price, 2015 for a summary). From this Wierzchoń et al. (2012) recommends using a Likert confidence rating scale along with the feelings of warmth and/or rule awareness scale to obtain the most broad and sensitive measure of awareness.

Taking this all together it seems that in terms of sensitivity most of the measures reviewed are broadly equivalent, although Wierzchoń et al.'s (2012) findings suggest that Likert style confidence ratings may give the broadest range of awareness detection, but with some finer grained detection loss when measuring higher levels of awareness (e.g., explicit knowledge). Using both the Likert scale with a feelings of warmth measure or the rule awareness scale could potentially improve this sensitivity.

3.3.3.3 Confidence ratings and non-reactivity.

Using a Likert version of the confidence ratings measure along with a feelings of warmth measure or rule awareness scale may be the most sensitive option. However, this may be problematic in terms of the non-reactivity criterion. The 'decision strategy attribution' test is potentially reactive as the options given explicitly refer to 'rule knowledge'. As the rule awareness scale asks participants whether they are aware of an underlying rule, this task is open to similar reactivity issues and thus would not be a good option here for the same reasons. A feelings of warmth task could then be used instead to increase sensitivity. However, the use of two subjective measures could also raise issues of reactivity.

Moroshkina et al. (2019) argue that the 'decision strategy attribution' task is reactive due to the reflection it prompts in participants on their knowledge. It is this reflection that could support the emergence of explicit knowledge and thus make a measure reactive. All subjective measures, including confidence rating measures require some form of reflection on one's own knowledge, that means reactivity will always be a potential issue within these measures. Despite this, efforts can and should be made to reduce this reflection and consequently potentially reduce reactivity. Including two subjective measures (e.g., Likert scale and feelings of warmth), provides two opportunities for reflection which in turn increases the potential reactivity of the subjective measures being used. This then suggests that to reduce reactivity, it would be preferable to only use one confidence rating measure. Since the Likert scale is proposed to measure the broadest range of awareness (Norman & Price, 2015; Wierzchoń et al., 2012) compared to the feelings of warmth measure, this style of confidence rating would be the most appropriate.

Using just one measure can help to reduce reactivity, with the use of a Likert confidence rating scale helping to ensure sensitivity. However, the use of this scale may still fall foul of the issue of distinguishing between knowledge of specific items vs regularities. If the Likert scale options need to avoid referring to the underlying regularity, then any explicit knowledge it may detect could be based on either explicit item or regularity knowledge. This is a pertinent issue for paradigms and/or measures that test knowledge with the items participants have been trained on, as is commonly done in artificial grammar paradigms (e.g., A. S. Reber, 1967), word-boundary paradigms as well as in the Monaghan et al. (2019) study discussed previously.

Alternatively, this issue could be overcome by paradigms which use different items in the training and testing tasks. The statistical learning paradigm used in this thesis tests knowledge of grammatical regularities through generalisation tasks using novel items that have not been encountered within the training tasks. Within this paradigm then a Likert confidence rating scale cannot be influenced by item recollection knowledge from the training tasks. This means that for a paradigm using generalisation tests, Likert confidence responses indicating explicit knowledge could only be based on knowledge of the underlying grammatical regularities.

Paradigms which consider grammatical knowledge in terms of generalisation then do not require a subjective measure to distinguish between explicit 'item' or 'regularity' knowledge. As such, within these paradigms subjective measures do not need to explicitly refer to the underlying 'rule' negating the need for measures such as the 'decision strategy attribution' task, which risk increasing the reactivity of the subjective measure. These generalisation paradigms can then adopt a more sensitive Likert confidence rating scale to better meet the non-reactivity criterion.

3.3.4 Implications for developing a subjective measure within a grammatical paradigm.

Overall, confidence rating measures of explicit knowledge have proven a more valid measure than verbal reports when considered against the criteria for subjective measures (Newell & Shanks, 2014). All confidence rating measures considered above meet the four main criteria put forward by Newell and Shanks (2014), however the type of confidence rating measure needs to also be considered carefully in relation to the criterion of non-reactivity.

To avoid the issues of reactivity, particularly highlighted by the 'decision strategy attribution' task (Moroshkina et al., 2019), confidence ratings need to avoid referring to the underlying rules. Additional efforts also need to be made to reduce opportunities for reflection by participants to reduce the chance of reflection aiding in the emergence of explicit knowledge. Thus, whilst some literature may suggest using multiple measures of explicit knowledge to increase sensitivity (e.g., Wierzchoń et al., 2012), this would come at the expense of non-reactivity due to the need of some reflection with any subjective measure. As such the use of only one confidence rating measure within a task would be preferable.

As reference to the underlying regularities should be avoided, efforts then need to be made to avoid issues of knowledge distinction highlighted by the 'remember/know' paradigm in section 3.2. This can be done using grammatical generalisation paradigms used in this thesis where the testing items are different from the training items. Here, if explicit knowledge is detected it can only be based on regularity knowledge, as item recollection knowledge from the training tasks would not aid responses here.

3.4 Developing a measure for a grammatical generalisation paradigm.

The conclusions from the discussion of the literature above will be used to develop a new subjective measure of explicit knowledge for use within the grammatical generalisation paradigm used in this thesis. The aim is to develop a measure that meets the stated criteria for subjective measures (Moroshkina et al., 2019; Newell & Shanks, 2014) and overcomes previous issues of reactivity within grammatical paradigms of statistical learning (Monaghan et al., 2019). It is hoped that this exercise will produce a sensitive and robust behavioural measure to better investigate the theoretical underpinnings of statistical learning, particularly in terms of the dual-system proposal (Conway, 2020; R. Frost et al., 2019).

The artificial language used by Mirkovic et al. (2021) and in Chapter 2 of this thesis incorporated both vocabulary and grammatical structure. A cross-situational task was used to both train and test participants for both vocabulary and grammatical knowledge. The language was based on grammatical gender and comprised novel nouns belonging to one of two semantic categories, animals, and artefacts. Nouns in these two categories were cued using deterministic determiners and suffixes. For example, animal nouns were indicated by the determiner *'tib'* and suffix *'eem'* (dog = *tib* mof*eem*), with nouns referring to artefacts by *'ked'* and *'ool'* (table = *ked* larsh*ool*).

Participants were trained on a set of the nouns in a word-picture matching task, where a picture and 'word' were presented simultaneously, and participants were asked to decide if the word and picture '*go well together*'. Trials consist mostly of correctly matched words and pictures with some incorrectly matched items and participants indicated their decision by pressing either a sad or happy face. Participants' grammatical knowledge was then tested in word-picture matching generalisation tasks, where participants still had to decide if words and pictures went well together. However, the items encountered were untrained and were either consistent or inconsistent with the grammatical regularities present in the nouns used for training. This use of new items within the testing task means grammatical knowledge is tested through the generalisation of grammatical regularities. This section will describe how a subjective measure can be integrated into this grammatical generalisation paradigm using the recommendations drawn from the existing subjective measures literature.

3.4.1 Consideration 1: Where to place the subjective measure.

To meet the subjective measures criteria (Newell & Shanks, 2014), the subjective measure being developed will be incorporated into each item of the training and generalisation word-picture matching task used in this thesis. This would ensure the immediacy criterion is met and would also help to ensure the subjective measure is asking about the test items themselves which helps to meet the relevancy criterion. The close association with individual items would also help to increase sensitivity and reliability. The word picture matching task requires participants to judge if a presented word and picture go well together, or if a word goes well with the artificial language the participant has been learning. To make this judgement, in previous studies the participant pressed a happy (it does go well with) or sad (it does not go well with) face button. It is this response that can be adapted to include a subjective measure.

3.4.2 Consideration 2: Ensuring the subjective measure meets Newell and Shanks's (2014) criteria and reduces reactivity.

To minimise reactivity, confidence ratings can be used in preference to a 'decision strategy attribution' test. The use of a grammatical generalisation paradigm also follows the proposed recommendations from section 3.3 as it means the testing tasks use different items from the training task. This means if the confidence rating task detects explicit knowledge, this can only represent explicit grammatical knowledge as recollections of items from the training task cannot be used to aid responses here.

These considerations also rule out the use of the 'remember/know' paradigm and 'process-dissociation procedure', despite their successful use within statistical learning paradigms (Batterink et al., 2015; Franco et al., 2011) and their lack of reference to the underlying regularities. Considering the 'process-dissociation procedure' first, to incorporate this measure a second artificial language would need to be developed. While an inclusion task could easily be developed using the existing language, the exclusion task which is the real measure of voluntary control, requires a second artificial language. This would involve increasing the length of an already long training phase, making its inclusion into this current grammatical generalisation paradigm inappropriate.

The 'remember/know' paradigm has been developed for use on stimuli that are present both in training and testing and in its current form it would not be suitable for a generalisation paradigm. It could be adapted for the training word-picture matching task, where participants are encountering training items multiple times. However, it would be harder to adapt it for the generalisation word-picture matching task, where knowledge of regularities is being tested on new, unencountered items. This measure could be adapted for the generalisation tasks by recontextualising the concepts of 'remember' and 'know'. However, this would mean participants would need to use two slightly different meanings for these responses between the training and generalisation versions of this task. This might then reduce the sensitivity and reliability of the scale, two of the criteria for successful subjective measures of awareness (Newell & Shanks, 2014).

Adapting the 'remember/know' paradigm also has the potential to increase reactivity within the generalisation testing task. To use the measure, efforts would need to be made to ensure participants understand the concepts of 'remember' and 'know' within the context of generalisation. They would also be asked to directly reflect on their knowledge, to assess if it is recollection or familiarity based. So, while the 'remember/know' paradigm options do not specifically refer to rule/pattern knowledge, it still falls victim to the same limitations as the 'decision strategy attribution' test. This is due to the instructional language needing to hint at or refer to regularities in some way along with the type of reflection required potentially prompting explicit knowledge emergence and strategy use (Moroshkina et al., 2019). Taking this into account, along with the potential to violate the reliability and sensitivity criteria (Newell & Shanks, 2014) this confidence measure will not be used here. Instead, a confidence rating measure developed within the implicit learning literature will be used.

Given the recent integration of the implicit and statistical learning literature (Batterink et al., 2019; Conway, 2020), a confidence rating measure developed within the implicit learning literature is likely to translate better into the grammatical generalisation paradigm being considered here (Mirkovic et al., 2021 and Chapter 2 of this thesis). As recommended by the literature, only one style of confidence rating will be used to avoid increasing opportunities for reflection and thus increasing reactivity (Moroshkina et al., 2019). Also recommended by the literature (section 3.3.3) a Likert confidence rating scale will be used as the measure that captures the broadest range of awareness (Wierzchoń et al., 2012) helping to increase the sensitivity of the measure being developed here.

Even following these recommendations, there is a contextual issue that could increase reactivity that is not resolved by using a Likert confidence rating measure or a grammatical generalisation paradigm. It is the same issue that necessitates the use of only one measure that was discussed in section 3.3.3. Any version of a subjective measure will require some knowledge reflection in order for a response to be given. Time taken to reflect on one's knowledge in any form, could also prompt the emergence of explicit knowledge and strategy use (Dienes, 2007) which means there is always an issue of reactivity within subjective measures (Moroshkina et al., 2019). Efforts can be made to reduce reactivity by reducing reflection time and opportunities, as is achieved by only using one subjective measure. A further step would be to integrate the response for the objective and subjective measures.

The usual protocol for subjective measures provides the participant with two opportunities to reflect per test item: first when they make their objective threshold response, and then when they make their subjective awareness response (see Table 23, 24, 25 & 26). Combining these responses, so participants provide both an objective and subjective threshold response at the same time would further reduce the opportunity for reflection. In the context of the word-picture matching task that is being adapted, it would mean participants are judging whether a word and picture go well together at the same time as judging their confidence for this word-picture matching judgement. By designing a combined response, this would reduce both the number of opportunities to reflect on knowledge as well as the overall amount of time spent reflecting on an item. Additionally, it may also help to better capture feelings of confidence that are transient and fleeting and so need to be captured quickly before they decay (Norman & Price, 2015; Tunney & Shanks, 2003).

3.4.3 Consideration 3: Choosing the right Likert scale for the job.

A Likert confidence rating scale is likely to be the most sensitive in terms of breadth (Wierzchoń et al., 2012), and would also be relatively easy to integrate into the word-picture matching response protocol used within this thesis. It can be developed to allow participants to endorse or reject word-picture pairings while also indicating a level of confidence for example, '*I have low/high confidence the word and picture do not go well together'* followed by '*I have low/high confidence the word and picture do go well together'*. There are several different Likert scale designs within the literature (see Table 23, 24, 25 & 26) that could be drawn from here. However, this section will focus on examples of combined objective and subjective threshold Likert scales, to help support measures to reduce reactivity (section 3.4.2). While there are only a few examples in the literature of combined scales, Tunney (2005) and Tunney & Shanks (2003) are examples of a combined Likert scale being used successfully. Both studies include the use of a four-point scale, which allows participants to endorse or reject an item as the objective threshold response. Then within the endorse and reject responses participants can indicate if they have high or low confidence in their objective threshold response (see Table 24). Their measure can be both a Likert and binary measure as while a four-point scale is used, the options only allow for a high or low confidence judgement within an endorsement or rejection response. Given the small number of studies that use a combined scale, it would be optimal to use the binary/Likert protocol that has been tried and tested in this regard.

While an example of successful use of a combined scale, Tunney (2005) and Tunney & Shanks (2003) used wording that referenced the underlying rules, something that needs to be avoided to reduce reactivity. Channon et al., (2002) did develop a six-point scale for use within an artificial grammar paradigm that does not reference the underlying rules (see Table 24). Channon et al. (2002) used the concept of 'certainty' in the scale wording to combine the objective and subjective measure. For example, *'certain endorsement*', '*fairly certain endorsement*' and '*guess endorsement*' with the same certainty options for reject options as well. This concept enables scale labels that direct participants to make a confidence judgement as well as a first order state judgement all in one response (e.g., *'certain endorsement'* or *'fairly certain rejection'*). Further to this, while this scale's language is directing participants to make both an objective and subjective response, it does so without referring to the knowledge they are being asked to make an objective and subjective judgement on.

Channon et al.'s (2002) use of 'certainty' helps to easily incorporate a combined Likert confidence rating into the word-picture matching task used by Mirkovic et al. (2021) in Chapter 2 of this thesis. In this task participants are asked to judge whether the word and picture *'go well together'* or if a word *'goes well with'* the artificial language that has been encountered. Within previous studies, this context has helped to disguise the grammatical aspect of the generalisation word-picture matching tasks, as it looks the same to the participant as the word-learning based task. The concept of 'certainty' can work within this established phrase to maintain this, for example *'I am certain the word and picture go well together'* or *'I am certain the word and picture do not go well together'.* This further helps to reduce reactivity. Thus, the Likert confidence rating scale developed by Channon et al. (2002) provides a strong basis from which to develop a scale that helps to reduce reactivity in two ways: firstly, by reducing knowledge reflection by combining the objective and subjective response, and secondly by using language which is less likely to prompt rule searching or strategy change (section 3.2).

3.4.4 Consideration 4: Creating a child-friendly scale.

While combining the objective and subjective responses can reduce reactivity, suitability for use with children is also a motive behind the development of this Likert confidence rating scale design. Separate objective and subjective responses are likely to increase the cognitive and memory demands on participants, a factor which may influence a child's ability to respond appropriately. In terms of the subjective measures criteria (Newell & Shanks, 2014), this could then reduce the relevancy and reliability of the subjective measure when used with children. Creating a combined response scale using child-friendly language is a way to reduce the cognitive demand on child participants. There is little existing research with children using subjective measures within an implicit learning paradigm, such as artificial grammar learning. However, child research from native language processing and memory studies can inform the design of child-friendly measures.

3.4.4.1 Can children use one scale to measure two concepts? Assessing the combined objective and subjective threshold measures with evidence from studies of language development.

Tasks used in studies of language development are a useful source for designing a child-friendly subjective measure. The studies described here all use grammatical judgement tasks, asking child participants to decide if stimuli are grammatical or not. These tasks are transparent about testing for grammatical knowledge, unlike the grammatical testing task being adapted from this thesis. In other respects, these native language grammatical judgement tasks are essentially asking for a similar judgement to the generalisation word-picture matching task. Thus, while these studies use Likert scales to measure acceptance of grammatical utterances rather than awareness, they support the use of Likert scales that measure two concepts within a grammatical judgement task. The key properties of the studies discussed below are presented in Table 26.

Katsos and Bishop (2011) and Katsos and Smith (2010) developed a Likert scale response protocol from a binary one for measuring children's (5-6- and 6–7-year-olds) pragmatic tolerance. Child participants were presented with a caveman character who is trying to learn English and they were asked to judge his grammatically correct or incorrect utterances on either a 3- or 5-point scale. These studies found that these Likert scales could be used by children and produced similar results in pragmatic judgements to binary, right/wrong response protocol. These findings support the suitability of combined Likert confidence ratings when making grammatical judgement for children but also for the transition from binary to scale responses.

However, these scales are slightly different to the Likert confidence rating scale proposed for measuring explicit and implicit knowledge use within grammatical generalisation. These studies use a reward-based system for participants to judge their acceptance of the utterance rather than directly rating the utterance. The participants are asked to decide on the reward they would give to the caveman based on 'how well' the caveman did, in terms of reward size (Katsos & Bishop, 2011) or number of rewards (Katsos & Smith, 2010; see Table 26 for details). So, participants are indirectly making grammaticality judgements via thinking about how to reward the utterer. Davies et al. (2016) and Pipijin and Schaeken (2012) successfully adapted this reward scale to a more direct happy and sad face scale. These studies were also investigating pragmatic tolerance in children (5-10- and 11–13-year-olds) and used a similar grammaticality judgement task with participants. These adapted happy/sad face scales asked participants to make a direct good/bad judgement about an utterance rather than indirectly rewarding the utterer. These scales are still assessing graded acceptance rather than explicit knowledge, but the more direct nature

Table 26. Empirical child studies from linguistics.

Reference	Stimuli Type	Key participant information	Subjective Awareness Measure	Combined Scale	Scale Measures:
Davies, Andrés-Roqueta & Norbury (2016)	Native Natural Language (Spanish)	N=18 (SLI) *age*=5;0– 10;11 N=18 (Controls) *age*=matched for chronological age	**Likert Scale using happy/sad faces:** *1= very good (muy buena)* *2 = good (Buena)* *3 = regular* *4= bad (mala)* *5= very bad (muy mala)*	No	Acceptance
Katsos & Bishop (Exp. 2; 2011)	Native Natural Language (English)	**Exp. 2:** N=18 Mage=5;8 (range=5;3-6;3) N=10 Mage=22;1 (range=19;9-25;1)	**Likert Reward scale using Strawberries:** *1 = small strawberry* *2 = big strawberry* *3 = huge strawberry*	No	Acceptance
Katsos & Smith (Exp. 2; 2010)	Native Natural Language (Dutch)	Exp. 2: N=30 Mage=7;1 (range=6;6-8;3) N=30 Mage=10;2 (range=8;7-11;5)	**Likert Reward scale using Strawberries:** *1 = give 1 strawberry* *2 = give 2 strawberry* *3 = give 3 strawberry* *4 = give 4 strawberry* *5 = give 5 strawberry*	No	Acceptance
Pipijn & Schaeken (2012)	Native Natural Language (Dutch)	N=22, *Mage*=11;3 (range=11-13)	**Likert Scale using happy/sad faces, 1-5:** *Only ends of the scale are labelled with a frowning face (left) and smiling face (right).*	Yes	Preciseness

brings it closer to the *'does/does not go well with'* judgements in the measures being developed here.

Both studies use a five-point Likert scale, although Pipijin & Schaeken (2012) only label the ends of the scale with a happy or sad face and Davies et al. (2016) label each scale point with both words and faces (see Table 26). Pipijin & Schaeken's (2012) found that their less detailed scale was not as sensitive as the reward scale. In contrast, Davies et al. (2016) found that their more detailed scale was more sensitive than a reward-based scale. Thus, while these studies support the suitability of combined response scales for children, Davies et al. (2016) findings suggest that information for each point of the scale is needed to help support child participants in using the scale. The authors also further confirmed the suitability of using a scale like this with children, through finding it to be accessible even for participants with developmental language disorder.

These findings from native language processing studies support the suitability of a combined Likert judgement scale with children. This supports their use for the grammatical decision element within Mirkovic et al. (2021) and Chapter 2's statistical learning paradigm. However, these studies are measuring pragmatic tolerance rather than awareness, so while children can make grammatical decisions using Likert scales can they also detect explicit knowledge use within children?

3.4.4.2 Can a Likert confidence rating scale measure explicit knowledge in children? Support from the memory literature.

Lockl and Schneider (2002) and Visser et al. (2014) both assessed explicit knowledge using Likert scales to explore the concept of 'feelings of knowing'. Feelings of knowing considers a person's meta-cognitive abilities, in this case around the awareness they have about what knowledge they have (Lockl & Schneider, 2002; Visser et al., 2014). 'Feelings of knowing' are fundamentally concerned with how well a person can assess what explicit item/event knowledge they have and do not have. So, unlike the 'remember/know' paradigm discussed earlier in relation to the memory literature, it is not asking whether knowledge is implicit or explicit, or in other words if second-order states have developed or not. Rather it is asking whether there are correct second-order states present for first-order states that may or may not exist. As such, the theoretical underpinnings diverge slightly from the concepts discussed so far in this review; however, there is some crossover in terms of subjective threshold measures for explicit knowledge.

Visser et al. (2014) investigated 'feelings of knowing' with general knowledge in children aged eight- and eleven-year-olds and Lockl and Schneider (2002) with native language vocabulary knowledge in children aged seven to ten years-old. Both studies used a Likert scale which incorporated happy and sad faces to measure levels of awareness. Lockl and Schneider, (2002) used a one-to-four scale and Visser et al. (2014) a one–to-five scale, both based on the concept of certainty. The sad faces were used to indicate uncertainty and the happy faces to indicate certainty (see Table 23). After making an objective threshold response, child participants were asked to use this scale to judge how certain they were that their response was correct. The scales used here were suitable for the child participants and were able to indicate that some children have awareness about what they know and do not know, and that this is linked to their performance accuracy (Lockl & Schneider, 2002; Visser et al., 2014). These findings support the use of a Likert scale to assess awareness in children and its potential to indicate the use of explicit knowledge within the statistical learning paradigm used in this thesis.

Lockl and Schneider, (2002) and Visser et al. (2014) did not specifically use the confidence rating scale to indicate the use of explicit knowledge within an implicit knowledge paradigm. However, as discussed, they investigated awareness through closely related concepts of first and second-order states and subjective and objective thresholds. The scale they used is also very similar to the Likert confidence rating scales that have been developed to assess explicit knowledge use within implicit learning paradigms (e.g., Channon et al., 2002; Tunney, 2005; Tunney & Shanks, 2003; Wierzchoń et al., 2012). This is particularly pertinent when considered with Channon et al.'s (2002) use of certainty to indicate confidence within adults, a concept both Lockl and Schneider (2002) and Visser et al. (2014) use. The use of a Likert confidence rating to assess the presence of explicit and/or implicit knowledge in children will need to be directly tested. However, related literature supports its use as a suitable measure of explicit knowledge indication within children.

3.4.4.3 Incorporating child-friendly considerations.

To sum up, studies of native language development support the suitability of Likert scales with children for measuring grammatical knowledge, when the scale is also assessing another concept (in this case pragmatic tolerance). This suggests a combined objective and subjective Likert response scale would be suitable for child participants. The memory-based studies exploring 'feelings of knowing' also support the suitability of using Likert confidence rating scales to assess explicit knowledge use within children. All studies drawn upon here use a word-based Likert scale but with visual aids, usually images of happy and sad faces. Thus, any response scales incorporated into word-picture matching tasks used in this thesis should also include this visual support. The previous binary response protocol already included a happy (endorsements) and sad face (rejections), so incorporating this into the new scale would also help to minimise the changes made to the task as well as aiding comparisons between studies.

The use of a certainty concept by Lockl and Schneider, (2002) and Visser et al. (2014) relates well to its use to indicate confidence in adults by Channon et al. (2002). It can then further support the use of this concept as a child-friendly way to avoid referring to the knowledge of the underlying regularity that is being assessed, in an effort to reduce reactivity. However, the findings from Davies et al. (2016) and Pipijin and Schaeken (2012) suggest that just labelling the ends of the scale (e.g., *'I am certain the picture and word go well together' or 'I am certain they do not go well together'*) may not provide enough sensitivity for use with children. While Channon et al. (2002) uses '*fairly certain'* to indicate the midpoint between certain and guessing, this wording may not be as suitable for children as it is for adults. Using the word *'think'* instead can indicate this less certain mid-point while being more child-friendly (e.g. *'I <u>think</u> the picture and word go well together' or 'I <u>think</u> they do not go well together'*).

3.4.5 Consideration 5: How the 'guessing criterion' and 'zero-correlation criterion' analysis shape the Likert confidence rating scale.

A resulting benefit of Likert confidence rating scales and the two analysis criteria (discussed in detail in section 3.2.1) is an underlying assumption that both explicit and implicit knowledge can be present (Dienes, 2007; Moroshkina et al., 2019). This is helpful for its use within the statistical learning paradigm used in this thesis, as previous studies looking at the role of explicit and implicit knowledge within statistical learning have found that adults use both (e.g., Batterink et al., 2015). This previous research was conducted in a word-boundary paradigm, thus using confidence ratings within a grammatical paradigm will help to show if this finding extends to grammatical regularity learning as well.

To gain this benefit, the adapted confidence rating scale should include some form of 'know' option(s) to express confidence and a 'guess' option to support the use of the two analysis criteria first described in section 3.2.1: zero-correlation and guessing criterion. Zero-correlation analysis can still be conducted if a 'guess' option is not included, but it has been argued that to get the full range of meta-knowledge, both are needed (Dienes & Perner, 1996). In relation to this, while the zero-correlation criterion can detect explicit knowledge by finding a relationship between accuracy and confidence, its ability to detect implicit knowledge is only indirectly implied by not finding this relationship. This leaves the criterion open to potentially overinterpreting null results when used to detect implicit knowledge (Dienes, 2015). It is then best to use the zero-correlation criterion mainly as a measure of explicit knowledge and the guessing criterion as the main measure for implicit knowledge.

To incorporate this, the scale used by Channon et al. (2002) can be drawn upon again here as it includes a guess option, as well as two 'know' options (low and high confidence) through using the certainty concept. Using this scale as a basis can thus make it possible to analyse the results using the guessing and zero-correlation criteria, as well as providing the basis for a combined scale which helps reduce issues of reactivity.

3.4.6 Conducting the analysis: The best fit for the paradigm.

The guessing criterion is relatively simple to calculate as accuracy is compared against chance performance for all guess responses. If participants are performing above chance when they are guessing, then the objective threshold has been met but the subjective threshold has not been, so the knowledge being used must be implicit (Rebuschat, 2013). In relation to the discussion in the previous section (3.4.5), this allows for the detection of implicit knowledge as it directly considers its presence within the accuracy of guess responses only. The 'zero-correlation' criterion can be measured using the 'Chan difference score' (Chan, 1992). This can be done by calculating the mean confidence score for correct and incorrect responses and then comparing these. If the mean confidence rating is significantly higher for correct than incorrect responses, this is an indication that explicit knowledge is being used as it indicates the presence of a relationship. Conventionally, if there is not a significant difference in mean confidence ratings between correct and incorrect response, this suggests that explicit knowledge is not being used, and above-chance performance should therefore be due to implicit knowledge (Rebuschat, 2013). However, as discussed above, the interpretation of a null result here needs to be considered carefully as it is open to potential overinterpretation (Dienes, 2015). As such, this particular analysis procedure will be used to focus on the detection of explicit knowledge.

These analysis procedures are recommended by Rebuschat (2013) for research investigating the role of explicit and implicit language knowledge during second language learning. These recommendations are relevant for the statistical learning paradigm used in this thesis, as while it uses an artificial language, the learning context for participants is more akin to second language learning than first language acquisition or native language processing. This is also the analysis protocol used in Channon et al.'s (2002) artificial grammar study, whose confidence rating scale has proved to be a strong basis upon which to develop the confidence rating scale for the paradigm used in this thesis. Using this analysis protocol also means no further changes or adaptations are needed for the scale or methodology as a whole.

It should however be noted that there are other, potentially more sensitive procedures that could be used (Norman & Price, 2015). The main alternative is signal detection analysis. This method proposes calculating two forms of *d'*, type one and type two, which relate to the objective (type one) and subjective (type two) thresholds (Barrett et al., 2013; Evans & Azzopardi, 2007; Kunimoto et al., 2001; Maniscalco & Lau, 2012; Norman & Price, 2015). Using a type two *d'* to calculate 'zero-correlation' is thought to overcome issues of response bias caused by individual difference in scale interpretation and option preferences (Kunimoto et al., 2001; Rebuschat, 2013). A new *meta-d* calculation has even been developed to further reduce the influence of response bias still found within type two *d'* (Barrett et al., 2013; Evans & Azzopardi, 2007; Maniscalco & Lau, 2012; Norman & Price, 2015).

While using a signal detection protocol would be favourable in terms of sensitivity, it is not a 'best fit' for the paradigm used within this thesis. Most of the research using this method, particularly when it comes to using *meta-d*, has been conducted within the perception detection literature (Barrett et al., 2013; Evans & Azzopardi, 2007; Kunimoto et al., 2001; Maniscalco & Lau, 2012). These studies typically use a large number of test items or trials (e.g., Maniscalco & Lau, 2012 uses 1000 trials) to gather the amount of data needed for this type of analysis, far more than is currently used within the generalisation, word-picture matching testing tasks (eight items per testing task). To use this analysis, the number of test items would need to be increased by an unrealistic amount. It is also worth noting that while Rebuschat (2013) discusses response bias and the use of signal detection for analysis, they do not recommend it for use in second language learning research. Therefore, whilst potential response bias confounds will need to be acknowledged and considered when interpreting results, using Channon et al.'s (2002) analysis protocol and Rebushat's (2013) recommendations is the best fit for the statistical learning paradigm used in this thesis.

3.5 Conclusions and final Likert confidence rating scale.

The recent merging of the implicit and statistical learning literature (e.g., Batterink et al., 2019; Conway, 2020; Frost et al., 2019; P. J. Reber et al., 2019) has prompted investigations into the assumption of implicitness in statistical learning research (Batterink et al., 2015; Franco et al., 2011; Monaghan et al., 2019). The current review has considered the relevant literature to guide the development of a subjective measure of explicit knowledge for use with the grammatical paradigm of statistical learning used in this thesis.

In the next chapter (Chapter 4), the new measure will be used to extend previously reported findings of explicit knowledge use (Batterink et al., 2015; Franco et al., 2011) into a more complex statistical learning paradigm, while also overcoming the limitations of the verbal reports used in recent studies of grammar-learning (Monaghan et al., 2019; Mirkovic et al., 2021 and Chapter 2 of this thesis). More broadly, the new measure can be used to investigate the theoretical underpinnings of statistical learning, in terms of the role of implicit and explicit knowledge within a multi-componential view (Conway, 2020; R. Frost et al., 2015; R. Frost et al., 2019).

To develop and adapt a subjective measure for use with the word-picture matching task used in this thesis, implications from the existing subjective measures literature were considered within the context of the criteria for subjective measures (Moroshkina et al., 2019; Newell & Shanks, 2014). Likert, confidence rating scales, adapted from the implicit learning literature, has provided a strong basis upon which to build a confidence response protocol. A Likert confidence rating scale was thus developed around five considerations. The first consideration detailed where best to place the scale to ensure the criteria of relevance, reliability, sensitivity, and immediacy are met. The second consideration considered further measures to best meet the subjective measures criteria (Newell & Shanks, 2014) and reduce reactivity. The third consideration explored the different types of Likert scales to assess which scale worked best within the grammatical generalisation paradigm used in this thesis. The fourth consideration looked at how the Likert confidence rating scale could also be suitable for child participants with the fifth consideration assessing how the analysis used may impact the scale.

This review resulted in the development of an integrated Likert confidence rating scale to investigate the contribution of explicit knowledge to grammatical generalisation within a statistical learning framework. As well as potentially being more sensitive, confidence ratings help to avoid the use of language which can prompt explicit knowledge emergence. Likert confidence ratings scales can also be easily combined with the objective threshold measure (Channon et al., 2002; Norman & Price, 2015; Rebuschat, 2013; Tunney, 2005; Tunney & Shanks, 2003) which can help to further reduce reflection time and opportunity. Likert based awareness scales have also been found to be suitable for children (Lockl & Schneider, 2002; Visser et al., 2014), with combined Likert scales also being child-friendly when making grammatical judgement (Davies et al., 2016; Katsos & Bishop, 2011; Katsos & Smith, 2010; Pipijin & Schaeken, 2012). These child-friendly Likert scales also use visual scale labels to aid understanding, supporting the continued use of a happy and sad face response protocol that has already been used within this thesis (Chapter 2).

A combined scale labelled using the concept of certainty (Channon et al., 2002; Lockl & Schneider, 2002; Visser et al., 2014) can also aid the use of language which does not refer to the underlying regularity of interest. This again helps to reduce reactivity. Finally, by using Rebuschat's (2013) recommendations for confidence rating analysis, ensures the use of a scale which includes a 'guess' option as well as 'confidence' options. This enables the use of both a 'zero-correlation' analysis to consider the presence of explicit knowledge and a 'guessing criterion' analysis to also directly consider the presence of implicit knowledge and (Dienes & Perner, 1996). Following Rebuschat's (2013) also means the most relevant analysis for the paradigm used in this thesis can be used without the need for major and unfeasible methodological changes.

The resulting scale is similar to that used by Channon et al. (2002; see Table 24 for details). However, minor adaptations will be added to incorporate visual happy and sad face labels along with adapted word labels to fit with the statistical learning paradigm used in this thesis. The wording will still utilise the concept of certainty, but instead of using *'endorse'* or *'reject'* participants will be asked to judge if the word-picture pair '*goes well with'* or *'does not go well with'*. To aid child suitability, instead of the option *'fairly certain'*, the word *'think'* will be used to denote a lower level of certainty/confidence. Taken together these adaptations result in a Likert confidence rating and word-picture matching judgement scale shown in Figure 11.

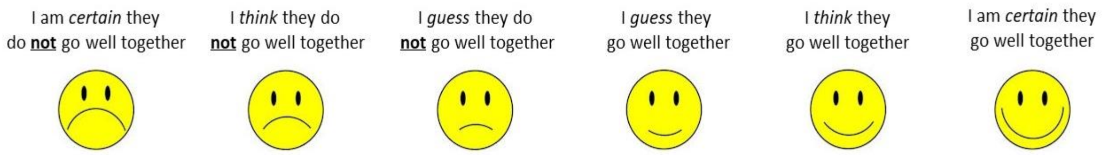

Figure 11. Finalised, combined confidence rating measure to be incorporated into the training and generalisation word-picture matching task.

Chapter 4. Grammatical Generalisation and Statistical Learning: The role of explicit and implicit grammatical knowledge.

4.1 Introduction

Statistical learning is commonly defined as the ability to implicitly extract statistical patterns from the environment (Aslin & Newport, 2012; Batterink et al., 2015; R. Frost et al., 2015; Gómez, 2002; Hall et al., 2018; Saffran, Aslin, et al., 1996). However, recent work has begun to consider the possibility that explicit awareness may also contribute to performance on statistical learning tasks, particularly in adults (e.g., Batterink et al., 2015; Franco et al., 2011; Mirkovic et al., 2021; Monaghan et al., 2019). This work has followed an increasing recognition of the parallels between the statistical and implicit learning literatures, and specifically the proposal of an interacting dual-systems model in which both implicit and explicit memory systems may be involved in learning (Batterink et al., 2015; Batterink et al., 2019; Conway & Christiansen, 2005; Franco et al., 2011; Monaghan et al., 2019; Perruchet & Pacton, 2006; P. J. Reber, 2013; P. J. Reber et al., 2019; P. J. Reber & Squire, 1994).

Despite these theoretical shifts in the literature, there is still limited direct evidence for the presence and use of explicit knowledge within language-based statistical learning tasks. The current evidence base comprises two studies focusing on the detection of word-boundaries using the classic word-boundary paradigm (Franco et al., 2011; Batterink et al., 2015), and three studies extending this approach to grammatical learning paradigms (Monaaghan et al., 2019; Mirkovic et al, 2021; Chapter 2 of this thesis). A recent review of the statistical learning literature by R. Frost et al. (2019) underlined the need for more empirical research to consider the role of both implicit and explicit learning systems, as well as extending the toolkit of statistical learning studies to include complex paradigms which more closely mirror the demands of natural language learning. The current study aims to help address these gaps in the literature by exploring the role of explicit and implicit knowledge within statistical learning by using a more complex grammatical generalisation paradigm.

4.1.1 Measuring explicit and implicit knowledge of grammatical regularities using subjective measures.

To properly explore the role of both explicit and implicit knowledge within more complex statistical learning paradigms, robust and sensitive measures are needed. Whilst previous studies have started to consider the role of explicit grammatical knowledge within grammatical learning paradigms (Mirkovic et al., 2021; Monaghan et al., 2019; Chapter 2 of this thesis), as discussed in Chapter 3 the methods used had various limitations. All cited studies used verbal report measures which lack sensitivity, relevancy, reliability, and immediacy; all key criteria for a good measure of explicit and implicit knowledge (Merikle & Reingold, 1992; Moroshkina et al., 2019; Newell & Shanks, 2014). Monaghan et al. (2019) also used a more sensitive measure of explicit knowledge which does meet these criteria, in the form of the 'decision strategy attribution' test. However, as detailed in Chapter 3, this measure also has a serious limitation: because the response options refer to the underlying grammatical structures of interest in the trained items, the measure could itself be eliciting explicit awareness, thus failing the requirement for non-reactivity (Moroshkina et al., 2019).

To overcome these limitations, the current study uses a confidence rating measure that does not refer to the underlying grammatical regularities either in the response options or in the wording used to introduce the measure to participants. This was possible due to the focus on generalisation in the paradigm used in this study, meaning the items used for the training are different from the (previously unseen) items used in the tests. The response options combine grammaticality judgement and confidence judgements in order to reduce the number of times participants are asked to reflect on their knowledge. Two forms of this confidence rating will be used across two experiments with Experiment 1 using a 5-point scale and Experiment 2 using a 6-point scale.

4.1.1.1 Zero-correlation analysis as an indicator of explicit and implicit knowledge.

As a subjective measure of consciousness, confidence ratings aim to detect when knowledge that is driving a behaviour goes from being implicit to explicit (Cheesman & Merikle, 1984; Dienes et al., 1995; Dienes, 2007; Dienes & Berry, 1997; Moroshkina et al., 2019). This transition point from implicit to explicit is known as the 'subjective threshold' and confidence ratings aim to detect whether this threshold has been reached. Zero-correlation analysis can indicate whether the subjective threshold has been reached or not. It is based on the idea that if a person has crossed the subjective threshold and explicitly knows about the knowledge that is informing their behaviour in a task, then they will be more certain when they 'know' and when they are 'guessing'. This results in a relationship between accuracy on the behavioural task and levels of reported confidence, with higher accuracy being associated with higher reported confidence. If this positive relationship is found it suggests that the subjective threshold has been reached and that explicit knowledge is present and being used (Chan, 1992, but also see Dienes, 2007; Dienes et al., 1995; Dienes & Berry, 1997; Moroshkina et al., 2019; Norman & Price, 2015).

When there is no relationship - that is, a 'zero correlation' - between accuracy and confidence, it is taken to mean the absence of explicit knowledge, and therefore that implicit knowledge must be driving behaviour. However, this is only indirect evidence of implicit knowledge. Including a 'guessing' option, can provide an alternative, more direct measure of implicit knowledge using confidence rating measures.

4.1.1.2 The guessing criterion and implicit knowledge.

The 'guessing criterion' focuses on the situation where the participant has not reached the subjective threshold and must therefore be using implicit knowledge to drive behaviour (Dienes et al. 1995). This analysis only includes trials in which the participant reports that they are guessing. If they perform above chance on accuracy on these trials, then this is a direct indicator of the use of implicit knowledge. Taken together the zero-correlation and guessing criterion analyses provide complementary approaches to assess the presence of both explicit and implicit knowledge, in the context of a task using subjective, confidence-rating measures. However, there are alternative, and potentially more sensitive types of measures which can be used to index implicit knowledge.

4.1.2 Measuring implicit knowledge: Moving away from subjective measures

Other measures of implicit knowledge have derived from the distinction between 'direct' and 'indirect' task types. Direct tasks require participants to come out of the immediate situation to think about and assess their knowledge before providing a response. These include commonly used tasks such as grammaticality judgements, or alternative-forced-choice tasks, in which the participant has to make a choice or a decision about the stimuli presented. Reflecting on one's knowledge to respond in this way means that direct tasks are likely to elicit explicit strategies in participants (Kelley & Lindsay, 1996; Moroshkina et al., 2019; Rebuschat, 2013; Timmermans & Cleeremans, 2015) although it is possible and even probable that implicit knowledge also contributes to performance (e.g., Cheesman & Merikle, 1984; Dienes, 2007; Dienes et al., 1995). In contrast, indirect tasks, sometimes referred to as 'online' tasks, indicate how knowledge is being used *in situ* rather than requiring participants to report on their knowledge directly. Indirect tasks typically use reaction-time, eye-tracking, or ERP measures to index cognitive processes as they occur in real time. The short (millisecond) time-frame that paradigms like this focus on means that they are highly likely to capture implicit processing, which is much more rapid than explicit, meta-cognitive processing (Conway, 2020; Destrebecqz & Cleeremans, 2001; 2003; Dienes, 2007). As such, indirect measures of grammatical knowledge may provide a more sensitive behavioural measure of implicit knowledge than confidence ratings.

4.1.2.1 Statistical learning and indirect tasks: A better measure of implicit knowledge

Drawing from the implicit and second language learning literatures, Batterink et al., (2015) used a range of measures that assessed the use of implicit as well as explicit knowledge, in a word boundary statistical learning paradigm. The authors based their stimuli on Saffran et al.'s (1996) transitional probability task and assessed learning across two experiments using an 'alternative forced choice recognition' task and a 'speeded target detection' task. The forced choice task is traditionally used within the statistical learning literature to test whether participants have learnt the transitional probabilities between artificial words and can be considered a direct task. In order to assess explicit and implicit contributions to performance, Batterink et al. (2015) incorporated confidence ratings into the alternative forced choice task.

The ‘speeded target detection’ task tested learning by using reaction times for correctly detected syllables that were either ‘correctly’ or ‘incorrectly’ placed at the end of a tri-syllabic word within a continuous stream. If participants had learnt the transitional probabilities during the familiarisation phase of the experiment, then reaction times would be faster for ‘correctly’ placed syllables compared to ‘incorrectly’ placed syllables. The real-time nature of this task is proposed to make it a more sensitive index of implicit knowledge (Conway, 2020; Destrebecqz & Cleeremans, 2001; 2003; Dienes, 2007). A further indirect measure of learning was conducted by taking ERP measures during both the ‘alternative forced choice’ and ‘speeded target detection’ tasks. Specifically, the authors used the ‘P300’ effect which can indicate how predictable target stimuli are. Thus, if a participant demonstrates a P300 effect, it indicates that they have learnt the underlying transitional probabilities to a level that enables them to predict when a syllable will occur. Due to the speed at which this predictability is detected (approximately 300 ms after the stimulus is presented), the knowledge this predictability measure is based on is assumed to be automatic and implicit.

By using these different measures, Batterink et al., (2015) found that both implicit and explicit knowledge is involved in adult word-boundary learning. More specifically, and in relation to the sensitivity of indirect measures for implicit knowledge, the authors found that the ‘speeded target detection’ task, was a more robust measure of implicit knowledge use in statistical learning: performance in this task correlated with the P300 effect, whereas performance in the ‘alternative forced choice’ task did not. This supports the use of the ‘speeded target detection’ as an indirect learning measure that is more likely to use implicit knowledge compared to the traditional, direct ‘alternative forced choice’ testing task. On the other hand, the use of explicit knowledge in the alternative forced choice task was indicated by responses on the confidence ratings: accuracy on this task was correlated with higher confidence ratings in experiment 1 (zero-correlation), as well as with ‘remember’ judgements (remember/know paradigm, see Chapter 3) in Experiment 2. Interestingly, analysis of guess responses in Experiment 2 did not show above-chance accuracy, indicating the absence of implicit knowledge, and suggesting that performance on the 2AFC task may be driven entirely by explicit knowledge.

The findings reported by Batterink et al. (2015) provide strong evidence that both implicit and explicit knowledge is used by adults in the traditional word-boundary detection task. They also suggest that indirect measures of learning are a more sensitive way to measure implicit knowledge compared to subjective measures applied to direct, offline tests of learning.

4.1.2.2 Adapting the indirect measure for a word-boundary paradigm to a grammatical learning paradigm.

The indirect task used by Batterink et al. (2015) tested learning of adjacent triplets of syllables making up the novel words in the word-boundary detection paradigm. By contrast, our grammatical learning paradigm uses non-adjacent dependencies between the determiner and suffix to cue grammatical categories. An indirect task to test implicit learning of this grammatical regularity requires a design that can capture learning of non-adjacent dependencies. Misyak et al. (2010b) developed an indirect measure of learning for a grammatical paradigm that considered non-adjacent dependencies (similar to distributional regularities). Their task drew upon the implicit learning literature and combined the properties of an artificial grammar learning task (e.g., A. S. Reber, 1967) and serial reaction time task (Nissen & Bullemer, 1987) within a statistical learning framework. This provided both a direct (artificial grammar learning) and indirect (serial reaction time) measure of learning. Misyak et al. (2010b) used the artificial language created by Gómez (2002), which was designed to examine and test the learning of non-adjacent dependencies. It consists of tri-syllabic words that form three grammatical categories through dependencies between the first and final syllable of the words (*a_d, b_e and c_f*) which are on either side of an arbitrary stem syllable (*X).* This forms strings with the structure *aXd* (e.g., *pel wadim rud*), *bXe (*e.g., *dak kicey jic)* and *cXf (*e.g., *vot pucer tood).*

The serial reaction time task was created by presenting a two by three grid on a screen, and stimuli were presented orthographically in the grid. A target and foil first word (*a, b* or *c*) were displayed in the top and bottom leftmost boxes, a target and foil interleaving stem (*X*) were displayed in the middle boxes and a target and foil final word (*d, e* or *f*) displayed in the two rightmost boxes. Once participants had viewed the written form of the items on this grid, an audio recording of the target word sequence (e.g., *aXd*) was played to the participants. Whilst the auditory version of the item was being played, the participant was asked to use the computer mouse to click the boxes on the grid that contained the target words being played to them. The speed with which participants clicked on the final element in the string indicated whether they were able to predict this element on the basis of the preceding elements; and therefore, whether they had learned the non-adjacent dependency. Participants were exposed to the language using three blocks, presented continuously; a training block where all strings were 'grammatical', then an ungrammatical block where strings violated the regularity on the final word and then a recovery block using grammatical strings. The difference in reaction time between the final and initial string elements, tracked across all three blocks, was used as the indirect measure of learning.

Following the serial reaction time task, participants completed a prediction task to provide a direct measure of learning of the language they had been exposed to during the preceding task. This task formed the artificial grammar learning part of the overall experimental design through the use of a two-alternative forced choice style task. Participants were told that there were 'rules' that specified the order of the words and were then presented with the grid and orthographic form of the first two words in the string. They had to then choose which was the ‘correct’ final word in the string from the two options presented in the right most boxes on the grid.

The findings indicated learning of the grammatical regularities in both the serial reaction time task, and the artificial grammar/two-alternative forced choice test. Of particular interest when considering indirect tasks is that reaction time speed for clicking the correct final word target significantly reduced across training (Misyak et al., 2010a, 2010b). This demonstrates the successful development and application of an indirect task within a statistical learning framework that considers learning at the grammatical level, and specifically is able to assess implicit learning of non-adjacent dependencies. As such, it provides a useful basis for developing a similar task for use in our grammatical learning paradigm. However, two key design elements need to be considered in any adaptation. Firstly, the paradigm used by Misyak et al. (2010b) relied on both auditory and orthographic presentation of the stimuli, whereas in our paradigm, stimuli are presented only auditorily. Secondly, the grammatical regularities in the language used by Misyak et al. (2010b) were restricted to a distributional cue (e.g., a_d dependency), while in our language, the grammatical regularities also have a semantic component.

4.1.2.3 Adapting the indirect for a cross-domain grammatical learning paradigm.

This current study uses an artificial language based on grammatical gender systems found in natural languages. Grammatical genders are associated with regularities in the semantic, phonological, and distributional domain (Corbett, 1991, 2013; Mirković et al., 2005; Zubin & Köpcke, 1981). To create semantic regularities, two semantic categories were used in our artificial language: animals and artefacts. Phonological regularities were incorporated using a “suffix” that was attached to a “stem” (e.g., *-eem:* mof<u>*eem*</u>). Finally, distributional regularities were incorporated as a co-occurring “determiner” and “suffix” (e.g., <u>*tib*</u> mof<u>*eem*</u>; see Table 1 for examples). Each co-occurring determiner and suffix was paired with a semantic category (animals or artefacts) using a picture referent. This provided an *a*X*b* structure for animals and *c*Y*d* structure for artefacts, with X and Y denoting the interleaving arbitrary stem, *a* and *c* the determiners and the *b* and *d* the suffixes (see Table 1). This creates a cross-domain artificial language with words that are only presented auditorily.

Based on the design of Batterink et al.’s (2015) ‘speeded target detection’ task, a ‘phoneme detection’ task was devised as the indirect measure of learning for our paradigm. In the two forms of indirect task described so far, it is the reaction time for detecting the final syllable or word in a string that is of interest, as preceding string elements provide the necessary information for predicting this final element. As the artificial language being used in this study also comprises of three elements, similar to Gómez (2002) and Misyak et al.'s (2010b) artificial languages, this focus on the final element will be incorporated into the ‘phoneme detection’ task being developed here. As the word stimuli will not be orthographically presented, creating a forced-choice, serial reaction time test as in Misyak et al. (2010b), would not be appropriate.

Instead, Batterink et al.’s (2015) target detection procedure will be utilised instead, with participants being asked to detect a phoneme found within one of suffixes used within our artificial language (*eem* and *ool*).

This design allows for a verbal-only presentation of the words (as in Batterink et al., 2015), along with the presentation of the visual/semantic element of the language that forms the semantic regularity. The co-occurrence between the determiner (e.g., *tib* and *ked*) and the semantic regularity from the picture provide available information for potential prediction which could support the detection of a phoneme found within the suffix. Faster reaction times would therefore be expected when the final suffix can be predicted from the available information compared to when it cannot be predicted. This follows a similar principle to other tasks which manipulate predictability and examine the effect on response times, including the serial reaction time task (Nissen & Bullemer, 1987), and phoneme identification tasks in which target phonemes are identified faster when presented in words which can be predicted from the context compared to when the context does not aid prediction (Dell & Newman, 1980; Morton & Long, 1976).

The effects of predictability on reaction times could also be viewed in terms of congruency. In our grammatical learning paradigm, when the suffix can be predicted from the determiner and picture within grammatical items, the item would be congruent with the trained regularities. However, items would be incongruent when the suffix cannot be predicted from the determiner and picture in ungrammatical items. Psycholinguistic literature that has explored the effects of lexical information (syllables) on phoneme detection, has also found faster reaction times for congruent versus incongruent conditions (e.g., Chetail & Mathey, 2013; Mehler et al., 1981). This provides further support for the use of reaction times to detect grammatical learning within the phoneme detection task constructed for this study.

By utilising different aspects of Batterink et al.’s (2015) speeded detection and Misyak et al.'s (2010b) serial reaction forms of indirect task, this new indirect measure of grammatical learning has been constructed for use within a cross-domain artificial language. This new task aims to provide a more sensitive measure of implicit knowledge, to support investigations into the role of implicit knowledge within a grammatical learning paradigm of statistical learning.

4.1.3 Current Study

The overall aim of this study is to investigate the role of both explicit and implicit knowledge on grammatical generalisation within a statistical learning framework. Across two experiments, the current study uses an adaptation of the paradigm used in Chapter 2 of this thesis, incorporating confidence ratings to provide a more fine-grained, sensitive, within-task measure of explicit and implicit knowledge than is possible with the verbal reports used previously. The specific goals of Experiment 1 are firstly, to replicate the findings of a partial role for explicit knowledge in grammatical generalisation, reported in Mirkovic et al. (2021) and Chapter 2 of this thesis, using confidence ratings instead of verbal reports; and secondly, to assess whether the inclusion of a subjective measure influences the learning and generalisation performance of participants as compared to previous studies using this paradigm - that is, to assess whether the measure is substantially reactive or not. In order to address these goals, Experiment 1 utilised a 5-point confidence rating scale for each participant response in both training and generalisation trials; explicit knowledge was assessed using the zero-correlation criterion.

Experiment 2 also considered the role of explicit knowledge using a confidence rating measure, but the focus was expanded to additionally consider the role of implicit knowledge. This included the addition of two measures of implicit knowledge: the guessing criterion and an indirect measure. To be able to conduct a guessing criterion analysis, a guessing response was added to the confidence rating, changing it from a five- to a six-point scale. A new, indirect measure was included in the form of a 'phoneme detection' task, using reaction times to capture the use of learnt grammatical regularities to predict upcoming phonemes.

4.2 Experiment 1

The experimental predictions for Experiment 1 were in the first instance concerned with replicating our previous findings. Specifically, we predicted that 1) participants would learn both the novel words and the grammatical regularities to a similar level to that found in our previous studies, 2) the level of explicit knowledge assessed by retrospective verbal report would be similar to that in our previous studies, and 3) that retrospectively reported explicit knowledge would be associated with the level of learning. In addition, we hypothesised that our new confidence rating measure would capture explicit awareness of the grammatical regularities, and so predicted 4) that participants would have greater levels of confidence in their correct responses, as compared to their incorrect responses in the generalisation tests (the 'zero correlation' criterion).

The two key design principles of the confidence rating scale used in Experiment 1 were: firstly, the grammaticality response is integrated with the confidence rating (similar to Channon et al., 2002), in order to minimise reactivity by reducing the number of opportunities for reflection per item, and by using language that does not refer to the underlying grammatical regularity of interest. Secondly, a certainty-based confidence scale was used (again similar to Channon et al., 2002), but with the addition of smiley and sad faces for each response item to aid understanding of the scale. This made for a comparable response protocol to previous studies using this paradigm (e.g., Chapter 2; Mirkovic et al., 2021) where a binary response protocol using a happy and sad face was used. It was also included to help make the scale child-friendly, as this experiment had originally been intended for use in a study with children, before the pandemic necessitated a change of plan.

4.2.1 Method

This experiment used similar stimuli and methodology from Chapter 2, Experiment 2 of this thesis. As such the below will only describe any differences or new elements for this current experiment and state specific sections from Chapter 2 where details can be found of what was kept the same.

4.2.1.1 Participants

Thirty-two adult participants with a mean age of 22.21 years (18.5-31.3*;* 8 male) took part in this study. The adult sample was drawn from the undergraduate and postgraduate population at the University of York, some were volunteer participants and others received course credits for their participation. The study protocol was approved by the ethics committee at the Department of Psychology, University of York.

4.2.1.2 Stimuli

All non-proprietary materials are provided on OSF. See also Appendix A for full stimuli lists (A1, A3, A4, A6 & A7).

The same stimuli sets were used for this Experiment as used in Chapter 2 (see section 2.2.1.2 & 2.3.1.2 for details). There was only a difference in the mismatched items used in the 'repetition and word-picture matching' training task detailed below.

Repetition and Word Picture Matching

This task included all 32 trained word-picture pairs with 8 additional trained item trials (drawn from the training item set). These additional trials represented the mismatched trials usually used within this task (Breitenstein et al., 2007) and incorporated into this task in Chapter 2 (see section 2.2.1.3 & 2.3.1.3). Mismatched trials were not used in this experiment as it is the first to use a confidence and judgement response protocol within this paradigm and this new protocol may have an effect on learning. So, to mitigate this potential effect, allow for replication of previous results along with assessing any impact the new response protocol may have, 'noise' was reduced by replacing mismatched items with matched. A different set of 8 trained items were used here for the four 'repetition and word-picture matching' tasks run during training (Figure 12).

4.2.1.3 Procedure

Participants completed all experimental tasks individually within the lab or a quiet home setting. The experiment was conducted in one session of approximately 50-70 minutes and consisted of a training protocol and a testing protocol. The training protocol consisted of 4 blocks of a combined repetition and word-picture matching task followed by a word testing task (see Figure 12 for the experimental protocol). Responses were recorded by the 'DMDX' programme

(Forster & Forster, 2003) on a PC laptop computer, unless otherwise stated. Participants were introduced to experimental tasks as a series of games involving 'alien' words introduced by a visiting extra-terrestrial. As such it was introduced to participants as a word-learning study and were not informed of the grammatical element of the study until the experiment was completed. After the computer-based experimental tasks, additional psychometric tests were conducted: Expressive vocabulary and Matrix Reasoning Ability were measured using the corresponding subtest in 'Wechsler Abbreviated Scale of Intelligence', Second Edition and a test of phonological short-term memory was conducted through a non-word repetition task using 'The Children's Test of Nonword Repetition'. Although results from these were not included in analysis.

The same or similar training and testing task procedures were used from Chapter 2, Experiment 2 (see section 2.3.1.3). The below only details any differences in this current experiment from the procedures used in this previous experiment.

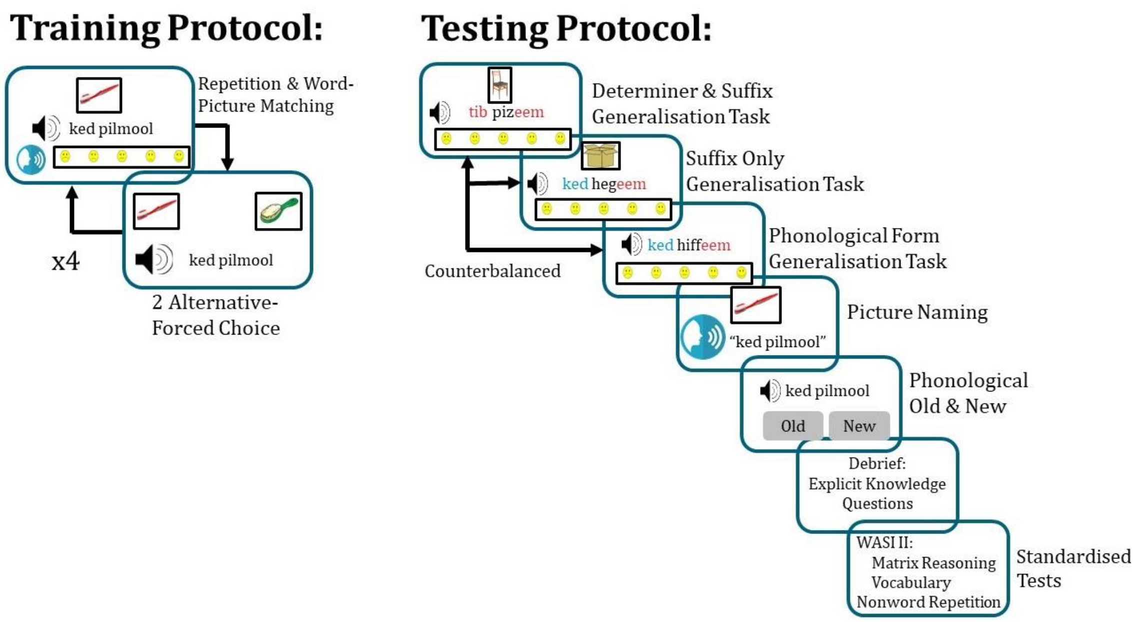

Figure 12. Experiment 1 procedure. The generalisation tasks depict an inconsistent item.

Training Tasks:

Repetition and Word-Picture Matching

The same combined repetition and word-picture naming tasks from Chapter 2, Experiment 2 was used (see section 2.3.1.3) with some changes made in order to address the specific aims of this current Experiment. Firstly, a criterion learning procedure was not used, instead this task was repeated 4 times during the training phase of the experiment (see Figure 12) providing the same exposure to the language as Chapter 2, Experiment 1 (see section 2.2.1.3). Secondly, while mostly the same trial procedure and timings were used, the audio of the 'alien' word was now played for a second time during a trial, after participants had repeated the word and pressed the spacebar to start the word-picture matching part of the trial. Also, this second section of the trial now timed-out after 6000 ms. Thirdly, a different response procedure was used for word-picture matching, in order to incorporate the new confidence rating scale.

To indicate their word-picture judgement response, participants were asked to choose one of five smiley/sad face options. These options were denoted by smiley/sad face pictures shown on the screen which corresponded to stickers placed on the computer keys for numbers 1-5. For the 'sad faces': key 1 had a 'very sad face' denoting the option '*I am certain the word and picture don't go well together'* and key 2 had a 'slightly sad face' denoting *'I think the word and picture don't go well together'*. For the 'happy faces': key 4 had a 'slightly happy face' denoting the option *'I think the word and picture go well together'* and key 5 had a 'very happy face' denoting *'I am certain the word and picture go well together'*. Key 3 had a 'straight face' denoting the option *'I am not sure whether the word and picture go well together'* (see Figure 13). These options allowed participants to give both a word-picture matching judgement as well as a confidence rating for this response. This scale was introduced on screen in the pre-task instructions and a copy of the smiley/sad faces, and their wording was placed next to participants for their reference. This task procedure took approximately 4 minutes to complete.

Across the four training tasks, each trained word-picture pair was presented for a total of 9 times (4 times in the repetition section of the trial, 4 times in the word-picture matching section and 1 across the mismatched replacement trials)

Figure 13. Five-point judgement scale for the training and generalisation word-picture matching tasks for Experiment 1.

Testing Tasks:

Word Learning: Picture Naming

The same task procedure was used from Chapter 2 (see section 2.3.1.3). However, responses were marked slightly differently. In this current experiment, to be correct, the utterance had to exactly match the phonology of or be phonetically close to the words produced in the training tasks. For example, productions such as /tid/ and /tee/ for '*tib',* along with /ket/ and /ken/ for '*ked'* were accepted as correct determiners, if the initial sound was correct. For the suffix, productions such as /een/ and /ee/ for '*eem'*, along with /all/ and /ole/ for '*ool*' were accepted as correct. Like the determiners, stems had to start with the correct initial sound with productions such as /gash/ for *'gatch'* (*tib* gatch**eem** – snake) and /snor/ for *'snar'* (*ked* snar*ool* – TV) being scored as correct.

Word Learning: Phonological Old and New

Similar trial procedure and timings were used from Chapter 2 (see section 2.3.1.3). However, in this current experiment, training items and foils were presented once, randomly by the DMDX programme. Participants also now indicated their response by using the 'left shift' (old) and 'right shift' (new) keys, with 'old' and 'new' stickers were placed on the relevant shift keys to aid participants.

Generalisation: Word-picture matching tasks

Similar trial procedure and timings were used from Chapter 2 (see section 2.3.1.3). However, there were differences incorporated into these testing tasks for this current Experiment, following the procedural details discussed for the word-picture matching training tasks.

4.2.1.4 Data Analysis

All analyses were conducted in R (R Core Team, 2020). All data, scripts and outputs are provided on OSF.

To assess performance change across training for word-learning, as assessed by the 2AFC tasks, linear, mixed-effect regressions were used (using the lmer function in R). Polynomial contrasts were set to the four 2AFC tasks using default contrast matrices within the contrast function in R (contrast.poly function; see Schad et al., 2020). This enabled both linear and quadratic comparisons of the tasks.

For the word-picture matching generalisation tasks the sad face responses (keys 1 & 2) were considered 'rejections' and happy face responses (keys 4 & 5) were considered 'endorsements'. Using this, endorsement rates for consistent vs. inconsistent items were used to derive an A' metric (Pallier, 2002) for analysis. Neutral face (3) responses were not included in analysis

To analyse the confidence ratings judgements for the word-picture matching generalisations task, judgements were converted to a score of 1 or 2 based on Channon et al.'s (2002) procedure. The '*I am certain'* judgements for both the sad (*'does not go well with'*; reject) and happy ('*does go well with'*; endorse) scale points (keys 1 & 5) are considered 'high confidence' and were scored '2'. The '*I think'* judgements for both the sad and happy scale points (keys 2 & 4) are considered 'low confidence' and were scored '1'. The neutral face (3) responses were again not included in analysis here.

These confidence ratings were then used in a 'zero-correlation' analysis to assess the use of explicit and implicit knowledge within grammatical generalisation. The rationale for this analysis is that if increased accuracy is related to increased confidence, this is an indication of explicit knowledge (Dienes, 2007; Dienes & Perner, 1999; Norman & Price, 2015; Rebuschat, 2013). If confidence and accuracy are not related (zero-correlation) this shows an absence of explicit knowledge. Based on this, confidence judgements scores will be compared for correctly and incorrectly judged items. It should be highlighted that this is different from the A' scores that are being used to judge whether generalisation has occurred in these tasks. The A' scores are based on the endorsement rates for both consistent and inconsistent items. To assess whether explicit knowledge is present both endorsement and rejection responses will be included in analysis. Endorsing consistent items and rejecting inconsistent items will be considered correct and endorsing inconsistent and rejecting consistent items will be considered incorrect. Mean confidence ratings will be calculated for correctly and incorrectly judged items.

If correctly judged items have a higher mean confidence rating score than the incorrectly judged items, this suggests a relationship between accuracy and confidence and indicates the use of explicit knowledge (Channon et al., 2002; Cheesman & Merikle, 1984; Dienes et al., 1995; Rebuschat, 2013). If no difference is found between correctly and incorrectly judged items, this suggests no relationship between accuracy and confidence, indicating an absence of explicit knowledge (and potential use of implicit knowledge).

4.2.2 Results

4.2.2.1 Word Learning

2AFC

Word learning at the end of training was examined through accuracy on the final 2AFC task. One-sample t-tests against chance (.5) showed that the participants performed significantly above chance (M=0.88, SD=0.09; $t(31)$=24.33, p<.0001, d=4.30), demonstrating a good level of novel word learning (see Figure 14).

The time course of word learning across training was also assessed using a mixed effects regression analysis which compared accuracy (outcome variable) across the four 2AFC tasks included in the training protocol (fixed factors set as polynomial contrasts). This analysis showed a significant linear increase in performance across training as well as a significant quadratic pattern (see Table 27). This shows that performance increased across training but started to plateau by the end of training as can be seen in Figure 14.

Phonological Form Old and New

The learning of the phonological form of the novel words was examined at the end of the testing session and yielded similar results to the final 2AFC task. One-sample t-tests showed that participants (M=0.86, SD=0.08) performed significantly above chance (0.5; t(31)=26.63, p<.0001, d=4.71). These results demonstrate that participants have a good level of phonological word-form knowledge.

Picture Naming

Vocabulary knowledge was also assessed using picture naming, a language production-based task. This task allowed us to separately assess the learning of the stem, and the learning of the determiners and the suffixes. Stem recall was low relative to the recall of the grammatical morphemes (

Figure 15), likely due to the difference in frequency of exposure. Despite this, a one-sample t-test demonstrated that stem recall was significantly above 0 ($t(31)=13.42$, $p<.0001$, $d=2.37$) suggesting at least some learning of stems. A repeated-measures t-test showed that recall was higher for determiners than suffixes ($t(31)=2.43$, $p=.021$, $d=0.43$). As the frequency of exposure was the same for both determiners and suffixes, this suggests the participants found the determiners to be easier to learn than the suffixes.

These word learning results are similar to those found in previous studies with adults using this experimental paradigm (Mirkovic et al., 2021; Chapter 2 of this thesis), suggesting the inclusion of the confidence rating has not significantly impacted word learning.

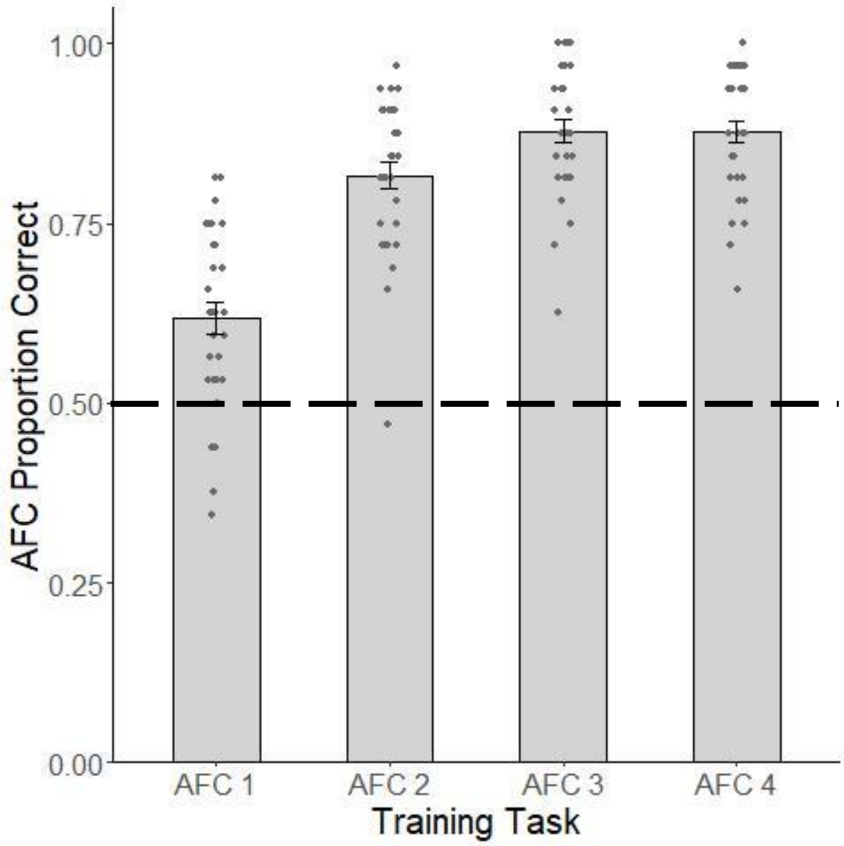

Figure 14. Word Learning: Accuracy on the 2AFC tasks across training in Experiment 1.

Table 27. Regression comparison for 2AFC performance across training in Experiment 1 using polynomial contrasts.

	Estimate	Std. Error	t	*p*	f^2
Linear Contrast	**0.19**	**0.01**	**15.43**	**<.0001**	**0.93**
Quadratic Contrast	**-0.10**	**0.01**	**-8.20**	**<.0001**	**0.26**

Significant results are highlighted in bold.

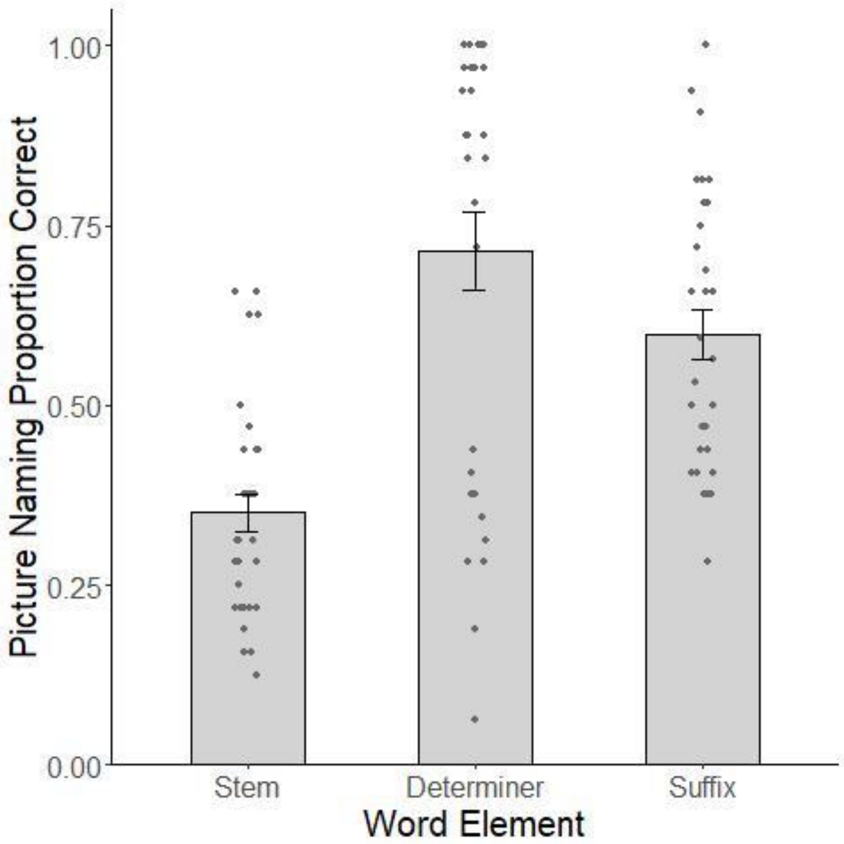

Figure 15. Word Learning: Recall accuracy for the Picture Naming task in Experiment 1.

4.2.2.2 Grammatical regularity generalisation performance

To analyse performance in the generalisation tasks, we derived an A' metric based on the endorsement rates for consistent and inconsistent trials as described in the Methods section. Figure 16 shows generalisation performance across all three generalisation tasks.

To test generalisation performance, one-sample t-tests against an A' of 0.5 were conducted for each generalisation task. Participants demonstrated evidence of successful generalisation in the Determiner and Suffix task (t(31)=7.77, *p*<.0001, *d*=1.37) and Suffix Only task (t(31)=3.36, *p*=.001, *d*=0.59). These tasks assess the generalisation of the distributional regularity and its mapping to the semantic cue, with the Suffix Only task focusing specifically on the suffix. Participants did not demonstrate evidence of successful generalisation in the Phonological Form task (t(31)=1.11, *p*=.137, *d*=0.20). This task specifically assessed learning of the distributional regularity with no reference to the semantic regularities.

These results suggest that adults can learn and generalise distributional, phonological, and semantic regularities. However, this is only the case when the semantic regularity is present. These results are similar to previous studies with adults which used the same experimental paradigm but with binary judgement response for the generalisation tasks rather than a five-point scale (Mirkovic et al., 2021; Chapter 2 of this thesis). As with word learning, these current results suggest that the use of an integrated confidence rating scale does not significantly impact grammatical generalisation performance.

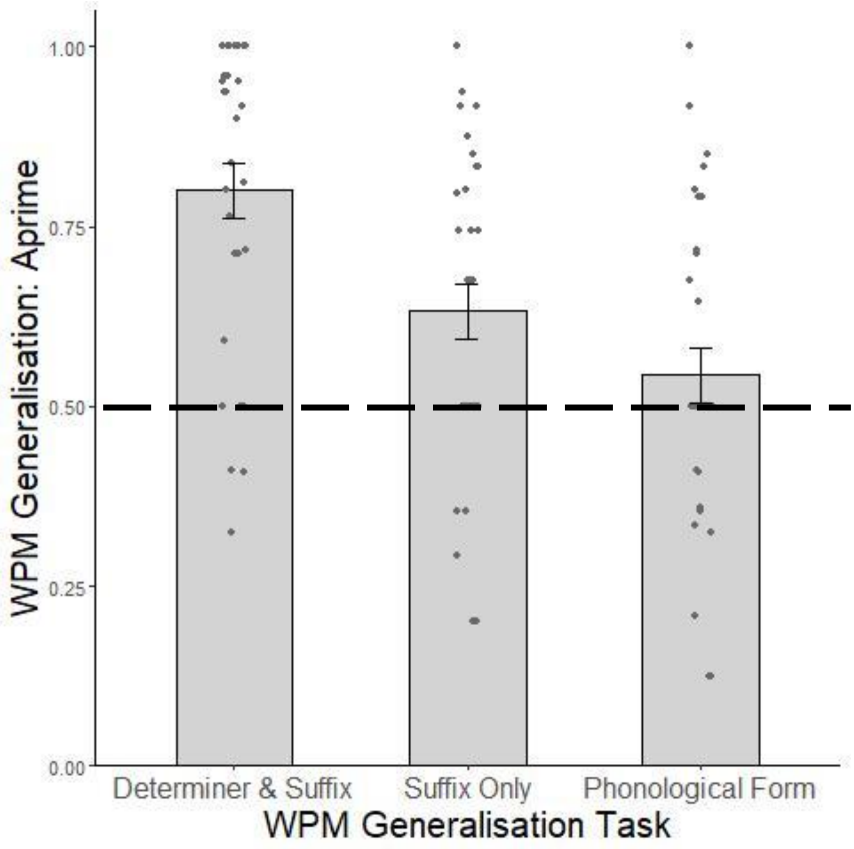

Figure 16. Generalisation: A' performance across all word-picture matching generalisation tasks in Experiment 1.

4.2.2.3 The role of explicit knowledge in grammatical regularity generalisation

The role of explicit knowledge in grammatical generalisation was explored in two ways. First, by replicating previous analysis (Mirkovic et al., 2021; Chapter 2 of this thesis) which considered the contribution of retrospective, verbally reported explicit knowledge on generalisation. Secondly, by using within generalisation task measures of explicit knowledge with confidence ratings.

Retrospective Verbal Reports

Retrospective explicit knowledge was measured using the 'Debriefing Questionnaire' at the end of the experiment. Answers were scored separately for determiner and suffix knowledge (see Table 3); reported explicit knowledge of determiners is numerically higher than for suffixes (Table 28). These levels of reported knowledge are numerically similar to those found in Chapter 2. The most relevant comparison is with adults from that study (Chapter 2, Experiment 1) who received the same level of exposure across training as in the current study (Ch 2: determiner M = 2.39; suffix M = 0.59; Current study: determiner M = 2.16; suffix M = 0.71).

Table 28. Descriptive statistics for verbally reported explicit knowledge in Experiment 1.

	M	SD
Determiner	2.16	1.07
Suffix	0.71	1.13

Examining the contribution of retrospective explicit knowledge on generalisation performance, was also considered at the individual level using regression analysis. This regression approach was applied to each of the three generalisation tasks, including the 'phonological form' task, in which group level performance did not exceed chance. The logic for this is that there was substantial variability within the group, and that it included individuals whose A' scores were above .5, and who thus showed evidence of generalisation (Figure 16). For each regression analysis, the outcome variable was the A' score for the respective task and the predictor variable(s) were the relevant morpheme score(s) for the given task. Both determiner and suffix knowledge were relevant for the Determiner and Suffix and Phonological Form tasks; only suffix knowledge was relevant for the Suffix Only task. Table 29 shows the results for these regression analyses.

Table 29. Multiple regressions for the role of verbally reported explicit knowledge on word-picture matching generalisation performance in Experiment 1.

	R^2	BSE	B	ß	*p*	f^2
Determiner & Suffix Generalisation:	0.22					
Determiner		0.04	0.07	0.34	.077	0.12
Suffix		0.04	0.05	0.26	.177	0.07
Suffix Only Generalisation:	0.11					
Suffix		**0.03**	**0.07**	**0.37**	**.039**	**0.16**
Phonological Form Generalisation:	0.01					
Determiner		0.04	0.01	0.05	.829	0.00
Suffix		0.04	0.05	0.25	.245	0.05

Significant results are highlighted in bold.

A significant contribution of explicit suffix knowledge was found for performance in the Suffix Only task. Neither the determiner nor the suffix significantly contributed to performance in the Determiner and Suffix or the Phonological Form generalisation tasks. Thus, in this experiment, participants only demonstrated a role of retrospective explicit knowledge within the Suffix Only task. This result replicated the finding of a partial contribution of verbally reported explicit knowledge to generalisation performance found in adults in Chapter 2, although with variation in exactly where contributions were found. For example, considered with the most relevant adult comparison (Chapter 2, Experiment 1), adults also show a contribution of explicit suffix knowledge to generalisation performance in the Suffix Only task, as found in this current experiment. However, adults from Chapter 2, Experiment 1 also showed a significant contribution of explicit determiner knowledge to performance in the Determiner & Suffix task.

Within task measures of explicit and implicit grammatical knowledge.

To assess the use of explicit knowledge within the word-picture matching generalisation tasks, confidence rating scores from the five-point response scale were used in a 'zero-correlation' analysis. Mean confidence rating scores were calculated for correct responses (endorsing consistent items and rejecting inconsistent items) and incorrect responses (endorsing inconsistent items and rejecting consistent items) for each generalisation task (see Figure 17).

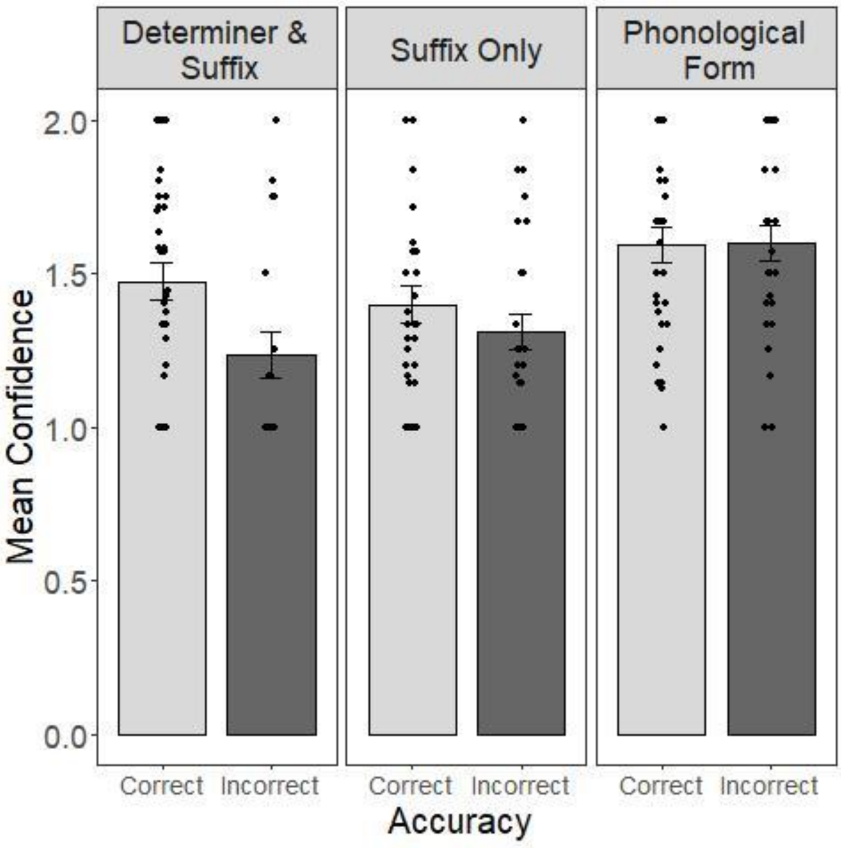

Figure 17. Explicit Knowledge: Mean confidence ratings for correct and incorrect responses across the word-picture matching generalisation tasks in Experiment 1.

A repeated measures t-test revealed a significantly higher mean confidence rating for correct compared to incorrect responses in the Determiner and Suffix task ($t(22)=2.39$, $p=.026$, $d=0.50$). Confidence ratings did not differ significantly between correct and incorrect responses for the Suffix Only task (t(30)=1.31, p=.200, d=0.24) or the Phonological Form task (t(30)=-0.47, p=.645, d=-0.08). These results provide evidence for within-task use of explicit grammatical knowledge for the Determiner and Suffix task, and no evidence for within-task explicit knowledge for the Suffix Only or Phonological Form tasks. According to the logic of a zero-correlation analysis, the successful generalisation performance found within the Suffix Only task and the lack of evidence found for explicit knowledge use, suggests that implicit knowledge may have contributed to performance here. However, the zero-correlation analysis does not directly assess the presence of implicit knowledge and so caution should be taken when assuming the presence of implicit knowledge here.

4.2.3 Discussion

Participants demonstrated good levels of word-form and vocabulary knowledge (Phonological Old/New and 2AFC tasks respectively). These results are similar to those found in both Mirkovic et al. (2021) and Chapter 2 of this thesis, suggesting that the use of a response scale instead of a binary response protocol has not affected word learning within this paradigm. Participants also demonstrated word learning across training with a linear increase in 2AFC performance and a plateau in performance by the end of training.

It was hypothesised based on previous findings, that participants would demonstrate successful generalisation across all word-picture matching generalisation tasks. This hypothesis was partially supported as evidence for successful generalisation was found for the grammatical regularities that include semantics but not when semantics were absent (Phonological Form Task). Notably, A' scores were numerically higher for the Determiner and Suffix task compared to the Suffix Only task, suggesting that the presence of the determiner is beneficial for learning. This broad pattern of results is consistent with previous studies using this paradigm, albeit with some variations: while phonological form learning was not significantly above chance in the current experiment or in Mirkovic et al. (2021), it was above chance in Chapter 2, Experiment 1. However, the A' scores for this task across experiments are very similar, and differences in significance should not be over-interpreted. Taken together, these findings suggest that learning of the distributional regularity (co-occurrence of determiner and suffix) in the absence of semantics is the hardest aspect of the grammatical regularities for participants to learn. In contrast, the presence of semantic cues, and the presence of the determiner, appear to support learning.

In addition to ensuring that the new confidence rating measure did not impact learning of words or grammatical regularities in the current paradigm, Experiment 1 also aimed to assess the potential impact on levels of explicit knowledge and its relationship to generalisation performance. The levels of retrospective, verbally reported explicit knowledge found in this experiment are similar to those found in Chapter 2 (Experiment 1 adults) and Mirkovic et al. (2021). Additionally, this experiment also demonstrates similar findings of a partial contribution of verbally reported explicit knowledge to performance in the word-picture matching generalisation tasks. Again, while the broad pattern of results is consistent across experiments, the specific effect sizes and patterns of significance do vary from experiment to experiment. In the current experiment, there was only evidence for explicit knowledge of the suffix contributing to generalisation performance (in the Suffix Only task), but not of the determiner. The lack of an association between determiner knowledge and performance on the Determine & Suffix task was surprising, as this effect has been found to be reliable in several previous experiments (Chapter 2, Experiment 1; Mirkovic et al., 2021). It is worth noting that numerically, the regression coefficient for the determiner was similar to that for the suffix, and so again, over-interpretation of the lack of significance would be inappropriate unless this finding is replicated.

The variability in specific contributions may also reflect the weaknesses of the verbal report measure, such as its retrospective nature, and the possibility that participants may not accurately or comprehensively verbalise every aspect of their explicit knowledge. Participants may set their own criteria about what knowledge is important or unimportant to the task and may not report explicit grammatical knowledge despite it being present if they judge it to be unimportant (Tunney & Shanks, 2003; Wierzchoń et al., 2012). The within-task confidence ratings measure was designed to overcome some of the weaknesses of retrospective verbal reports. This measure also showed evidence of explicit knowledge in grammatical generalisation; however, in contrast to the findings from the verbal report measure in this experiment (but consistent with the findings from previous experiments), the zero-correlation analysis suggested that explicit knowledge contributed to generalisation performance in the Determiner & Suffix task but there was no evidence for its use in the Suffix Only task.

Taken together, the results of Experiment 1 provide further support for the literature reporting the presence and use of explicit knowledge within statistical learning paradigms (Batterink et al., 2015, 2019; Franco et al., 2011; Monaghan et al., 2019). The relatively small effect sizes, and the discrepancies reported in Experiment 1 between verbal report and confidence-ratings, as well as those between the verbal report measure here compared to previous experiments, could be the result of noisy data and imperfect measures. Alternatively, they may reflect a more complex underlying reality, in which both explicit and implicit knowledge contribute to performance. The experiments reported in this thesis so far can only indirectly and cautiously infer the presence of implicit knowledge. Experiment 2 aims to capture the use of implicit knowledge more directly, by using the 'guessing criterion' in the confidence-rating measure, as well as a novel indirect reaction time measure.

In conclusion, Experiment 1 showed that introducing a confidence-rating measure that was incorporated into each training trial did not disrupt either the level of learning, or the final level of explicit awareness of the grammatical regularities. It also showed that the new measure can capture explicit awareness of regularities and that this is associated with performance on some aspects of grammatical generalisation.

4.3 Experiment 2

The overarching aim of the current study is to explore the potential contribution of explicit knowledge to the learning of grammatical regularities, and complementing this, to capture implicit knowledge of these regularities. Building on the findings of Experiment 1, Experiment 2 aimed to directly measure implicit knowledge of grammatical regularities, using two different measures: firstly, the confidence ratings themselves, and secondly, a new reaction time measure.

Capturing implicit knowledge in the confidence rating measure was done by adding a 'guess' option, which allows for the use of the 'guessing criterion'. This allows access to implicit knowledge as it focuses only on the responses which participants say are a guess: that is, the presence of explicit knowledge is ruled out in these responses, so that accurate performance must necessarily reflect implicit knowledge instead (Cheesman & Merikle, 1984; Dienes et al., 1995). In order to allow for a 'guess' response option, we expanded our confidence rating scale from a five-point scale (as used in Experiment 1) to a six-point scale (see Figure 11). This enabled the inclusion of a 'guess' confidence rating for both an endorsement and a rejection response to the word-picture matching judgement; above chance accuracy for guess responses indicates the use of implicit knowledge (Cheesman & Merikle, 1984; Dienes, 2007; Dienes & Perner, 1999; Norman & Price, 2015; Rebuschat, 2013). This six-point scale is still suitable for a 'zero-correlation' analysis, allowing for replication of this analysis from Experiment 1.

In addition to incorporating the 'guessing criterion' into the confidence rating scale, Experiment 2 also uses a novel reaction time measure, to attempt to directly capture online, real-time use of implicit knowledge of grammatical regularities. Based on the 'speeded target detection' task used by Batterink et al. (2015), we developed a phoneme detection task in which participants are asked to detect one of the two suffix vowel phonemes (/ee/ or /oo/; see Table 1) within trained items and untrained items that are consistent or inconsistent with the grammatical regularities of the artificial language. Phoneme detection has been found to be faster in lexical contexts that allow for prediction of upcoming phonemes compared to non-predictive contexts (Chetail & Mathey, 2013; Dell & Newman, 1980; Mehler et al., 1981; Morton & Long, 1976; Nissen & Bullemer, 1987). This is also similar to a congruency effect, where reaction times are faster for congruent (predictive) than incongruent (non-predictive) items.

In our phoneme detection task, untrained but grammatically consistent items, in which the upcoming vowel in the suffix can be predicted from the determiner and semantic (picture) cues, would be considered 'congruent'. In contrast, untrained grammatically inconsistent items would be 'incongruent', as the suffix cannot be predicted from the picture and determiner. Thus, we expected participants to demonstrate a faster reaction time when detecting a phoneme within the untrained consistent items, compared to the untrained inconsistent items. Although the main motivation for developing this task was to provide a sensitive index of implicit knowledge of grammatical regularities, the same task can also provide an online measure of word learning, since trained items that have been learnt should elicit faster responses than untrained items, even if they are consistent. The use of this task to provide a measure of word learning is also supported by findings that phonemes are faster to detect in real words than non-words (Rubin et al., 1976).

A final addition to Experiment 2 was to include a new Determiner Only word-picture matching generalisation task, in order to assess the learning of the specific regularity between the determiner and the semantic category. The robust generalisation in the Determiner & Suffix task in Experiment 1, as well as in previous experiments using this paradigm (Mirkovic et al., 2021; Chapter 2, Experiment 1), strongly suggests that learning the determiner and its association with a semantic category is one of the easiest aspects of the grammatical regularities for adults to learn. However, the Determiner & Suffix task also incorporates all the other cues in the language: the co-occurrence between the determiner and suffix as well as the suffix-semantic association. Isolating the role of the determiner therefore requires a Determiner Only generalisation task. Finally, because the Determiner Only generalisation task, and the phoneme detection task both required new stimuli, and because we only have a finite set of stimuli that meet our design criteria, we decided to repurpose the novel words from the phonological old and new task. Apart from these changes the same artificial language (see Table 1) and experimental design as Experiment 1 was used (see Figure 18).

In addition to the expectation that participants would be able to learn the novel words and at least some of the grammatical regularities, as has been found previously, the specific predictions for Experiment 2 were 1) that there would be an increase in explicit knowledge of grammatical regularities between the first and last blocks of training, as indicated by the 'zero correlation' criterion, 2) that explicit knowledge as indexed by the confidence ratings would be evident for each of the grammatical generalisation tests following training (based on the 'zero correlation' criterion), 3) that implicit knowledge of the regularities would be evident in the confidence ratings for each of the grammatical generalisation tasks, based on the 'guessing criterion', 4) that implicit knowledge of the newly learnt words would be shown in slower responses for untrained compared to trained items in the phoneme detection task, and finally 5) that implicit knowledge of grammatical regularities would be indexed by the phoneme detection task, in slower responses to untrained items that are inconsistent compared to those that are consistent with the language. Specific hypotheses and a data analysis plan for this study were pre-registered and can be found on the OSF (https://osf.io/xdvyg). Two additional analyses to the pre-registered analysis plan have been added and these will be indicated within the results section.

4.3.1 Method

4.3.1.1 Participants

Sixty-three adult participants took part in this experiment. Following pre-registered exclusion criteria, four participants were excluded from data analysis completely. This left a total of fifty-nine participants, 26.78 years (18.25-39.58 with one non-disclosure; 19 male) included in data analysis. Fourteen of these participants were partially excluded from analysis for some of the tasks, again based on pre-registered exclusion criteria. For further details of the exclusions made and the criteria used, please see the OSF pre-registration (or Appendix B: B1 – B3). The adult sample were drawn from the online participant panel ‘Prolific’ and were paid for their participation. The study protocol was approved by the ethics committee at the Department of Psychology, University of York.

4.3.1.2 Stimuli

All non-proprietary materials are provided on OSF. See also Appendix A for full stimuli lists (A1-A2, A4-A12).

The same training and generalisation stimuli sets as Experiment 1 were used for the equivalent tasks conducted in Experiment 2. New word-picture pair stimuli were created for two new tasks. To create the artificial words for the new word-picture pairs, the ‘Phonological Old and New’ task was removed from the experimental protocol and the thirty-two foil items redeployed to the new tasks. Artificial words were created from the English database of pronounceable nonwords (Rastle et al., 2002). The construction of these new words followed the same procedure used in Chapter 2, and Experiment 1 of this chapter. All artificial words used in the current experiment were re-recorded by a native English speaker to avoid presentation differences between the original and new word stimuli. To create the pictures for the new word-picture pairs, an additional thirty-six colour pictures were drawn from Rossion and Pourtois (2001) object database (281x173ppi) and ‘Pixaby’ an online, copyright free image search engine. The below will describe the word sets for new tasks developed specifically for Experiment 2 along with the reintroduction of the mismatched items in the training word-picture matching task.

Word-Picture Matching Training Word Set

To investigate the potential emergence and use of explicit and implicit grammatical knowledge across training, mismatch items were now included into each ‘repetition and word-picture matching’ training tasks following the procedure and construction from Chapter 2. Participants were randomly exposed to the 32 trained word-picture pairs (matched trials) once, interspersed with 8 mismatched items. The mismatched items contained trained words paired with previously unpaired trained pictures, half of the words were mismatched with pictures from the correct semantic category (e.g., *tib* zeap*eem* – tiger) and half mismatched with pictures from the incorrect category (e.g., *tib* mof*eem* – lamp). Four different sets of these mismatched items were used for the four iterations of the ‘repetition and word-picture matching task’ (Figure 18). Constructing mismatched items in this way provided trials where grammatical and not just word knowledge could be used to guide a response.

Word-Picture Matching Generalisation Word Set – Determiner Only

In addition to the three generalisation tasks used in Chapter 2 and Experiment 1 of this chapter, a new fourth generalisation task was constructed, the Determiner Only task. This set of 12 previously unseen words and pictures aimed to test the co-occurrence between the semantic category and the determiner, as well as the distributional co-occurrence with the suffix. Half of the items contained a word-picture pairing with a consistent mapping conforming to the regularities in the training set. For the inconsistent items, the suffix matched the semantic category of the picture, but the determiner did not match either the suffix or the semantic category of the picture. For example, *tib* mal*ool* was paired with a picture of a key; here the co-occurrence of *ool* with the picture of an artefact conformed to the trained regularities, but the determiner *tib* was inconsistent with both the suffix *ool* and the semantic category of artefact (see also Table 30).

Table 30. Inconsistent item construction for the word-picture matching generalisation tasks and phoneme detection task in Experiment 2.

Task	Determiner	Suffix	Picture Category	Example
Determiner & Suffix	***tib***	***eem***	artefact	*tib* darl*eem* - bowl
	ked	***ool***	animal	*ked* soid*ool* - wolf
Determiner Only	***tib***	*ool*	artefact	*tib* mal*ool* - key
	ked	*eem*	animal	*ked* vug*eem* - bird
Suffix Only	*tib*	***ool***	animal	*tib* sen*ool* - goat
	ked	***eem***	artefact	*ked* dorg*eem* - bell
Phonological Form	*tib*	*ool*	-	*tib* jit*ool*
	ked	*eem*	-	*ked* narp*eem*

*Items in **bold** indicate the morphemes which are inconsistent with the semantic category shown in the picture pair (as compared to the training items).*

Phoneme Detection Word Set

A set of new items was designed to test word learning and generalisation performance after training, in an indirect, phoneme detection task. This task includes items from the trained set but further items were created for this task that had not been presented at training. The new, previously unseen and/or unheard items were constructed to create either consistent or inconsistent items.

i) Consistent Items

This set of 32, previously unencountered items were constructed to conform with the trained set (e.g., ked roivool or tib narpeem; see also Table 1), with half using '*ked_ool*' and the other half using '*tib_eem*'. Eight of these consistent items contained just an artificial word with the other twenty-four containing both a word and semantically consistent picture pairing. These 8 word-only items were created to match proportionally with the word-only inconsistent items (see below).

ii) Inconsistent Items

This set of 16, previously unencountered items were constructed to test generalisation through comparison with performance on the consistent items. These items were constructed to be inconsistent with the trained items and were constructed in 4 different ways; following the inconsistent item construction process for the 4 word-picture matching generalisation tasks.

Four of the inconsistent items followed the Determiner and Suffix inconsistent item construction, where both the determiner and suffix are consistent with each other, but inconsistent with the semantic category shown in the picture. The next four followed the Determiner Only inconsistent item construction, where the determiner is inconsistent with the semantic category, but the suffix is consistent with it. Four items followed the Suffix Only inconsistent item construction, where only the suffix was inconsistent with the semantic category, but the determiner is consistent. The final four items followed the Phonological Form inconsistent item construction, where a picture was not incorporated into the item and as such only the co-occurrence between the determiner and suffix was inconsistent with the training items (see Table 30 for examples of all the inconsistent items).

Within all four types of inconsistent items, half were based on the animal semantic category and the other half the artefact category as shown by the picture pairings. This was equivalent to half the items starting with tib and the other half starting with ked for the 'Phonological Form' inconsistent set.

iii) Constructing the phoneme detection word lists

Using the newly created consistent and inconsistent word sets along with the trained word set, 4 lists of 32 items were created. Each list contained a different 8 items from the trained set and a different 8 items from the untrained, consistent set. Half of these were '*ked_ool*' and the other half '*tib_eem*' words. Within these two grammatical categories for the consistent items, one of the items contained just the word-form and the other three contained both a word and picture. The same 16 inconsistent items were added to each of the four lists. These three item types enabled a comparison of reaction times between trained and untrained consistent items to measure word learning; as well as comparisons between untrained consistent and untrained inconsistent items to measure generalisation.

These four lists were counterbalanced across participants. The use of different lists and the counterbalancing procedure aimed to reduce the emergence of explicit grammatical knowledge, which exposure to a large number of consistent items may aid.

4.3.1.3 Procedure

Experiment 2 followed a broadly similar protocol and task order as Experiment 1, with the following differences (Figure 18). Firstly, Experiment 2 was conducted online, using the online experiment platform 'Gorilla' (Anwyl-Irvine et al., 2020). Although every effort was made to keep task procedures as close to those used in Experiment 1, using this new testing software meant that some timing and response procedures did differ, as detailed below for individual tasks. All tasks started with written instructions presented on the screen, which participants could move through by clicking a 'next' or 'previous' button at the bottom of the screen. All tasks were started by clicking a 'start' button on screen rather than pressing the spacebar. An additional debriefing questionnaire was added to the end of the study, asking if participants had experienced any technical issues as well as an honesty question to aid in implementing the pre-registered exclusion criteria.

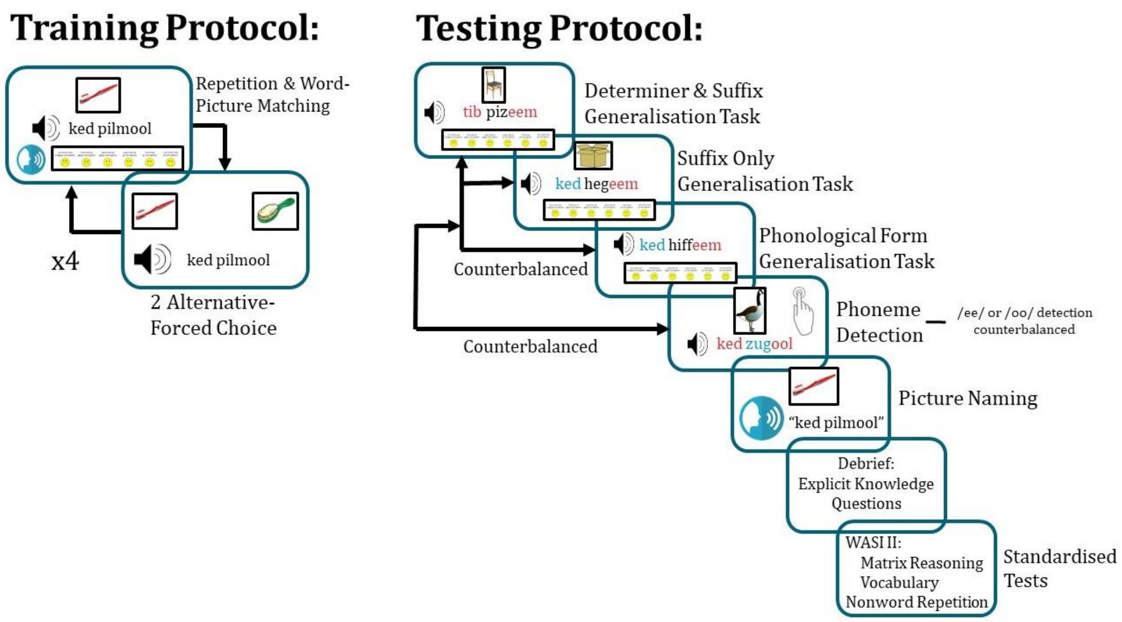

Figure 18. Experiment 2 procedure. The generalisation tasks and phoneme detection task depict an inconsistent item.

Training Tasks:

Repetition and Word-Picture Matching

The procedure for this task was the same as Experiment 1 but with two exceptions. Firstly, participants' repetitions were recorded with a time limit for a response. Participants moved onto the second, word-picture matching response section of the trial once 4000 ms had elapsed. Recording participants allowed for quality checking. Including a time limit aimed to reduce participants ability to write down any of the words, but with enough time for a verbal repetition to occur. Secondly, the word-picture matching response procedure differed. For this second section of the trial, the stimuli picture stayed on the screen, the audio recording of the word was repeated and the smiley/sad face scale pictures appeared at the bottom of the screen as before. This time the scale wording was also included above the smiley/sad face scale to mimic this being available next to the computer for Experiment 1 participants. As noted, the scale now incorporates six points.

For the 'sad faces' (reject): a 'very sad face' denoted the option '*I am certain the word and picture don't go well together',* a 'regular sad face' denoted *'I think the word and picture don't go well together'* and a 'slightly sad face' denoted *'I guess the word and picture don't go well together'*. For the 'happy faces' (endorse): a 'slightly happy face' denoted the option *'I guess the word and picture go well together',* a 'regular happy face' denoted *'I think the word and picture go well together'* and a 'very happy face' denoted *'I am certain the word and picture go well together'* (Figure 11). Participants could indicate their choice by clicking on the appropriate scale-point picture.

Across the four iterations of this task during training (see Figure 18), participants were exposed to matched training items 8 times (once in the repetition section of the trial and once in the word-picture matching section of the trial. Each mismatched item was presented once across the training protocol.

Testing Tasks:

Word Learning: 2AFC

This task's procedure was the same as Experiment 1, except participants gave their response by clicking on one of the two pictures presented on the screen.

Word Learning: Picture Naming

This task's procedure was the same as Experiment 1 but with two exceptions. Firstly, responses were recorded by Gorilla for later scoring by the experimenter using the same scoring procedure from Experiment 1. Secondly, a 6000ms time limit for a response was introduced to prevent participants accessing any potential written material they may have taken. Thus, participants moved onto the next trial once this time limit had elapsed. The scoring of responses followed the procedure from Experiment 1 of this chapter.

Word Learning and Generalisation: Phoneme Detection Task

This task aimed to test word learning and grammatical generalisation within an online, processing context after the learning phase. To measure this, participants were asked to detect one of the two phonemes present in the artificial language's suffix, /oo/ for '*ool'* or /ee/ for '*eem'*. The phoneme to detect was counterbalanced across participants. This task was conducted after training, with order counterbalanced with the 'word-picture matching generalisation' tasks (Figure 18). This task used one of the four 'phoneme detection' word lists which were counterbalanced across participants (as described in the stimuli section). The 8 training items, 8 untrained consistent items and 16 untrained inconsistent items were presented randomly in one block.

Each trial consisted of an orientation cross displayed for 500 ms, followed by an audio-only presentation of a word and, if present in the item, the simultaneous presentation of the paired picture. Participants were instructed to press the spacebar on their computer keyboard as quickly as possible if they heard the target phoneme (/ee/ or /oo/ depending). If they did not hear the phoneme they were instructed to not press or click any buttons. Participants' responses and reaction times were recorded and if a picture was present it stayed on screen until the end of the trial. The next trial commenced after a response was made or 2000 ms had elapsed.

The task started with two practice trials, to help introduce the task and phoneme to be detected. The practice items consisted of the same semantic category as before (fruit pictures) but now used artificial words which contained one of the target phonemes (/ee/ or /oo/) in the 'suffix' position, but did not conform to the regularities present at training (e.g., *edd harleesh* and *dyt pryloog*). After the practice trials participants could click the 'next' button to start the task. This task took approximately 3 minutes to complete with reaction times for correct phoneme detection (hits) being used as the outcome measure.

Generalisation: Word-Picture Matching

The word-picture matching generalisation tasks took place after the training protocol, counterbalanced with the Phoneme Detection task. The order of the four generalisation tasks were also counterbalanced for each participant (Figure 18). The Determiner and Suffix, Suffix Only and Phonological Form tasks used the same procedure from Experiment 1 but with the same response scale exception detailed in the 'Repetition and Word-Picture Matching' training task (e.g., six-point instead of five-point response scale; Figure 11). The new 'Determiner Only' task followed the same procedure as the 'Determiner and Suffix' and 'Suffix Only' tasks.

Explicit Knowledge Questionnaire:

This questionnaire followed the same procedure as Experiment 1, except responses were typed by participants into 'Gorilla'. The same scoring system as Experiment 1 was also used (Table 3).

4.3.1.4 Data Analysis

All analyses were conducted in R (R Core Team, 2020). All data, scripts and outputs are provided on OSF. The presented analysis follows the pre-registered analysis plan. Effect sizes will be reported as Cohen's *d* or f^2 (Cohen, 1992).

2AFC Analysis

To assess changes in accuracy across training for the 2AFC and training word-picture matching tasks, linear, mixed-effect regressions were conducted (using the lmer function in R). The same polynomial contrasts as Experiment 1 were set to the four 2AFC and word-picture matching training tasks, using default contrast matrices within the contrast function in R (contrast.poly; see Schad et al., 2020). This enabled both linear and quadratic comparisons of the tasks.

Word-Picture Matching Analysis

For both the training and generalisation word-picture matching tasks, all sad face responses were considered 'rejections' and all happy face responses were considered 'endorsements'. Using this, the same procedure from Experiment 1 was used to calculate A' scores for the word-picture matching generalisation tasks for use in analysis. To analyse the confidence ratings for the word-picture matching training and generalisations task, responses were converted to a score of 0, 1 or 2 based on Channon et al.'s (2002) procedure. The '*I am certain'* judgements for both the sad (*'does not go well with'*; reject) and happy ('*does go well with'*; endorse) scale points were considered 'high confidence' and were scored '2'. The '*I think'* judgements for both the sad and happy scale points were considered 'low confidence' and were scored '1'. The '*I guess'* response for both the sad and happy judgments were considered 'guessing' and were scored '0'. These scores were used for the 'zero correlation' and 'guessing criterion' analysis.

To emphasise, as for Experiment 1, different metrics were used to assess generalisation performance and explicit grammatical knowledge use within the direct, word-picture matching measure of generalisation. Generalisation performance within the word-picture matching task was assessed using the A' scores that were calculated using endorsement rates for consistent and inconsistent items. For the zero-correlation analysis, confidence ratings were compared between all correct and all incorrect responses. Correct responses included endorsing consistent and rejecting inconsistent items and incorrect responses included endorsing inconsistent items and rejecting consistent items. Additionally, the guessing criterion analysis was conducted on accuracy performance and not the A' score.

For analysis considering the emergence of explicit and implicit grammatical knowledge use during the word-picture matching training task, analysis was only conducted on the mismatched items that used semantically incorrect word-picture pairs. The training items and items mismatched within the same semantic category were not used as accurate responses for these items depended on word knowledge only. Thus, they could not assess the use of grammatical knowledge. Items mismatched by semantic category allowed for the use of grammatical knowledge to inform a response, although it is acknowledged that word knowledge can also contribute to accurate responses. However, these items were designed to give some insight into the emergence of explicit and implicit grammatical knowledge over the course of training. Confidence ratings were analysed using both ‘zero correlation’ and ‘guessing criterion’ to investigate this.

Phoneme Detection Analysis

Analysis of the phoneme detection task was conducted on reaction times for correct ‘hit’ responses only; that is, when participants correctly detected their target phoneme when it was present in an item, using a ‘go, no-go’ procedure. This task was designed to assess word learning and grammatical generalisation. To assess word learning, reaction times were compared for trained and untrained consistent items. To assess grammatical generalisation, reaction times for untrained consistent and untrained inconsistent novel items were compared. A mixed-effects regression was used to compare the three item types (using the lmer function in R). To do this, a set of contrasts using simple comparisons (contrast.treatment) were assigned to the three item types which were then inputted as fixed factors within the model. The first contrast compared ‘untrained consistent’ (0) with ‘trained’ (1) items to give an indirect measure of word learning. The second contrast compared ‘untrained consistent’ (0) with ‘untrained inconsistent’ (1) items to give an indirect measure of grammatical generalisation. Contrasts were set using default contrast matrices within the contrast function (see Schad et al., 2020).

4.3.2 Results

4.3.2.1 Direct measures of learning and generalisation

Word Learning: 2AFC

As for experiment 1, performance on the 2AFC task conducted at the end of training was one measure of word learning. One-sample t-tests against chance (.5) showed that the participants performed significantly above chance (M=0.78, SD=0.14; $t(58)$=18.45, p<.0001, d=2.40), demonstrating a good level of novel word learning (Figure 19).

In addition, changes in word learning during training were measured using the four 2AFC tasks conducted across training. A mixed effects regression analysis was conducted to compare accuracy (outcome variable) across the four 2AFC tasks (fixed factors set as polynomial contrasts). As was found in Experiment 1, there was a significant linear increase in performance, as well as a significant quadratic pattern across training (Table 31). That is, a linear increase in performance across early training blocks began to plateau towards the end of training (Figure 19).

Word Learning: Word-picture matching training tasks (WPM)

Word learning performance change during training was also measured using the four WPM tasks conducted across training (Figure 20). The same mixed effects regression analysis used to analyse 2AFC performance was conducted to compare accuracy (outcome variable) across the four WPM tasks in the training phase (fixed factors set as polynomial contrasts). The results of this regression can be seen in Table 32. This analysis showed a significant linear increase in performance across training as well as a significant quadratic pattern. Similar to the 2AFC task results, performance increased across training but with a levelling off by the end as can be seen in Figure 20.

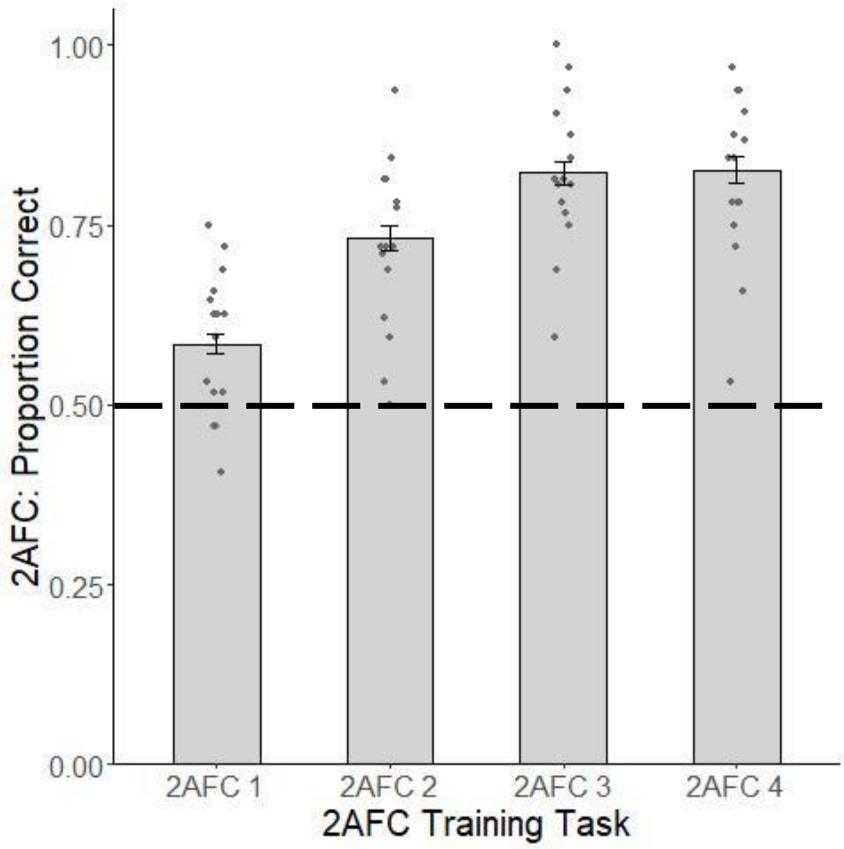

Figure 19. Word Learning: Accuracy on the 2AFC tasks across training in Experiment 2.

Word Learning: Picture Naming

Vocabulary knowledge was also assessed using picture naming, a language production-based task. As in Experiment 1, this task allowed us to separately assess the learning of the stem, and the learning of the determiners and the suffixes. Similar to Experiment 1, stem recall was low relative to the recall of the grammatical morphemes (Figure 21), likely due to the difference in frequency of exposure. In an additional analysis to the pre-registered plan, stem recall was

Table 31. Regression Comparison for word learning performance on the 2AFC task across training in Experiment 2 using polynomial contrasts.

	Estimate	Std. Error	t	*p*	f^2
Linear Contrast	**0.18**	**0.01**	**15.65**	**<.0001**	**0.56**
Quadratic Contrast	**-0.07**	**0.01**	**-6.14**	**<.0001**	**0.09**

Significant results are highlighted in bold.

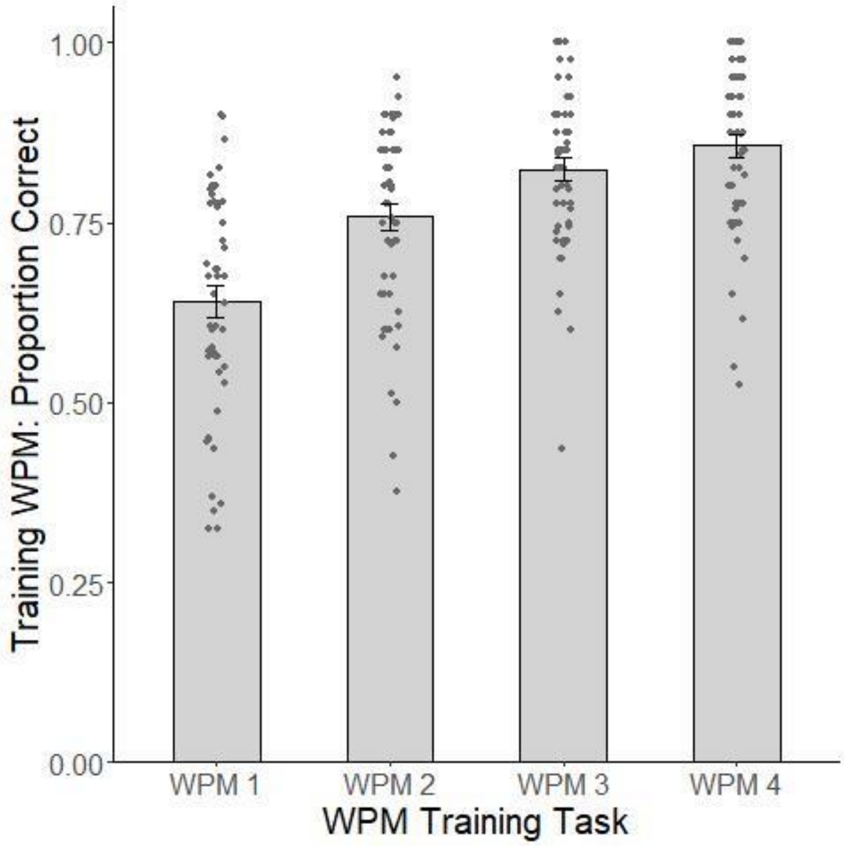

Figure 20. Word Learning: Accuracy on the word-picture matching task across training in Experiment 2.

Table 32. Regression comparison for word learning performance for the repetition and word-picture matching task across training in Experiment 2 using polynomial contrasts.

	Estimate	Std. Error	t	*p*	f^2
Linear Contrast	**0.16**	**0.01**	**13.74**	**<.0001**	**0.39**
Quadratic Contrast	**-0.04**	**0.01**	**-3.64**	**<.0001**	**0.03**

Significant results are highlighted in bold.

compared against 0 using a one-sample t-test to assess if there was a significant level of learning for this word element. This analysis suggests that there was a significant level of learning for the stem (t(57)=8.64, *p*<.0001, *d*=1.13).

Recall of the determiners and suffixes were similar, with a small but non-significant numerical advantage for determiners (t(57)=1.75, *p*=.086, *d*=0.23). The similar levels of recall for each morpheme may be a result of equal exposure to these two morphemes. However, the trend towards better determiner recall is consistent with the picture naming findings from Experiment 1 and may suggest that determiners are easier for participants to learn than suffixes.

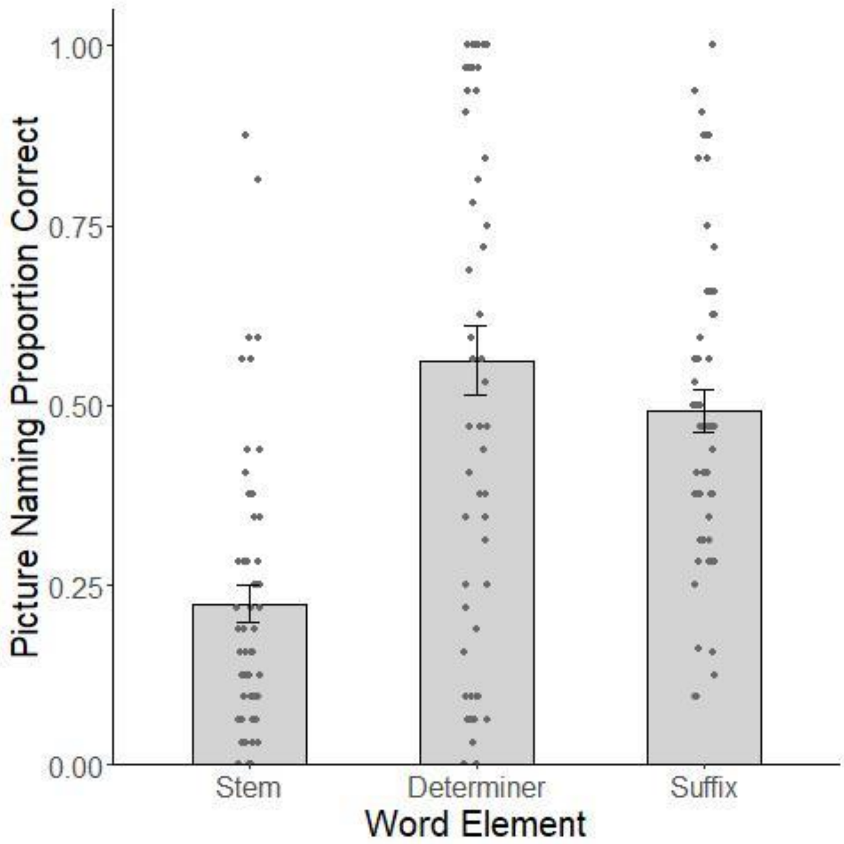

Figure 21. Word Learning: Recall accuracy for the picture naming task in Experiment 2.

Grammatical Generalisation: Word-picture matching generalisation tasks

To analyse performance in the word-picture matching generalisation tasks, we again derived an A' metric based on the endorsement rates for consistent and inconsistent trials as described in the method's section. Figure 22 shows generalisation performance across all four generalisation tasks.

To test generalisation performance, one-sample t-tests against an A' of 0.5 were conducted for each task. Participants demonstrated evidence of successful generalisation in the Determiner and Suffix task ($t(55)=7.86$, $p<.0001$, $d=1.05$), Determiner Only task ($t(55)=6.14$, $p<.0001$, $d=0.82$) and the Suffix Only task ($t(55)=2.32$, $p=.012$, $d=0.31$) but not in the Phonological Form task ($t(55)=0.43$, $p=.335$, $d=0.06$). These results replicate and extend the findings from Experiment 1, suggesting that adults are able to learn and generalise distributional, phonological, and semantic statistical regularities, but only when the semantic cue is present.

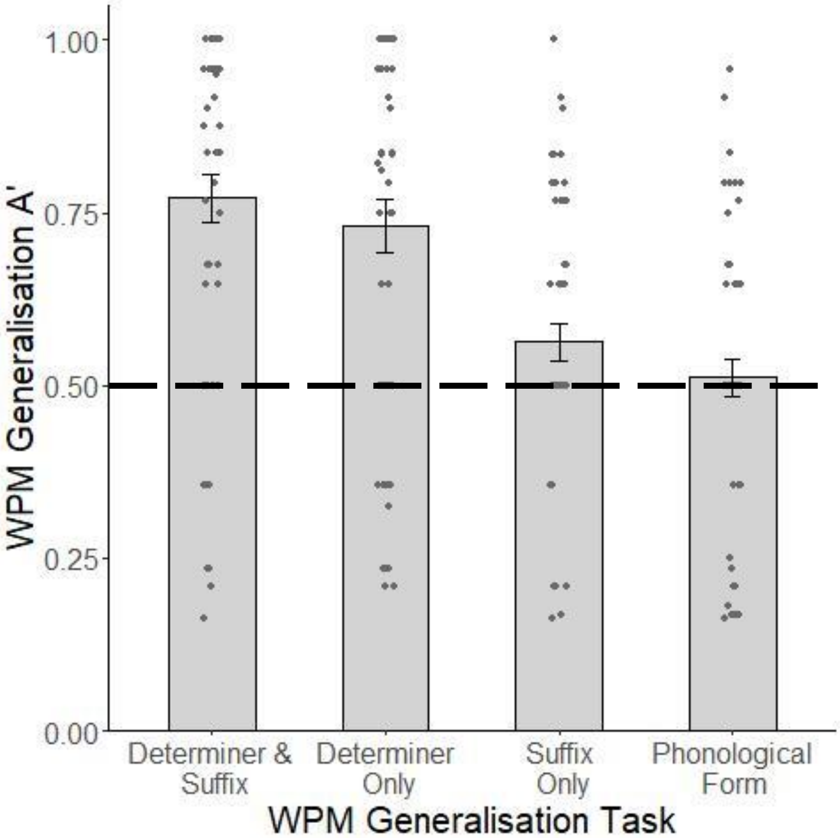

Figure 22. Generalisation: A' performance across all word-picture matching generalisation tasks for Experiment 2.

4.3.2.2 Explicit grammatical knowledge and grammatical generalisation within direct measures of learning:

The role of explicit knowledge in grammatical generalisation was again explored within the word-picture matching generalisation task in two ways. First, by replicating Experiment 1 and previous analysis (Mirkovic et al., 2021; Chapter 2 of this thesis) which considered the contribution of retrospective, verbally reported explicit knowledge on generalisation. Secondly, by using within-task confidence ratings.

Retrospective, verbally reported explicit grammatical knowledge and its contribution to direct measures of grammatical generalisation.

Retrospective explicit knowledge was measured using the 'Debriefing Questionnaire' run at the end of the experiment. Answers were scored separately for determiner and suffix knowledge (see Table 3). Table 33 shows the descriptive statistics for explicit knowledge scores where reported explicit knowledge of determiners is numerically higher than for suffixes. This pattern is similar to those found in Chapter 2 (Experiment 1) for adults who received the same level of exposure to the language (determiner M = 2.39; suffix M = 0.59) and in Experiment 1 of this study (Table 28). Whilst the same pattern has emerged, scores overall are numerically lower in this experiment than for previous experiments. This could be a result of testing online rather than in person.

Table 33. Descriptive statistics for verbally reported explicit knowledge in Experiment 2.

	M	SD
Determiner	1.58	1.42
Suffix	0.21	0.65

As before, the contribution of explicit knowledge to generalisation performance was considered at the individual level by conducting a regression analysis for each generalisation task. The outcome variable was the A' score for each task and the predictor variable(s) were the relevant morpheme awareness score(s). Table 34 shows the results of this analysis.

A significant contribution of explicit knowledge of the determiner was found for performance in the Determiner & Suffix and Determiner Only tasks. A significant contribution of explicit suffix knowledge was found for performance in the Phonological Form task. Evidence was not found for a contribution of determiner knowledge to performance in the Phonological Form task and suffix knowledge did not significantly contribute to performance in the Determiner and Suffix task or in the Suffix Only task. Thus, participants demonstrated a significant role for retrospective explicit knowledge on generalisation performance in all bar the Suffix Only task.

Within task measures of explicit grammatical knowledge use for direct measures of grammatical generalisation.

Zero-correlation analysis was again used to assess the use of explicit knowledge within the generalisation tasks. Responses from the six-point scale were used to compare mean confidence ratings for correct and incorrect responses (Figure 23).

A significantly higher mean confidence rating for correct compared to incorrect responses was found within the Determiner and Suffix task ($t(42)=2.03$, $p=.049$, $d=0.31$). A significant difference was not found for the Determiner Only task ($t(40)=0.01$, $p=.992$, $d=0.00$), the Suffix Only task ($t(55)=0.56$, $p=.579$, $d=0.07$) or the Phonological Form task ($t(56)=-0.18$, $p=.856$, $d=-0.02$). These results replicate the findings of Experiment 1 and provide evidence for within-task use of explicit grammatical knowledge for the Determiner and Suffix task, but not the Determiner Only, Suffix Only or Phonological Form tasks. It also suggests the potential use of implicit grammatical knowledge within the Determiner Only and Suffix Only tasks where participants demonstrated successful generalisation at the group level (Figure 23). Although a direct measure of implicit knowledge is needed to support this inference.

Table 34. Multiple regressions for the contribution of verbally reported explicit knowledge on word-picture matching generalisation performance in Experiment 2.

	R^2	BSE	B	ß	*p*	f^2
Determiner & Suffix Generalisation:	0.47					
Determiner		**0.02**	**0.13**	**0.70**	**<.0001**	**0.95**
Suffix		0.04	0.02	0.04	.677	0.00
Determiner Only Generalisation:	0.25					
Determiner		**0.02**	**0.10**	**0.52**	**<.0001**	**0.36**
Suffix Only Generalisation:	0.03					
Suffix		0.04	0.07	0.22	.107	0.05
Phonological Form Generalisation:	0.08					
Determiner		0.02	0.01	0.08	.552	0.01
Suffix		**0.04**	**0.10**	**0.32**	**.017**	**0.12**

Significant results are highlighted in bold.

Emergence of explicit grammatical knowledge use during training

To assess the emergence of explicit knowledge during training the following analysis was run on the first and last (1 and 4) training word-picture matching tasks. As discussed in the method section, only items mismatched by semantic category were used in this analysis. A ‘zero-correlation’ analysis was carried out across both the first and last training word-picture matching tasks, using the confidence ratings from the six-point scales responses (Figure 24). To run the ‘zero-correlation’ analysis, a mixed effect, 2x2 ANOVA was conducted, comparing mean confidence ratings across the fixed effects of task (first and last task), and accuracy (correct, incorrect). A task x accuracy interaction would indicate a significant difference for mean confidence ratings between correct and incorrect responses over the course of training.

Significant main effects of task ($F(1,65.10)=15.51$, $p<.0001$, $f^2=0.11$) and accuracy ($F(1,48.19)=18.01$, $p<.0001$, $f^2=0.02$) were found; the interaction effect just missed significance ($F(1,64.09)=3.95$, $p<.051$, $f^2=0.02$). These results mean that overall, confidence ratings were higher for the final training word-picture matching task and higher overall for correct compared to incorrect responses across tasks, with a trend showing an increase in confidence for correct responses at the end of training, compared to the first training block (illustrated in Figure 24).

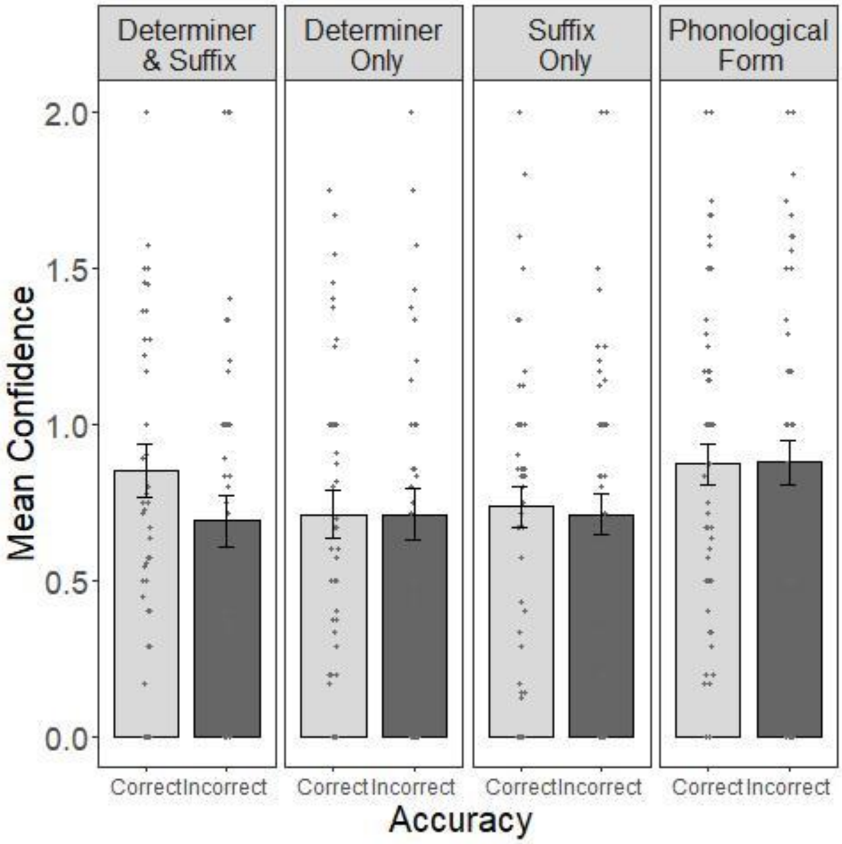

Figure 23. Explicit knowledge: Mean confidence ratings for correct and incorrect responses across word-picture matching generalisation tasks for Experiment 2.

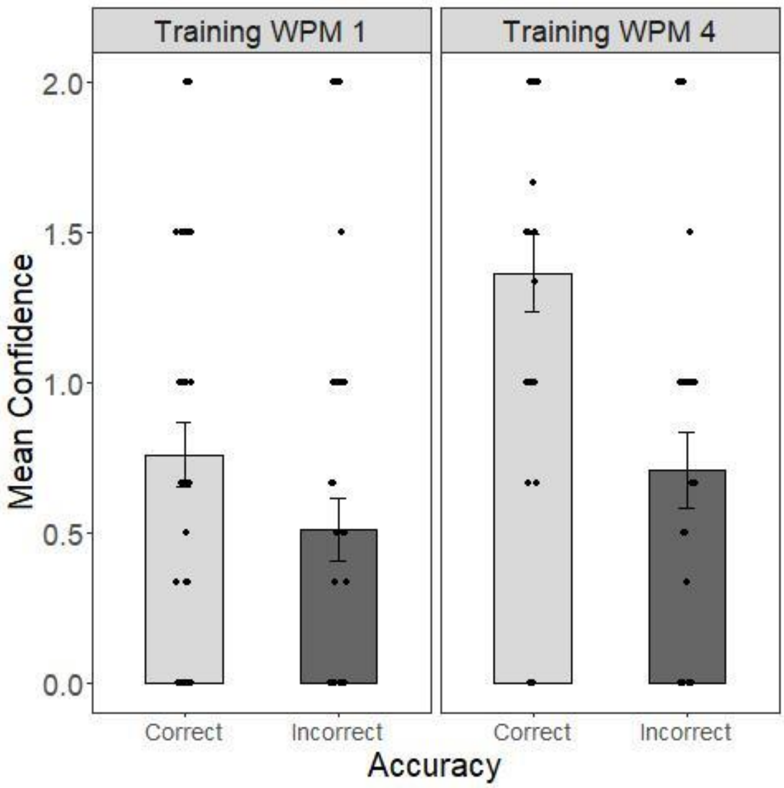

Figure 24. Explicit knowledge: Mean confidence ratings for correct and incorrect response for the first and last word-picture matching training tasks in Experiment 2.

4.3.2.3 *Implicit grammatical knowledge and grammatical generalisation within direct measures of learning*

Implicit grammatical knowledge using confidence ratings was explored in two ways. Firstly, its use was considered within the word-picture matching generalisation tasks. Secondly, the emergence of implicit knowledge across the training word-picture matching tasks was considered. Both investigations used a 'guessing criterion' analysis, which considers the accuracy of guess responses (0) only, as extracted from the six-point response scale. If accuracy is above chance for these responses it is considered evidence for implicit knowledge use.

Within task use of implicit grammatical knowledge for direct measures of grammatical generalisation

A 'guessing criterion' analysis was run on each word-picture matching generalisation task. To do this, guess only responses were extracted for each task and one-sample t-tests against chance (0.5) were conducted on the accuracy of these guess responses (not the A' scores). Figure 25 shows accuracy performance for these guess responses across all of the word-picture matching generalisation tasks.

Participants performed significantly above chance for guess responses in the Determiner Only task ($t(32)=4.45$, $p<.0001$, $d=0.77$), Determiner and Suffix task ($t(32)=2.81$, $p=.004$, $d=0.49$), Suffix Only task ($t(32)=2.54$, $p=.008$, $d=0.44$) and the Phonological Form task ($t(32)=1.99$, $p=.028$, $d=0.35$). These results provide evidence that participants were using implicit knowledge of the grammatical regularities across all of the word-picture matching generalisation tasks.

Emergence of implicit grammatical knowledge use during training

To assess the emergence of implicit knowledge during training the guessing criterion analysis was carried out on the first and last (1 and 4) training word-picture matching tasks (Figure 26).

Participants did not perform significantly above chance in the first task ($t(19)=0.86$, $p=.200$, $d=0.19$) or the final task ($t(19)=-1.40$, $p=.911$, $d=-0.31$). These results do not provide evidence for the emergence of implicit grammatical knowledge during training, despite the fact that there was evidence of implicit knowledge being used during the grammatical generalisation tests which immediately followed training. However, this may be due to the availability of explicit knowledge of the trained items in this task.

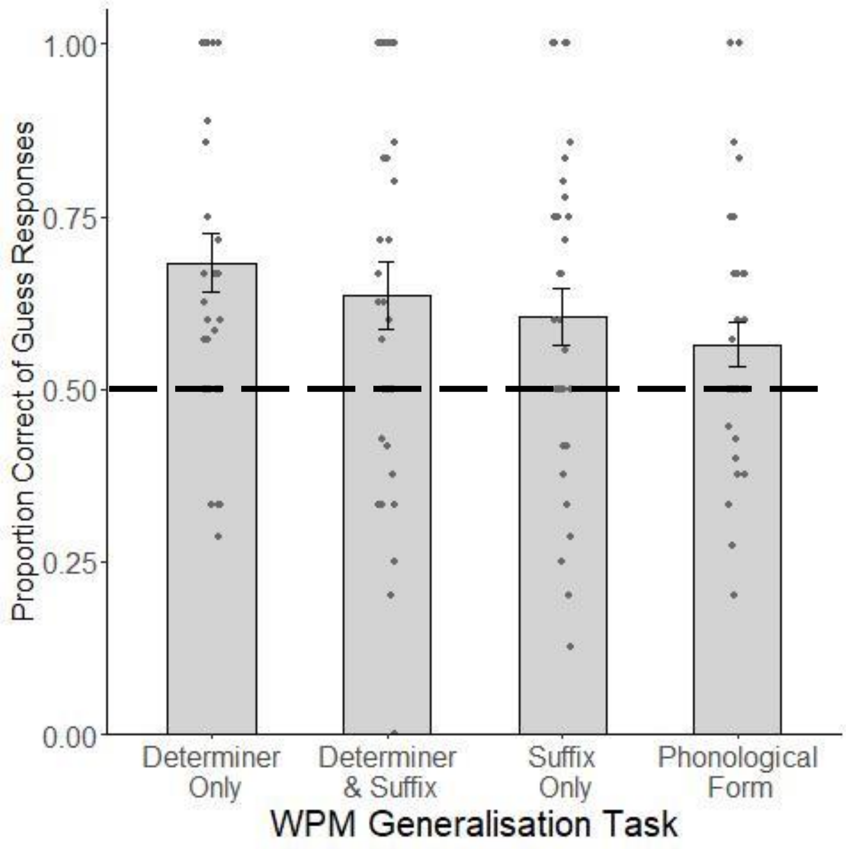

Figure 25. Implicit knowledge: Accuracy for guess responses across all word-picture matching generalisation tasks in Experiment 2.

4.3.2.4 Phoneme detection task: Indirect measure of word learning and grammatical generalisation

Both word learning and grammatical generalisation were assessed using the Phoneme Detection Task. Reaction times for hit responses were compared between three item types: trained, untrained consistent and untrained inconsistent (illustrated in Figure 27). A mixed-effects regression was conducted with the outcome variable set as hit reaction times and the predictor variables set as the simple (treatment) contrasts applied to the three item types as discussed in the method section. Table 35 shows the results of this analysis.

Reaction times were significantly faster for 'trained' items when compared with 'untrained consistent' items. This result demonstrates a word learning effect and provides evidence of implicit knowledge of the trained novel words. Contrary to our experimental prediction however, reaction times were not significantly different between untrained inconsistent and consistent items, meaning that participants did not demonstrate evidence of generalising grammatical regularities in this task.

Pre-registered analysis comparing the reaction times of the four types of inconsistent items within the phoneme detection task had been planned. However, this analysis was dependent on finding a significant generalisation effect (faster reaction times for untrained consistent than untrained inconsistent items) in the previous analysis. As a generalisation effect was not found, this analysis was not conducted.

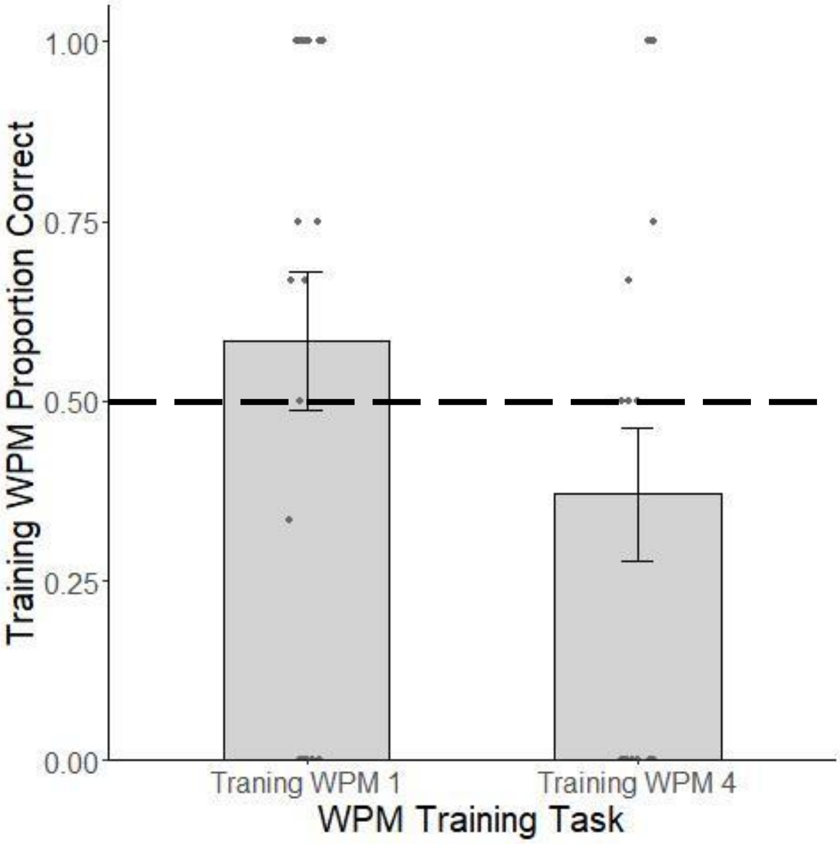

Figure 26. Implicit knowledge: Accuracy for guess responses for the first and last word-picture matching training tasks in Experiment 2.

4.3.3 Discussion

Experiment 2 replicated previous findings, including from Experiment 1, in a number of areas. Firstly, participants demonstrated a good level of word knowledge in the 2AFC task, with performance showing a linear increase across training before levelling off by the end of training. This pattern of word learning was also shown in the word-picture matching tasks during training. Secondly, participants also demonstrated evidence for successful generalisation performance in the word-picture matching tasks following training, for regularities that involve semantics but not when semantics were absent (phonological form task). Next, participants in Experiment 2 also showed a partial contribution of verbally reported explicit grammatical knowledge on generalisation as measured by the word-picture matching generalisation tasks. Finally, evidence for explicit knowledge using the confidence rating measure was found for the Determiner & Suffix word-picture matching generalisation task, but not for any of the other tasks. Going beyond previous findings, Experiment 2 also considered the emergence of explicit knowledge across the word-picture matching training tasks. When confidence ratings were compared between the first and last word-picture matching task, there was an overall increase in confidence ratings across

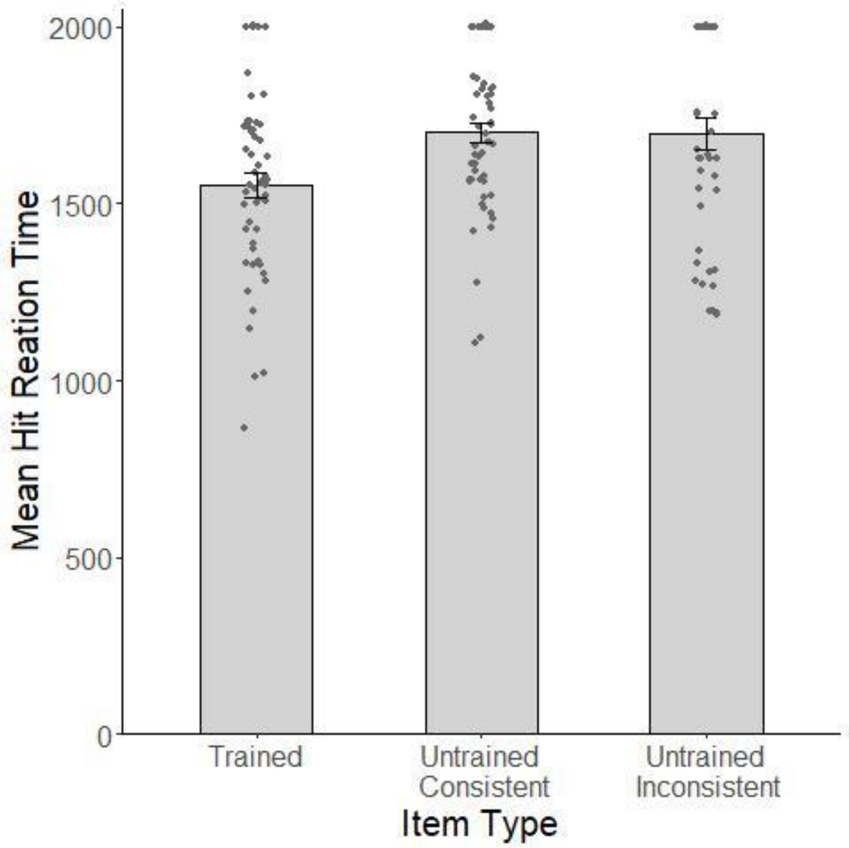

Figure 27. Implicit knowledge: 'Hit' reaction times for the three item types in the phoneme detection task in Experiment 2.

Table 35. Regression comparison for the hits in the phoneme detection task in Experiment 2 using simple (treatment) contrasts.

	Estimate	Std. Error	t	*p*	f^2
Trained vs. Untrained Consistent	**-148.31**	**46.70**	**-3.18**	**.002**	**0.06**
Untrained Consistent vs. Untrained Inconsistent	-1.78	48.60	-0.04	.971	0.00

Significant results are highlighted in bold.

the two tasks with overall higher confidence ratings for correct compared to incorrect responses. There was also a non-significant trend for greater confidence in accurate responses at the end of training, compared to the beginning, which potentially suggests both that explicit awareness emerges early in training, and also that it increases over the course of training. However, only a small number of items could be used for this analysis (as it focused on ‘mismatched’ items), and as these items were part of the training set, it is possible that responses reflected item-level knowledge rather than knowledge of the grammatical regularities per se. Alternative methods to investigate the time course of the emergence of explicit knowledge would be an interesting avenue for future studies to explore.

In addition to considering explicit knowledge, a major aim of Experiment 2 was to also examine the presence and use of implicit knowledge in grammatical learning and generalisation. Evidence of implicit knowledge was found within all the direct generalisation tasks that followed training, although not during the training tasks themselves. In contrast, the phoneme detection task did not show evidence of a grammatical generalisation effect, although participants did demonstrate word learning in this task. This suggests that participants have learnt the training words to a high enough level to utilise implicit word knowledge, but that implicit grammatical knowledge is either not present or is not robust enough for use within this task.

While the word learning findings in this current experiment were predicted and, in most respects, replicated Experiment 1, it is worth noting that overall levels of word knowledge were slightly lower (e.g., for the final 2AFC task M = .78, compared to M = .88 in Experiment 1, and M = .91 in Chapter 2 Experiment 1). This could reflect the different testing environments between face-to-face and online testing and/or the more variable population used in Experiment 2 (drawn from the Prolific participant pool, rather than undergraduate students). It was also predicted that participants would demonstrate successful generalisation across all word-picture matching generalisation tasks. As with Experiment 1, this prediction was only partially supported as participants did not demonstrate successful generalisation in the phonological form task. While this finding, like Experiment 1, diverges from Chapter 2 in terms of significance, it does follow the general pattern for this task to show the lowest A' score compared to the other generalisation tasks. Taken together, it seems likely that the finding of successful generalisation in the phonological form task in Chapter 2 was a false positive.

The main aims and predictions for Experiment 2 centred around the presence and use of explicit and implicit grammatical knowledge. With respect to explicit knowledge, both verbal reports and the 'zero-correlation' analysis on the generalisation tasks yielded evidence of explicit contributions to grammatical generalisation. In addition, the 'guessing criterion' revealed that implicit knowledge was being used in every one of the grammatical generalisation tasks. This supports and extends the findings from Experiment 1 as well as from Batterink et al. (2015), and attests to the implicit nature of this statistical learning paradigm. It also supports the idea that 'direct' tests can utilise both explicit and implicit knowledge when judgements are being made.

However, contrary to our predictions, the phoneme detection task, which we expected would be able to capture online implicit processing of the learnt grammatical regularities, did not reveal evidence of a generalisation effect. It did show the predicted word learning effect (faster responses to trained compared to untrained consistent items), indicating that the failure of this task to show grammatical learning was not due to an inability to capture learning *per se*. This means we have not been able to extend Batterink et al.'s (2015) 'indirect' task findings into a grammatical paradigm. So, while we know that our participants are able to learn the grammatical regularities of our novel language, and the 'guessing criterion' analysis demonstrated that implicit knowledge contributes to that learning in our direct task, evidence of implicit knowledge detection was not found within the phoneme detection task.

There are a number of potential reasons for this result. One potential reason is the complexity of our task, which included four different types of inconsistent items. Furthermore, rather than simply being 'predictable' or 'unpredictable', some of our items were 'partially predictable'; for example, in the 'Determiner & Suffix' construction, the suffix could be predicted on the basis of the determiner but be inconsistent with the picture, or vice versa for the 'Determiner Only' construction. This is in contrast to the speeded target detection task used by Batterink et al. (2015), in which items were either predictable or not. Similarly, in Misyak et al. (2010b), following a serial reaction time task procedure, all items were predictable for the participant, and they only needed to decide between two choices.

In a similar vein, the artificial language used in this paradigm is more complex than the one used in Batterink et al. (2015) and Misyak et al. (2010b). Both of these studies only considered regularities within the phonological domain (although this was also orthographically represented by Misyak et al., 2010b). In contrast, our paradigm incorporates multiple cues, including semantic regularities. Although the advantage of this complexity is that it is more naturalistic, and akin to that found in natural languages, it may be that it takes longer for learning - perhaps particularly implicit learning - to take place. However, this explanation cannot account for our finding of implicit contributions to performance in the generalisation tasks, using the guessing criterion. A final possibility may be related to the fact that our phoneme detection task focused on prediction of the suffix. Indeed, as language processing proceeds in real time, it has to be the later elements (in our case the suffixes), that are predicted by preceding information (the determiner, and the semantic cue/picture). However, we also know from our previous experiments that learning of the suffix in this paradigm is less robust than learning of the determiner, and it may be that the implicit representation of the suffix is not yet robust enough to support online processing.

4.4 General Discussion

The main aim of this study was to explore the role of explicit and implicit grammatical knowledge in grammatical generalisation within a statistical learning paradigm. Evidence was found for retrospective, verbally reported explicit knowledge contributing to grammatical generalisation in both experiments, which replicates previous findings (Mirkovic et al., 2021; Chapter 2 of this thesis). Extending this finding, confidence rating measures also showed evidence of within-task use of explicit knowledge, specifically for the mapping between the determiner/suffix and the semantic category (Determiner & Suffix task). Taken together, these findings support the growing literature demonstrating a role for explicit knowledge within adult statistical learning (Batterink et al., 2019; Batterink et al., 2015; Franco et al., 2011; Mirkovic et al., 2021; Monaghan et al., 2019 and Chapter 2 of this thesis) and extends it to grammatical generalisation.

Within the 'direct' generalisation tasks, the lack of relationship between accuracy and confidence in the Suffix Only task, despite successful generalisation, suggests the use of implicit knowledge. However, this can only be inferred when using a zero-correlation analysis and runs the risk of overinterpreting the null (Dienes, 2015). Experiment 2 thus expanded on this result by finding more direct evidence of implicit knowledge by introducing a 'guessing' response to the confidence rating measure. This revealed the use of implicit knowledge in all the generalisation tests, and supports the idea that both implicit and explicit knowledge can be used when responding within 'direct' style tasks (Franco et al., 2011; Jacoby, 1991). In particular, it suggests that both implicit and explicit knowledge is used by adults within statistical learning tasks (Batterink et al., 2015), supporting and extending this to a grammatical paradigm. Experiment 2 also used a newly developed 'indirect' reaction time task as an additional index of implicit knowledge. However, this did not show evidence of implicit knowledge of the regularities.

Although across both experiments we found clear evidence of both explicit and implicit contributions to grammatical learning, there were some unexpected discrepancies which need to be considered. In both experiments, within-task explicit knowledge was only found for the Determiner and Suffix task. This is despite retrospective, verbally reported explicit knowledge being found to contribute to performance the Suffix Only task in Experiment 1 and all bar the Suffix Only task in Experiment 2 (and in the Suffix Only and Determiner & Suffix task in Experiment 1 of Chapter 2). Discrepancies between these two measures of explicit knowledge are not entirely unexpected, as the within-task confidence ratings measure is proposed to be a more sensitive measure than verbal reports (Dienes & Perner, 1999; Tunney & Shanks, 2003; Wierzchoń et al., 2012). However, if it was more sensitive, it should arguably have revealed use of explicit knowledge in more - not fewer - of the generalisation tasks than verbal reports.

One potential reason for this discrepancy is methodological: the association between verbal reports of explicit knowledge and grammatical generalisation is an individual differences analysis, in which the focus is on individual variability in awareness and performance. The 'zero-correlation' analysis of confidence ratings, on the other hand, is a group-level analysis which focuses on mean levels of confidence and performance across the group as a whole, but is not sensitive to individual variability within the group. A second possibility is that the emergence and use of explicit knowledge in adults may be related to the differences found between regularities when generalising. In this study, and in previous experiments using the same paradigm, the most robust evidence of generalisation has always been found for the Determiner & Suffix task. It may be that explicit awareness is more likely to emerge for more robust representations or memory traces; alternatively, it may be that the regularities indexed by the Determiner & Suffix task are particularly accessible to explicit awareness, and that this in turn enhances performance on the generalisation task.

With respect to our implicit indices of learning, the ‘guessing criterion’ showed clear evidence of implicit knowledge of all of the trained regularities in the direct, offline generalisation tasks. This finding is seemingly in contrast to those found by Brown et al. (2022). Here, the authors used a similarly complex language to consider grammatical generalisation, which incorporated real nouns which were known to the participants and provided the semantic cues to two grammatical categories (vehicles and animals). Artificial words were then used to create distributional cues to these two categories in the form of a particle which followed the noun. Brown et al. (2022) found that explicit grammatical knowledge, measured via verbal reports, was the driving force behind successful grammatical generalisation in adults, suggesting implicit knowledge was likely not to play a role here.

These results are, however, likely due to methodological differences. Firstly, Brown et al. (2022) use real nouns and as the authors discuss, this allows for the participant’s prior language knowledge to support learning. As discussed in Chapter 2 of this thesis, in sections 2.4.1 and 2.4.2.3, this access to prior knowledge in the construction of the semantic categories, may better support the emergence of explicit grammatical knowledge and so may potentially be contributing to this finding. Furthermore, the authors did not directly test for the presence of implicit knowledge. As discussed by Brown et al. (2022) their tests are ‘direct’ in nature in that they are more sensitive to explicit learning processes, but implicit knowledge can also be used when responding to these tasks (Franco et al., 2011; Jacoby, 1991) as supported by the current study’s results. As such, implicit knowledge may have also been present in Brown et al.’s (2022) participants, as was found in the current study. This is particularly pertinent as multi-componential theories of statistical learning suggest that both explicit and implicit knowledge can develop and may potentially interact (Batterink et al., 2019; Conway, 2020). Results from our indirect measure of generalisation could speak to this potential interaction. Despite finding evidence of implicit knowledge use within our direct measure of generalisation, evidence of implicit knowledge was not detected in the indirect, online phoneme detection task. As discussed earlier, it may be that the focus of this task on the suffix, which appears to be more difficult for participants to learn, may have contributed to this finding. Relatedly, it may be that learning of this regularity was not sufficiently robust to support its use in real-time.

Given Brown et al.’s (2022) findings and having found evidence of both explicit and implicit contributions to the learning of novel grammatical regularities, it is important to consider how these two processes may interact. In their review, R. Frost et al. (2019) propose a multi-componential view of statistical learning, a perspective also adopted by other recent reviews considering the neurobiological underpinnings of statistical learning (e.g., Batterink et al., 2019; Conway, 2020). Conway (2020) specifically suggests an interacting dual-system including both implicit and explicit learning mechanisms. The implicit learning mechanisms are based in the lower perceptual regions of the brain and are proposed to reflect the learning of 'simple', time-constrained and domain-specific regularities. However, these are underpinned by domain-general principles of neural plasticity that improves processing through learning. The explicit learning mechanisms reflect downstream brain systems (such as the prefrontal cortex) and are proposed to reflect the learning of more complex and global regularities which require learning over longer time periods. The two systems can work in parallel but can also work in collaboration or in competition with one another (Batterink et al., 2019; Conway, 2020).

In finding evidence for both explicit and implicit contributions to the learning of grammatical regularities, our current study supports this multi-componential view of statistical learning. This is particularly in light of the multiple regularities present within our paradigm, which cross modalities and domains (visual and auditory, phonological and semantic) and thus involve the type of complexity that more closely approximates natural language. This complexity, seen through the lens of the multi-componential view, may help to explain the lack of generalisation within our 'indirect' phoneme detection task, since cross-modal/domain regularities are proposed to require a longer time period for learning (Conway, 2020). Brown et al. (2022) did train their participants over 4 consecutive days, a longer learning period than used in this current study. However, the authors do also discuss the potential that this still is not long enough for implicit knowledge to emerge in regards their findings, particularly as they employ learning measures more sensitive to explicit knowledge use. The availability of prior language knowledge, as discussed earlier, may also be a contributing factor here. As such, this theoretical model and findings from Brown et al. (2022) supports the possibility that the exposure and learning time may not have been enough to build sufficiently robust implicit knowledge representations for use in real-time processing.

The successful generalisation found in the 'direct' generalisation tests could also be interpreted as support for parallel use of explicit and implicit systems, since there is evidence for use of both systems in this task. The interaction between them could be collaborative, or competitive. If implicit knowledge is gained but is not strong enough for real-time use, then it is possible that the explicit system supports its use in our direct tasks. The involvement of the explicit system could be aiding the use of implicit knowledge, enabling generalisation to occur in a shorter time period when knowledge can be reflected on rather than needed for real-time processing. Alternatively, it could be hindering the development of implicit knowledge either globally within the artificial language or for specific regularities. The lack of successful 'indirect' generalisation in the phoneme detection task could still be the result of no or relatively weak implicit knowledge representations, but due to the explicit system impeding its development rather than lack of sufficient learning time.

4.4.1 Future Directions

By using a more complex and naturalistic artificial language to explore the role of explicit and implicit knowledge on grammatical generalisation, this study has helped to highlight that both these learning systems may be involved in statistical learning and grammatical generalisation. Future work could build on these initial findings. For instance, while confidence ratings measures are a more sensitive measure of explicit and implicit knowledge than verbal reports, they still do not fully escape the issues of response bias (Barrett et al., 2013; Dienes, 2007; Norman & Price, 2015; Rebuschat, 2013). Thus, further research using more sensitive measures would help to confirm current findings and explore them in more detail. ERP (as used by Batterink et al., 2015) and eye tracking measures could be utilised for this.

When considering the role of explicit and implicit knowledge in statistical learning, both word-boundary and grammatical paradigms have so far only considered adults. Statistical learning has traditionally been seen as invariant across age (Raviv & Arnon, 2017; Saffran et al., 1997). However, how statistical learning is utilised across development may change, as may the roles of explicit and implicit knowledge in language learning in general, and grammatical learning in particular. Similarly, the neural underpinnings of statistical learning put forward by Conway (2020) are likely to go through different development trajectories and thus become available at different stages, as has been suggested from infant statistical learning research (Gómez, 2017).

In Chapter 2 children (9-10-year olds), did not show any retrospective, verbally reported explicit knowledge contribution to grammatical generalisation. Adults across all experiments in Chapter 2 did show a partial contribution here, even when they demonstrated the same lower level of training word knowledge and generalisation performance as the children. This does suggest that the role of explicit and implicit knowledge in statistical learning and for grammatical generalisation may be different for children; it would therefore be valuable for future research to investigate the developmental trajectory of explicit and implicit knowledge use within statistical learning.

To conclude, this study aimed to explore the role of explicit and implicit knowledge within grammatical generalisation using a statistical learning framework. The findings support a growing literature that suggests adults use both explicit and implicit knowledge within statistical learning tasks and has extended findings into a grammatical generalisation paradigm. This has helped to contribute to literature that considers statistical learning in more complex and naturalistic stimuli, as well as providing empirical evidence in support of recent theoretical proposals of multi-componential models of statistical learning (Batterink et al., 2019; Conway, 2020; Frost et al., 2019).

Chapter 5. General Discussion

The overarching aim of this thesis was to contribute to the broad question of how grammatical generalisation occurs within language. This question has been extensively considered from a nativist perspective (e.g., Chomsky, 1965), proposing that grammar knowledge is innate and it is from this knowledge base that grammar is generalised. However, as discussed in Chapter 1, the connectionist and constructivist (e.g., Bates & Goodman, 1997; Elman et al., 1998; Goldberg, 2005, 2009; MacDonald et al., 1994; Seidenberg & MacDonald, 2018), and more recent statistical learning (e.g., Saffran, Aslin, et al., 1996) literatures have questioned the innate nature of grammatical knowledge posed by nativist theories. Instead these literatures suggest that grammatical knowledge is learnt from the environment through statistical computations and/or ways of chunking item knowledge for processing (e.g., Christiansen & Chater, 2015; Isbilen et al., 2020). Grammatical generalisation then occurs once this knowledge base has been acquired. Constructivist based theories have explored how these chunking processes or statistical computations are generalised (e.g., Goldberg, 2005, 2009; Wonnacott et al., 2012) but literature exploring this within statistical learning is limited.

Thus, to help explore the '*how*' of grammatical generalisation, this thesis used a statistical learning framework, and specifically a grammatical category paradigm which incorporated multiple regularities within an artificial language that mimicked grammatical gender (e.g., Mirkovic et al., 2021; Mirkovic & Gaskell, 2016). This type of grammatical category paradigm provided a well-documented example of the use of multiple regularities to build language knowledge (e.g., Gómez, 2002; Lany, 2014; Lany & Saffran, 2010, 2011; Mintz, 2003; Monaghan et al., 2005). It also enabled the use of a more complex artificial language, through incorporating multiple regularities from different domains (e.g. phonological, distributional, and semantic) and allowing for the training and testing of both vocabulary and grammatical knowledge. There are only a small number of studies that use this more complex, and more naturalistic, design (e.g., Lany, 2014; Lany & Saffran, 2010, 2011; Monaghan et al., 2019), a limitation in the statistical learning literature that has been commented upon (R. Frost et al., 2019). This current lack of complexity within the statistical learning literature limits the generalisation of findings to natural language processing. As such, by using a grammatical category paradigm this thesis is helping to address this limitation and provide a language learning context that more closely links to natural language learning.

Using this paradigm, this thesis considered the 'how' of grammatical generalisation by aiming to explore the role of vocabulary, and the role of explicit and implicit knowledge of the grammatical regularities. As introduced in Chapter 1, both of these aims were considered from the perspective of connectionist theories of language processing, which were operationalised through the 'chunk and pass' model (Christiansen & Chater, 2015). This model proposes that grammatical knowledge emerges through the building of synaptic connections to create networks (representations) that support successful language behaviour. These representations are formed through interactions between innate learning mechanisms (e.g., statistical learning), language experience and prior language knowledge (Elman et al., 1998).

Bates & Goodman (1997) and Goldberg (2005, 2009) use this underlying principle to propose that different aspects of language are not stored and processed separately, rather they are stored and processed together and impact one another. Connections will be formed between all areas of language, in whatever way best supports language behaviour. Bates & Goodman (1997) and Goldberg's (2005, 2009) theories specifically consider the impact of vocabulary and grammar on one another, termed item knowledge and generalisation by Goldberg (2005, 2009). From these theories it was hypothesised that within the statistical learning framework used in this thesis, vocabulary would have an impact on grammatical generalisation performance.

Linked to these theories but specifically highlighted by the 'chunk and pass' theory (Christiansen & Chater, 2015) used to operationalise them in Chapter 1, is the assumption of implicitness underlying constructivist theories of language and in particular the statistical learning literature (Batterink et al., 2015; Perruchet & Pacton, 2006; P. J. Reber et al., 2019). This assumption implies that learning grammatical regularities within this context contributes towards the building of a language processing system. They contribute to these processing systems by providing the methods of processing that support successful language behaviour. In terms of the 'chunk and pass' theory, the method of processing is in how language is increasingly abstracted and chunked to allow for in-the-moment language processing that works within the limits of memory for real-time language use (Christiansen & Chater, 2015). Thus, grammatical regularities are not represented in and of themselves (or in other words, explicitly) within this system, they are implicitly represented through the connections made with items or chunks of the language.

This assumption of implicitness has been characteristic of the statistical learning literature until fairly recently. Similarities between the statistical learning literature and the implicit learning literature (Batterink et al., 2015; Monaghan et al., 2019; Perruchet & Pacton, 2006; P. J. Reber et al., 2019) have however questioned this assumption. Within the implicit learning literature, the implicit nature of what has been learnt has been directly examined for a large part of this literature's history (for reviews see Batterink et al., 2015; A. S. Reber, 1989; P. J. Reber et al., 2019) with evidence of explicit (conscious) knowledge impacting performance within implicit learning studies being found (Kelley & Jacoby, 2000; A. S. Reber, 1989; P. J. Reber et al., 2019). Recent research within statistical learning frameworks that have been prompted by this parallel have also found a role for both implicit and explicit knowledge in adults, both within a traditional word-boundary paradigm (e.g., Batterink et al., 2015; Franco et al., 2011) and a grammatical learning paradigm (Monaghan et al., 2019). Further to this, a recent review of the neural correlates for both implicit learning and statistical learning tasks demonstrate similar neural systems are being used (Batterink et al., 2019). Recent reviews of statistical learning have incorporated implicit learning research and proposed a role for both implicit and explicit processes (Conway, 2020; R. Frost et al., 2019). This suggests that within adults at least, explicit as well as implicit grammatical knowledge may influence grammatical generalisation performance, a hypothesis that was considered in this thesis.

The following will address the findings from this thesis in regards to these two aims: the role of vocabulary and the role of explicit and implicit grammatical knowledge in grammatical generalisation. While this thesis had originally aimed to consider these questions in relation to children as well as adults, the recent pandemic prevented much of the planned child-based research. As such, the findings regarding adults will be addressed first for both questions regarding vocabulary and explicit/implicit knowledge, before considering the limited child findings from Chapter 2. Secondary findings and further implications will then be discussed.

5.1 The role of vocabulary - is it all about the 'word'?

Chapter 2 details findings that demonstrate how vocabulary within the learner's lexicon, but not within the learning context can contribute to adult grammatical generalisation. Experiment 1 in this chapter shows how increased vocabulary variability within the learning context did not affect grammatical generalisation. While this was an unexpected result given previous findings of a variability effect within grammatical category learning in statistical learning paradigms (e.g. Gomez, 2002) it does support more recent findings by Brown et al. (2022). This more recent study uses a more complex language compared to Gomez (2002), similar to the artificial language used throughout this thesis, albeit semi-artificial rather than fully artificial. Taken together with the findings from Experiment 1 of Chapter 2, it suggests that the benefits gained from higher variability within the learning context become less accessible the more complex the stimuli are. Experiment 2 of Chapter 2 demonstrated that the level of vocabulary knowledge within the lexicon does however contribute to adult grammatical generalisation. When vocabulary knowledge of the training language was reduced in adults their grammatical generalisation performance was reduced as well.

Despite the lack of a vocabulary variability affect, these findings support the hypothesis that vocabulary knowledge impacts grammatical generalisation, specifically that vocabulary knowledge within the lexicon plays a supportive role here. As this hypothesis was based on lexicalist theories, these results also provide evidence for the integration of both vocabulary and grammatical knowledge. More specifically, these findings support the suggestion that grammatical knowledge cannot emerge until a critical mass of vocabulary has been reached (Bates & Goodman, 1997). According to this and other linguistic theories (e.g. Goldberg, 2005, 2009), once this critical mass of vocabulary knowledge has been reached this provides the needed amount of experience, use and variability to allow for similarities to become more salient, enabling the abstraction and generalisation of grammatical regularities.

If future research continues to support a role for vocabulary within grammatical generalisation, then how exactly is vocabulary supporting grammar? According to implicit connectionist/constructivist theories (e.g., Christiansen & Chater, 2015; Goldberg, 2005, 2009), it would be supportive by providing the item knowledge needed to allow for similarities (grammatical regularities) to become salient for abstraction and generalisation. When this idea is operationalised within the 'Chunk and Pass' theory (Christiansen & Chater, 2015), similarities are shown through 'well-worn' connections with multiple items within a network, connections which are implicit within a language system. As discussed earlier in terms of critical mass and variability, vocabulary here would be supporting the learning of implicit grammatical knowledge that is supporting processing within a language system.

A secondary analysis conducted in Chapter 2, considered the role of verbally reported explicit grammatical knowledge that emerged in the course of the study. This secondary analysis found that when vocabulary knowledge was reduced in adults, so was explicit grammatical knowledge. This finding coupled with the found contribution of explicit knowledge to grammatical generalisation performance could suggest another way in which vocabulary supports grammatical generalisation in adults. It could be supporting the emergence of explicit grammatical knowledge which in turn supports grammatical generalisation. Whether vocabulary is coupled with this or not, the contribution of verbally reported explicit knowledge in adults, even when vocabulary knowledge is low support similar previous findings in adults within statistical learning (e.g., Batterink et al., 2015; Brown et al., 2022; Franco et al., 2011; Monaghan et al., 2019) as well as recent multi-componential theoretical proposals (e.g., Batterink et al., 2019; Conway, 2020; Frost et al., 2019). However, the use of retrospective verbal reports to measure explicit grammatical knowledge is limited, particularly in terms of response bias and lacking sensitivity (Merikle & Reingold, 1992; Moroshkina et al., 2019; Newell & Shanks, 2014; Tunney & Shanks, 2003; Wierzchoń et al., 2012). An issue that was discussed in detail in Chapter 3. Thus, more sensitive measures of explicit grammatical knowledge are needed to confirm the findings here.

5.2 Explicit and implicit grammatical knowledge

To start to address these issues the second aim of this thesis focused on the role of explicit and implicit grammatical knowledge on adult grammatical generalisation. To do this, Chapter 3 detailed the development of a more sensitive and within-task measure of explicit knowledge, which also incorporated a measure of implicit knowledge use: confidence ratings. Confidence ratings have been developed and used within the implicit learning, memory and perception awareness literatures (Dienes et al., 1995; Dienes, 2007; Dienes & Berry, 1997; Dienes & Perner, 1999; Moroshkina et al., 2019) and provide a behavioural measure of the subjective threshold: the point at which a person becomes aware of the knowledge that is driving their own behaviour

(Cheesman & Merikle, 1984; Dienes et al., 1995; Dienes, 2007; Dienes & Berry, 1997; Dienes & Scott, 2005; Moroshkina et al., 2019).

Confidence ratings are also better able than verbal reports to meet the criteria for a good subjective awareness measure (e.g. Newell & Shanks, 2014): relevancy, sensitivity to different levels of consciousness, reliability, and immediacy. In addition, we were careful to try to reduce reactivity, or the potential for the confidence rating measure itself to influence the emergence of explicit knowledge (Moroshkina et al., 2019). This resulted in a confidence rating measure being integrated into the existing word-picture matching task, that was used during both the training and testing phases of the grammatical learning paradigm. The new measure was thus better able to index explicit and implicit knowledge than the verbal reports we had used previously, while requiring only minimal changes to our grammatical learning paradigm.

This measure was used across the two experiments reported on in Chapter 4. In both experiments a zero-correlation analysis was used to examine whether participants' increased confidence in their own responses, was associated with greater accuracy on the grammatical generalisation tests. (Chan, 1992, but see also Dienes, 2007; Dienes et al., 1995; Dienes & Berry, 1997; Moroshkina et al., 2019; Norman & Price, 2015). Across both experiments in Chapter 4, this analysis confirmed the use of explicit knowledge during generalisation, specifically for the generalisation task that tested for both the determiner and suffix mapping with the semantic referent (Determiner and Suffix task). Surprisingly, zero-correlation analysis did not show evidence of the presence of explicit knowledge for the other generalisation tasks, even when this had been suggested by verbal reports. This may be because of methodological differences, in that verbal reports were considered in a regression-based individual differences analysis, while the zero-correlation reflects group-level performance. Alternatively, this discrepancy could reflect the retrospective nature of verbal reports, where participants are able to consider their experience as a whole when assessing their awareness of grammatical regularities. Nonetheless, taken together, the results of the zero-correlation analysis and the verbal reports, support and extend the recent literature reporting the presence of explicit knowledge within statistical learning tasks (Batterink et al., 2015; Brown et al., 2022; Franco et al., 2011; Monaghan et al., 2019).

Although it is important to recognise the presence and potential contribution of explicit knowledge to grammatical generalisation and to statistical learning tasks more broadly, it cannot fully account for learning. Most pertinently, while performance on the Suffix Only task showed evidence of successful generalisation, it did not consistently seem to be associated with explicit knowledge. Experiment 2 in Chapter 4 aimed to directly assess the presence of implicit knowledge, rather than inferring it from the lack of explicit knowledge. The six-point version of the confidence rating scale allowed for a 'guessing criterion' analysis, in which above-chance performance on trials where the participant says they are guessing, is taken to indicate implicit knowledge use (Cheesman & Merikle, 1984; Dienes et al., 1995). Using this analysis within Experiment 2 found evidence of implicit knowledge use across all generalisation tasks. Taken together with findings regarding explicit knowledge use, this demonstrates the presence and use of both learning systems within grammatical generalisation, supporting and extending findings from previous word boundary (Batterink et al., 2015; Franco et al., 2011) and grammatical learning (Monaghan et al., 2019), statistical learning studies.

It also suggests, that while including semantic cues in our artificial language may bias participants towards the use of explicit knowledge (e.g., Brown et al., 2022) implicit knowledge still seems to be developing and supporting responses when knowledge is being tested. It may also suggest, that when participants are less able to access and use prior knowledge to support learning, implicit learning processes may be utilised more. Brown et al. (2022) used a similarly complex language to the one used in the current thesis and while they did not directly test for the presence of implicit knowledge, relying on verbal reports, they demonstrate evidence that their successful generalisation results are driven by explicit knowledge in participants. The authors acknowledge that their use of a semi-artificial language (using real English nouns, known the English-speaking participants, with artificial grammatical particle words) means participants can access prior language knowledge to support learning. The artificial language used across this thesis takes the next step, in that artificial noun words were used but with known referents and semantic concepts, so some prior knowledge can be used to support learning but to a lesser extent than when real words are also used. This potentially could have meant less reliance on explicit knowledge and a larger draw on implicit knowledge in this current paradigm compared to Brown et al. (2022). It could be that explicit learning processes are drawn upon more when prior knowledge is more accessible and available to support learning. With the flip side being that implicit learning process are drawn upon more when prior knowledge cannot be drawn upon or is less available to support learning. This idea will be revisited later in this chapter (section 5.4.1) but does need to be directly tested, where access to prior knowledge is directly manipulated and compared along with the use of more sensitive measures for both implicit and explicit knowledge.

While confidence ratings provide a more sensitive measure of explicit and implicit knowledge, it still has notable limitations. Firstly, it is still subject to response bias. This is an issue that has been extensively discussed and considered within the perception detection literature, where a ‘d-prime’ based analysis has been developed to address this (meta-d; Barrett et al., 2013; Evans & Azzopardi, 2007; Maniscalco & Lau, 2012; Norman & Price, 2015). However, the smaller number of trials used in artificial language paradigms compared to the larger number of trials (100s to 1000s) that are more standard within the perception detection paradigm, mean this method of analysis was unsuitable for our studies. As such, the recommendations made by Rebuschat (2013) for the second-language literature were chosen for this thesis due to the closer similarities to the grammatical learning paradigm used here. Research methods that avoid response bias altogether could address this issue, including measures such as eye-tracking, ERP (as used by Batterink et al., 2015) and measure which incorporate reaction times.

5.2.1 Exploring new ways to measure implicit grammatical knowledge

An eye-tracking or ERP approach was originally planned for Experiment 2 of Chapter 4, but the COVID-19 pandemic prevented the use of these measures. Instead, a reaction time task was developed, based on the 'speeded target detection task' used in Batterink et al. (2015). Batterink et al. (2015) developed this task as a better method for capturing implicit knowledge use, drawing from the second language learning literature and the distinction between direct and indirect tasks. Indirect tasks measure the use of knowledge in real time, and because implicit knowledge is assumed to be the primary driver of processing in faster time-frames, tasks that require this type of response are posited to be better measures of implicit knowledge (Kelley & Lindsay, 1996; Rebuschat, 2013; Timmermans & Cleeremans, 2015). Direct tasks directly assess the knowledge of interest, requiring the need for a participant to review their knowledge before responding and operating over a longer time-frame; as such, direct tasks are suggested to be biased towards explicit knowledge use, although it is recognised that implicit knowledge is also likely to contribute to performance on these tasks (Kelley & Lindsay, 1996; Rebuschat, 2013; Timmermans & Cleeremans, 2015).

Batterink et al. (2015) used both a direct and indirect style tasks, with a traditional statistical learning two-alternative forced choice task as the direct task and the 'speeded target detection' task as the indirect task. While evidence of successful word-boundary learning was found within both tasks, it was only the indirect, 'speeded target detection' task that correlated with ERP measures of predictable target detection within participants (P300 effect). This correlation supports the proposition that tasks of this kind are a better reflection of implicit knowledge use. This finding can also be seen in studies of natural language learning, where native language users are better able to utilise implicit knowledge of grammatical gender than second language users, in reaction-time based tasks (Lew-Williams & Fernald, 2010).

The word-picture matching generalisation tasks used in the experiments throughout this thesis could be considered a 'direct' task, as they require participants to reflect on their knowledge in order to respond. Thus, with the lack of opportunity to use ERP or eye-tracking methods, the findings from Batterink et al. (2015) and Lew-Williams & Fernald (2010) influenced the development of a new reaction-time based task, as a more sensitive measure of implicit knowledge use. This task took the form of a phoneme detection task, in which participants' responses to the phoneme characterising the suffix (/oo/ or /ee/), were influenced by how well this could be predicted by the preceding cue (the determiner, and the picture indicating the semantic category).

Reaction times for correct phoneme detections were faster for trained words than grammatically consistent untrained words, demonstrating successful word learning (and providing evidence that the task was sensitive to learning *per se*). However, reaction times did not significantly differ for grammatically consistent versus inconsistent untrained words, suggesting a lack of evidence for grammatical generalisation. Taken at surface value, this would suggest the absence of implicit knowledge of grammatical regularities. Yet this is at odds with the confidence rating findings from the word-picture matching generalisation tasks. It may be that the phoneme detection task was too complex in terms of task design (four different types of inconsistent stimuli), or in terms of the multiple cues in the language itself. Alternatively, the focus of this task on the suffix, which we know is a relatively difficult aspect of the language to learn, may have been problematic. Relatedly, it may be that implicit knowledge of the suffix was present, but the representation was not sufficiently robust to be accessible under time-pressure in a real-time, online task. If that is the case, the longer time-frame of the offline word-picture matching generalisation tasks may have allowed even weak (implicit) representations of the suffix to be activated. The lack of successful generalisation found within this reaction time measure, may also fit with the proposed idea that when semantic cues are incorporated into an artificial language, it biases participants towards the use of explicit learning processes (e.g., Brown et al., 2022). As discussed in the previous section (5.2), this may be linked to the availability of prior knowledge for supporting learning. Research using a similar paradigm and reaction time measure, but where both the words and semantic information is novel and unknown to participants would be of interest in considering this hypothesis. This line of enquiring along with the suggestions proposed in section 5.2, could aid in exploring how and when explicit and implicit learning processes are used within language acquisition and learning.

While acknowledging these limitations, overall the results from Chapter 4 support and extend growing evidence of the use of both implicit and explicit knowledge within statistical learning (Batterink et al., 2015; Brown et al., 2022; Franco et al., 2011; Monaghan et al., 2019). More broadly, they contribute to the empirical evidence base supporting recent theoretical proposals of a multi-componential view of statistical learning (Batterink et al., 2019; Conway, 2020; Frost et al., 2019).

5.3 Grammatical generalisation in children - What can we tell so far?

Whilst children were only tested in one experiment of Chapter 2, the results from this study can provide some insights regarding the two main aims of this thesis. The first finding of note to consider is that although 10-year old children showed a good level of word-learning, they did not show evidence of generalisation for the grammatical regularities within the training language. As discussed in Chapter 2, this finding appears to be counter to previous research within statistical learning, both using word-boundary and grammatical paradigms where children of a similar age and younger demonstrate successful grammatical learning and generalisation (Brown et al., 2022; Gómez, 2002; Hall et al., 2018; Lany, 2014; Lany & Saffran, 2010, 2011; Wonnacott et al., 2012). One key difference is the nature of the language being learnt. The majority of previous studies use simpler artificial languages that incorporate only phonological and/or distributional cues (e.g. Gómez, 2002; Hall et al., 2018). In the cases where more complex languages are used, which include semantic referents, these are trained after the phonological aspects of the language (e.g. Lany, 2014; Lany & Saffran, 2010, 2011). Presenting all grammatical regularities simultaneously as the current paradigm does, may have been too complex a task for children, particularly in the limited time-frame used for training.

As discussed, Brown et al. (2022) uses a similarly complex language which is presented simultaneously to child participants, who demonstrated successful generalisation within the deterministic version of their language. However, the use of real English nouns, which were known to the child participants, means that children could draw upon prior language knowledge to support generalisation here. Known semantic referents were used across this thesis, presenting prior knowledge that children could potentially use to support generalisation. However, this may not have been enough when presented with novel noun words, supporting the argument that this higher level of complexity may have prevent finding evidence for successful grammatical generalisation. This would be an interesting line of enquiry to pursue as natural language acquisition and learning does require children to learn and generalise multiple, simultaneously presented regularities. In the case of language acquisition this happens when little to no previous language or other knowledge can be drawn upon to support learning here. However, this typically occurs immersively and over a much longer time-frame. Brown et al. (2022) uses a paradigm which incorporates learning over four sessions which occurred over four consecutive days, potentially contributing to their findings here. Thus, more exposure to the regularities, potentially over multiple days to allow for consolidation opportunities, may have enabled children to generalise within our paradigm.

The specific idea investigated in Chapter 2, however, was that grammatical learning and generalisation builds on a ‘critical mass’ of vocabulary, in line with lexicalist theories (e.g. Bates & Goodman, 1997). While the children in Chapter 2 showed good levels of word-learning, this was still significantly lower than that of adults who had received the same amount of training (as was also found in Mirkovic et al., 2021), and this level of vocabulary knowledge may have been insufficient for evidence of grammatical generalisation to have been found. This hypothesis was supported by the further findings reported in Chapter 2, with adult participants showing a similar lower levels of grammatical generalisation performance when their vocabulary levels were reduced to child levels. However, more direct evidence from child participants is ideally needed to confirm this hypothesis, with the clear prediction that increasing vocabulary knowledge in children should also improve grammatical generalisation. If that were found to be the case, it would provide further support for the lexicalist perspective, suggesting that there are similar requirements of a critical mass across development, and that one driver of developmental differences is the capacity for learning at the item-level, rather than at the level of grammar.

In addition to finding developmental differences in word-learning, we also found a striking difference between adults and children with respect to explicit knowledge of the grammatical regularities, showing mean scores of lower than 1 for both determiners and suffixes (see Table 15). Although as a group the 10-year old children in Chapter 2 showed a lack of evidence for generalisation, there was substantial variability within the group (as there was in adults), with some children who did show evidence of generalisation. However, the regression analysis showed that at the individual level, even when children did generalise, there was no evidence of an explicit knowledge contribution. This was in contrast to the adults, for whom there was consistently evidence of a contribution of explicit knowledge, even when as a group vocabulary and grammatical generalisation levels had been reduced to a similar level to that of children. This suggests potential developmental differences in the learning systems utilised in statistical learning.

With the notable exception of Brown et al. (2022), the role of explicit knowledge within statistical learning for children has so far not been considered in the statistical learning literature. In contrast to the current findings, Brown et al. (2022) found that explicit knowledge was the driving force behind successful grammatical generalisation, even in their child participants. However, the discussed use of real nouns and the access to prior knowledge this provides could be influencing this result. Further supporting research, directly considering the role of prior knowledge in the use of explicit and implicit knowledge would be if interest here. Research from the related learning paradigms, Hebb-learning, where access to prior knowledge to support learning is limited, do support the current results. Smalle et al. (2018) used a Hebb-learning paradigm to examine word learning for an artificial language in adults and children. When this language was implicitly trained, that is when participants did not receive any information about the underlying regularities (similar to the statistical learning paradigm used in this thesis), explicit knowledge emerged in both groups. However, it emerged at an earlier point in adults and only contributed to adults - but not children's - word learning performance. In terms of implicit knowledge, children demonstrated better word learning performance compared to adults when trained implicitly compared to an explicit training protocol in which participants were given information about the underlying regularities.

The findings from Chapter 2, together with those reported by Smalle et al. (2018), warrant further research considering the role of both implicit and explicit knowledge in children. This could make use of the confidence rating measure from Chapter 4, potentially combined with ERP and/or eye-tracking methods. Again, this had been planned as a study before the pandemic. If future research of this kind confirms the current findings it could help to inform theories of statistical learning and grammatical generalisation. As discussed with regard to the adult data, it may be that implicit knowledge takes longer to fully develop, particularly when the stimuli involved are more complex. If children are more reliant on this than explicit knowledge, or are less able to utilise explicit knowledge, this could help to explain the lack of generalisation found within this paradigm for children.

5.4 Not all regularities are learnt equally

An advantage of the paradigm used throughout this thesis is that its increased complexity better reflects naturalistic language processing than traditional statistical learning tasks, something that has been lacking from much statistical learning research (R. Frost et al., 2019). Using a paradigm which incorporates multiple cues - phonological, distributional and semantic - also raises interesting questions about how these cues interact during learning. Empirically, the studies reported in this thesis, in conjunction with those reported by Mirkovic et al. (2021), have demonstrated that some regularities are prioritised over others: The Determiner and Suffix task, which indexes the mapping of the 'determiner and suffix' co-occurrence to the semantic cue, consistently shows the most robust generalisation. This is followed by the determiner-only mapping to semantics, and then the suffix-only mapping to semantics. The weakest performance is always for the Phonological Form task, which indexes learning of the co-occurrence between the determiner and suffix, without reference to the semantic cues. That is, regularities which incorporate semantics are consistently learnt better than those that do not. Within the regularities that do incorporate semantics, the determiner seems to be prioritised over the suffix.

Recent research by Vujović et al. (2021) using a similar paradigm has started to consider this interaction between cues and vocabulary and its effects on learning. This study used an artificial language which incorporated both vocabulary and grammatical information through the use of phonological, distributional and semantic cues to denote one of two grammatical categories. This was achieved through the use of a stem including one of two vowel sounds providing the phonological cue, which was then uniquely paired with a visual referent. The visual referent provided the semantic cue along with this pairing providing vocabulary information. Stems were then either paired with a prefix or suffix to provide the distributional cue. Using this artificial language, the authors considered the interaction of not just the different cues but also the cues with different aspects of vocabulary on learning. A finding of interest to this thesis, is that when vocabulary type frequency was high (high variability), this supported the generalisation of prefixes but not suffixes. However, this effect was only found when the semantic cue was present during testing.

This result could help to explain why in the highly deterministic artificial language used in this thesis, which incorporated a high vocabulary type frequency (variability), seems to find that generalisation of the determiner and semantic mapping to be one of the most robust learning effects. The lack of a type frequency (variability) effect found by Vujovic et al. (2021) when semantics are absent reflects the overall lack of generalisation found in this context. This is a similar finding to that reported in Chapter 4 of this thesis, and in Mirkovic et al. (2021), and provides further support for the conclusion that when regularities include semantics they are learnt better. Findings from Vujovic et al. (2021) and this thesis supports proposals that cue type and context can interact with grammatical generalisation, as well as with vocabulary context, in line with lexicalist theories (Bates & Goodman, 1997). Further investigation would be warranted here to see exactly where and how this interaction influences grammatical generalisation.

5.4.1 A further interaction with explicit and implicit knowledge?

Further to this, it is worth considering the possibility that the learnability of regularities could interact with the explicitness or implicitness of grammatical knowledge. In the experiments reported in this thesis, and in Mirkovic et al. (2021) explicit awareness is most reliably found for the task which shows the best levels of generalisation performance (the determiner & suffix test), suggesting a positive relationship between explicit knowledge and generalisation performance. On the other hand, in the study by Vujović et al. (2021), performance - and particularly the effect of vocabulary type frequency on prefix generalisation - did not appear to be related to explicit knowledge: although nearly half of participants reported knowledge of the language rules by the end of the experiment, the pattern of results did not change when these participants were excluded from the analysis.

One possible explanation for this difference relates to the distinction between first language acquisition and second-language learning. The paradigm used in this thesis could be argued to be more akin to second language learning, as the semantic referents used are animals and objects that participants already know and have words for in their native language. Vujović et al. (2021) use of frebbles for referents on the other hand, uses novel semantic referents that are previously unknown to participants. While Vujović et al. (2021) tested adult participants who bring fully developed systems and prior language knowledge to the task, their design in terms of novel labels for novel referents, is more akin to first language acquisition. It might be that the role of explicit knowledge is more pronounced within a second language learning context, where the underlying semantics are already known to the participants. When linked to discussions in regards to Brown et al. (2022), this idea also supports hypotheses regarding the role of prior knowledge.

Starting with Brown et al. (2022), moving next to the paradigm used in this thesis and then to Vujović et al. (2021) it could be argued that this represents a continuum of prior knowledge availability. Brown et al.'s (2022) semi-artificial language could represent the highest level of prior knowledge availability where participants can use the real nouns as an anchor to support learning. The paradigm used in this thesis reduces the availability of prior knowledge through the use of novel artificial nouns but referents and semantic categories that are known to participants. Vujović et al. (2021) then has the least prior knowledge availability through the use of both novel words, referents and semantic categories. The found lack of explicit knowledge influence on grammatical generalisation in Vujović et al. (2021), supports the proposed hypothesis that higher prior knowledge availability may bias towards explicit learning processes and lower prior knowledge availability biases towards implicit learning processes (see section 5.2).

This is currently a speculative idea and so some caution is needed when making these interpretations. For instance, Vujović et al. (2021) only used verbal reports to measure explicit knowledge and did not directly test implicit knowledge, meaning direct research considering this idea using more sensitive measures of both explicit and implicit knoweldge is needed. However, it does strongly support future investigation of the potential interaction between explicit and implicit knowledge, the learnability of regularities and the role of prior knowledge availability with this. Investigations of this nature could help to reveal potential differences in the mechanisms underlying first and second language acquisition, may also have real-world implications for second language teaching and learning as well as helping to better understand the theoretical underpinnings of statistical learning. It would also be interesting to consider whether and how these processes change across the course of development, and the implications for children's learning of not only their first, but also additional languages at different ages.

5.5 Limitations and Implications

While this thesis presents some novel findings, particularly within the field of statistical learning, there are limitations which need to be considered. This section will summarise some important limitation for consideration from empirical chapters 2 and 4 before moving onto discussing some overall limitations from the thesis.

5.5.1 Chapter 2: Exposure frequency and children verbalising explicit knowledge

The experiments presented in Chapter 2 aimed to consider the role of variability during exposure and the role of vocabulary knowledge within the lexicon. A key limitation within these experiments is that frequency of exposure was not controlled for and as such could be a confounding variable at play within the results. These three aspects of vocabulary: exposure variability, exposure frequency and knowledge, are interrelated factors which are potentially hard to disentangle from one another. For instance, vocabulary knowledge cannot be obtained without some kind of exposure to the vocabulary. Despite the interrelatedness of these factors, teasing apart their roles within grammatical generalisation would help better understand the mechanisms at play, particularly in terms of the lexicalist theory proposed by Bates and Goodman (1997). Future studies looking at exposure variability can control exposure frequency by repeating stimuli in low variability conditions to equal the exposure frequency in highly variable condition, as implemented by Gomez (2002).

Controlling for exposure frequency is harder to do when considering vocabulary knowledge. As highlighted you need exposure for knowledge to develop and across Chapter 2 levels of knowledge was manipulated by also manipulating levels of exposure frequency. Manipulating levels of knowledge without changing exposure would not be possible, however an individual differences approach may help here. With a larger sample of participants, where levels of vocabulary knowledge would naturally vary even with exposure frequency kept constant, would allow for analysis which could look at the role of vocabulary without exposure frequency being a confound. Exploratory analysis within Chapter 2 considered these individual differences approach by combining adult data from across groups and experiments to look at the correlations between vocabulary knowledge and generalisation performance. This analysis showed a partial relationship between knowledge and generalisation which is hopeful, however it is exploratory and as such further studies are needed with larger samples for more robust results.

Exploratory analysis in Chapter 2 also showed an interesting, potential developmental difference in the use of explicit knowledge within grammatical generalisation. This is due to the finding that children showed lower levels of verbalised explicit knowledge compared to adults and did not show any evidence of an explicit knowledge contribution to grammatical generalisation performance unlike adults. This is even when adults demonstrated a similar level of vocabulary knowledge and generalisation performance. While an intriguing finding, it needs to be acknowledged that as well as being an exploratory finding, the limitations of verbal reports as a measure for explicit knowledge (as discussed in Chapter 3) limit what we can draw from this. Particularly in terms of being able to verbalise explicit knowledge, a limitation posed for adult participants (Dienes & Perner, 1999) which is likely to be more exacerbated in children. Thus, the conclusions that can be made in regards to the use of explicit grammatical knowledge for grammatical generalisation in children is limited and warrant further investigation with more sensitive measures, as discussed in section 5.3.

5.5.2 Chapter 4: The illusive subjective threshold

The main limitations within the experiments reported in Chapter 4 centre around the difficulties in detecting the subjective threshold. Both in terms of before it is reached (implicit knowledge) and when it is reached (explicit knowledge). Across Chapter 4, two measures were used to try and detect explicit knowledge, so when the subjective threshold had been reached: verbal reports and confidence ratings using the zero-correlation analysis. Whilst both measures showed evidence of explicit knowledge use within the word-picture matching generalisation tasks there were discrepancies in exactly where explicit knowledge use was found. As discussed in section 5.2, this discrepancy may be due to methodological differences and/or the retrospective nature of verbal reports in comparison to the within-task nature of confidence ratings. It highlights difficulties in considering when the subject threshold is reached as well as when this may play an active role in grammatical generalisation within this paradigm. The findings from this thesis cannot currently speak to exactly why there is a discrepancy here, an understand of which would help to better understand the mechanisms at play in terms of when the subject threshold is reached and informs behaviour.

In Experiment 2 of Chapter 4, first attempts to look at the emergences of explicit and implicit knowledge across training was implements using confidence ratings within the repetition and word-picture matching training task. While it showed a trend towards the emergence of explicit knowledge but no evidence of implicit knowledge use by the end of training, the findings here are extremely limited. Firstly, there was also only a very small number of items this analysis could be conducted on, namely the mismatched items that were mismatched by grammatical regularities, four items within each training task (see sections 4.3.1.2 and 4.3.1.4 for more details). Secondly, the confidence rating measure used in this experiment was specifically designed for generalisation tasks, where the items being test had not been encountered before (see Chapter 3). It cannot then distinguish between explicit item recollection knowledge or explicit grammatical regularity knowledge. The items in the training tasks contain, particularly by the end of training would obviously have been encountered before by participants. This means that the trend towards the emergence of explicit knowledge by the end of training could be driven by explicit item and not regularity knowledge. Additionally, implicit regularity knowledge may be present in participants, but their item level knowledge may negate the need to use this when responding. Thus, the confound issue of explicit item knowledge obscures this measures ability to detect whether the subjective threshold has or has not been reached.

In terms of detecting when the subjective threshold has not been reached during grammatical generalisation, two measures were employed to look at this within Experiment 2 of Chapter 4: Confidence ratings using the guessing criterion analysis and reaction times within the phoneme detection task. Discrepancies were found between these two measures in that evidence was found for implicit knowledge use using confidence ratings but not within the phoneme detection task when reaction times were analysed. This discrepancy currently limits the conclusions that can be made about the underlying mechanisms at play and highlights the difficulty in detecting implicit knowledge. As discussed in section 5.2, there are a number of potential reasons for this discrepancy with suggestions for further research to help overcome this limitation.

Overall these limitations highlight the difficulty in capturing the subjective threshold both when it has and has not been reached. While each behavioural measure used across this thesis has advantages they all still have limitations. For example, confidence ratings are more sensitive than verbal reports (Chapter 3), however they are still subject to response bias (see section 5.2 for more details). It highlights the need for the use of multiple measures, both behavioural as well as introducing neuroimaging measures. Efforts to better understand exactly what each measure is capturing would also better aid understanding in the role of implicit and explicit knowledge within grammatical generalisation.

5.5.3 General limitations

There are three main limitations that should be considered when interpreting the findings from across this thesis: when considering first language acquisition, the complexity of the grammatical regularities and the time course of learning. The methods and paradigm used throughout this thesis are more akin to second language learning in it's use of familiar referents within the artificial language. Participants, even the child participants, are thus able to bring prior linguistic knowledge to the task which will have an impact on how the grammatical regularities are learnt. Something infants are unable to do when acquiring their first language. While understanding second language learning and processing is still a needed endeavour, the findings are limited in what they can conclude regarding first language acquisition. Connected to this, while child participants are used in Experiment 1 of Chapter 2, they are markedly older than the typical infancy age for first language acquisition. This is pertinent as language processing does change and develop across childhood into adulthood (e.g. Newport, 1990). Future research could start to overcome this limitation through considering younger participants as well as through the use of unknown referents within the artificial language.

The artificial language used across this thesis is fairly novel for statistical learning research in its use of multiple regularities (although there are more examples of this within linguistic research, see section 1.1.3.4). The use of multiple regularities has helped to answer the call for more complexity within statistical learning (R. Frost et al., 2019) and consider grammatical generalisation within a more naturalistic context. However, the artificial language uses grammatical regularities which are deterministic in nature. While this may better reflect natural languages in terms of semantic regularities, distributional and phonological regularities are usually probabilistic (e.g. Braine, 1987). This obviously puts a limit on how naturalistic the artificial language is, which needs to be considered when generalising the results found here into the real-world. The next step towards a more naturalistic artificial language using this paradigm, would be to create probabilistic distribution and phonological regularities to better reflect the pattern found in natural languages.

Finally, throughout this thesis grammatical generalisation has been consider within one learning session. However, real-world language learning usually occurs through 'bathing' in language over a long period of time. A wealth of research has consistently shown a role for consolidation within vocabulary learning (e.g., Henderson et al., 2015; Henderson et al., 2012; Henderson & James, 2018; James et al., 2017)). Whilst findings are mixed when it comes to grammatical generalisation (e.g., Mirkovic & Gaskell, 2016; Tamminen et al., 2012; Tamminen et al., 2015), the proposed relationship between vocabulary and grammar (e.g. Bates & Goodman, 1997, Goldberg, 2005; 2009; Chapter 2 of this thesis) warrants efforts to investigate the time course grammatical regularity learning and generalisation. Particularly when considering the role of explicit and implicit knowledge, aspects of grammatical knowledge which have had little considered when examining grammatical generalisation across longer time periods.

5.6 Overall Conclusions

This thesis builds on and extends previous work in statistical learning to explore the mechanisms involved in the generalisation of grammatical regularities, focusing on the role of item-level vocabulary knowledge, and the contributions of explicit and implicit knowledge. Consistent with lexicalist theories (e.g., Bates & Goodman, 1997; MacDonald et al., 1994), the experiments presented in Chapter 2 demonstrate that vocabulary is important for grammatical generalisation, when considering the vocabulary within a learner's lexicon. In adults, reducing vocabulary levels within the lexicon reduced grammatical generalisation performance. This finding can also be interpreted in terms of constructivist approaches (e.g., Goldberg, 2005, 2009) which suggest that when a critical mass of vocabulary knowledge has been reached within the lexicon, the more salient and regular aspects of the language can be abstracted and generalised.

In relation to the contributions of explicit and implicit knowledge, the experiments presented in Chapter 4 demonstrated the involvement of both these learning systems in grammatical generalisation. This extends the very small body of recent work which has shown both implicit and explicit contributions to performance on the word-boundary detection paradigm (Batterink et al., 2015), as well as to grammatical learning (but not generalisation, Monaghan et al., 2019). The findings challenge the long-held assumption in the statistical learning literature that learning is unitary and implicit, and instead align with recent theoretical proposals for a multi-componential view of statistical learning (Conway, 2020; R. Frost et al., 2019). Conway (2020) specifically suggests an interacting dual-system including both implicit and explicit learning mechanisms. The implicit learning mechanisms are based in the lower perceptual regions of the brain and are proposed to reflect the learning of 'simple', time-constrained and domain-specific regularities. However, these are underpinned by domain-general principles of neural plasticity that improves processing through learning. The explicit learning mechanisms reflect downstream brain systems (such as the prefrontal cortex) and are proposed to reflect the learning of more complex and global regularities which require learning over longer time periods. The theory proposes that the two systems can work in parallel, but can also work in collaboration or in competition with one another (Batterink et al., 2019; Conway, 2020). The findings from this thesis provide behavioural evidence in support of this theory of statistical learning, although they do not speak directly to the possible interactions between explicit and implicit processes.

As with all work in this field, the broad motivation for the work presented in this thesis is to shed light on the process of language learning, including first language acquisition in young children, but it is important to acknowledge that the majority of the findings within this thesis speak to statistical learning and grammatical generalisation in adults. Furthermore, the finding of explicit contributions to learning in adults but not children, highlights the need to take developmental changes into account, and not to assume that adult learners can be used as proxies for younger learners. Speculatively, we propose that the same implicit learning processes are at play in adults and children, with later-emerging explicit processes playing a role only as the system matures which potentially may be related to the availability of prior knowledge within specific learning contexts. We also propose that vocabulary knowledge supports grammatical generalisation in a similar way in adults and children, based on the finding that when adult's vocabulary knowledge was reduced to child levels, they showed a similar lack of grammatical generalisation. Further studies with children (which had been planned before the pandemic), would enable more direct investigation of these possibilities, and could help to further develop the recent multi-componential theories of statistical learning, which do not currently address issues of developmental change (Batterink et al., 2019; Conway, 2020; Frost et al., 2019).

Appendices

Appendix A

Table A1: Training items used across all Experiments in Chapter 2 and 4.

Word	Picture
tib mofeem	dog
tib hormeem	sheep
tib doudeem	horse
tib shemeem	chicken
tib tercheem	parrot
tib gatcheem	snake
tib norneem	lion
tib flarneem	owl
tib vedeem	cow
tib zeapeem	duck
tib viffeem	tiger
tib phlaveem	squirrel
tib jerleem	crocodile
tib ropeem	elephant
tib morcheem	fox
tib dupeem	mouse
ked borchool	knife
ked roivool	toaster
ked snarool	tv
ked larshool	table
ked felchool	hairbrush
ked dranool	umbrella
ked pilmool	toothbrush
ked migool	box
ked tormool	window
ked jorool	lamp
ked shillool	ipod
ked cassool	guitar
ked vevool	pencil
ked geshool	spoon
ked gniddool	bed
ked wilnool	clock

Table A2: Mismatched items used in the word-picture matching training task in Chapter 2 and Chapter 4 Experiment 2.

Word	Picture
tib vedeem	window
tib zeapeem	lamp
tib viffeem	ipod
tib phlaveem	guitar
tib jerleem	cow
tib ropeem	duck
tib morcheem	tiger
tib dupeem	squirrel
ked tormool	pencil
ked jorool	spoon
ked shillool	bed
ked cassool	clock
ked vevool	parrot
ked geshool	snake
ked gniddool	lion
ked wilnool	owl
tib mofeem	knife
tib hormeem	toaster
tib doudeem	tv
tib shemeem	table
tib tercheem	dog
tib gatcheem	sheep
tib norneem	horse
tib flarneem	chicken
ked borchool	hairbrush
ked roivool	umberella
ked snarool	toothbrush
ked larshool	box
ked felchool	crocodile
ked dranool	elephant
ked pilmool	fox
ked migool	mouse

Table A3: New items used in the phonological form old and new task in Chapter 2 and Chapter 4 Experiment 1

Word
ked bisool
ked zurrool
ked dowthool
ked furgool
ked garrool
ked jurbool
ked perrbool
ked plaunool
ked shreelool
ked treggool
ked tremool
ked zomool
ked goomool
ked vaddool
ked rukeool
ked craunool
tib blorcheem
tib chebeem
tib cherkeem
tib chowtheem
tib dreteem
tib gorteem
tib heefeem
tib lupeem
tib sarbeem
tib scoiffeem
tib zarfeem
tib zimbeem
tib yawgeem
tib twalleem
tib smowneem
tib kigeem

Table A4: Determiner and suffix word-picture matching generalisation task items

Word	Picture	Consistency	Experiment used
tib lekeem	rabbit	consistent	Chapter 2 & 4
tib taseem	giraffe	consistent	Chapter 2 & 4
tib newleem	penguin	consistent	Chapter 4
ked rournool	glasses	consistent	Chapter 2 & 4
ked tarbool	door	consistent	Chapter 2 & 4
ked snoveool	balloon	consistent	Chapter 4
ked soid**ool**	wolf	inconsistent	Chapter 2 & 4
ked sird**ool**	pig	inconsistent	Chapter 2 & 4
ked floach**ool**	gorilla	inconsistent	Chapter 4
tib darl**eem**	bowl	inconsistent	Chapter 2 & 4
tib bup**eem**	camera	inconsistent	Chapter 2 & 4
tib guch**eem**	peg	inconsistent	Chapter 4

Bold morphemes highlight where inconsistencies are with the picture pairing

Table A5: Determiner Only word-picture matching generalisation task items

Word	Picture	Consistency	Experiment used
tib perneem	kangaroo	consistent	Chapter 4
tib wholeem	monkey	consistent	Chapter 4
tib wrewsheem	bear	consistent	Chapter 4
ked volmool	anchor	consistent	Chapter 4
ked hylmbool	basket	consistent	Chapter 4
ked vazeool	bike	consistent	Chapter 4
tib grurmbool	button	inconsistent	Chapter 4
tib malool	key	inconsistent	Chapter 4
tib haiceool	kite	inconsistent	Chapter 4
ked poveem	hedgehog	inconsistent	Chapter 4
ked henkeem	frog	inconsistent	Chapter 4
ked vugeem	bird	inconsistent	Chapter 4

Bold morphemes highlight where inconsistencies are with the picture pairing

Table A6: Suffix only word-picture matching generalisation task items

Word	Picture	Consistency	Experiment Used
tib thileem	panda	consistent	Chapter 2 & 4
tib vangeem	cat	consistent	Chapter 2 & 4
tib daikeem	swan	consistent	Chapter 4
ked shubool	chair	consistent	Chapter 2 & 4
ked geechool	book	consistent	Chapter 2 & 4
ked knonnool	sofa	consistent	Chapter 4
tib sen**ool**	goat	inconsistent	Chapter 2 & 4
tib sheg**ool**	camel	inconsistent	Chapter 2 & 4
tib scis**ool**	peacock	inconsistent	Chapter 4
ked chush**eem**	ruler	inconsistent	Chapter 2 & 4
ked heg**eem**	computer	inconsistent	Chapter 2 & 4
ked dorg**eem**	bell	inconsistent	Chapter 4

Bold morphemes highlight where inconsistencies are with the picture pairing

Table A7:Phonological form word-picture matching generalisation task items

Word	Consistency	Experiment Used
ked thrairool	consistent	Chapter 2 & 4
ked rutchool	consistent	Chapter 2 & 4
tib pesseem	consistent	Chapter 2 & 4
tib goiteem	consistent	Chapter 2 & 4
tib norkeem	consistent	Chapter 4
ked litchool	consistent	Chapter 4
tib jit**ool**	inconsistent	Chapter 2 & 4
tib lerth**ool**	inconsistent	Chapter 2 & 4
ked rish**eem**	inconsistent	Chapter 2 & 4
ked narp**eem**	inconsistent	Chapter 2 & 4
tib spom**ool**	inconsistent	Chapter 4
ked voap**eem**	inconsistent	Chapter 4

Bold morphemes highlight the inconsistent grammatical regularities

Table A8: List 1 - Consistent trained and untrained items for the phoneme detection task in Chapter 4, Experiment 2.

Word	Picture	Trained or Untrained
tib mofeem	dog	Trained
tib hormeem	sheep	Trained
tib doudeem	horse	Trained
tib shemeem	chicken	Trained
ked borchool	knife	Trained
ked roivool	toaster	Trained
ked snarool	tv	Trained
ked larshool	table	Trained
ked bisool	helicopter	Untrained
ked zurrool	hammer	Untrained
ked dowthool	glass	Untrained
ked rukeool	-	Untrained
tib blorcheem	cockerel	Untrained
tib chebeem	deer	Untrained
tib cherkeem	bee	Untrained
tib smowneem	-	Untrained

Table A9: List 2 - Consistent trained and untrained items for the phoneme detection task in Chapter 4, Experiment 2.

Word	Picture	Trained or Untrained
tib tercheem	parrot	Trained
tib gatcheem	snake	Trained
tib norneem	lion	Trained
tib flarneem	owl	Trained
ked felchool	hairbrush	Trained
ked dranool	umberella	Trained
ked pilmool	toothbrush	Trained
ked migool	box	Trained
ked furgool	envelope	Untrained
ked garrool	fence	Untrained
ked jurbool	drum	Untrained
ked craunool	-	Untrained
tib chowtheem	ant	Untrained
tib dreteem	cheetah	Untrained
tib gorteem	lobster	Untrained
tib kigeem	-	Untrained

Table A10: List 3 - Consistent trained and untrained items for the phoneme detection task in Chapter 4, Experiment 2.

Word	**Picture**	**Trained or Untrained**
tib vedeem	cow	Trained
tib zeapeem	duck	Trained
tib viffeem	tiger	Trained
tib phlaveem	squirrel	Trained
ked tormool	window	Trained
ked jorool	lamp	Trained
ked shillool	ipod	Trained
ked cassool	guitar	Trained
ked shreelool	candle	Untrained
ked treggool	bottle	Untrained
ked tremool	boot	Untrained
ked perrbool	-	Untrained
tib sarbeem	donkey	Untrained
tib scoiffeem	racoon	Untrained
tib zarfeem	rhino	Untrained
tib heefeem	-	Untrained

Table A11: List 4 - Consistent trained and untrained items for the phoneme detection task in Chapter 4, Experiment 2.

Word	Picture	Trained or Untrained
tib jerleem	crocodile	Trained
tib ropeem	elephant	Trained
tib morcheem	fox	Trained
tib dupeem	mouse	Trained
ked vevool	pencil	Trained
ked geshool	spoon	Trained
ked gniddool	bed	Trained
ked wilnool	clock	Trained
ked zomool	bin	Untrained
ked goomool	belt	Untrained
ked vaddool	bat	Untrained
ked plounool	-	Untrained
tib zimbeem	beetle	Untrained
tib yawgeem	butterfly	Untrained
tib twalleem	turtle	Untrained
tib lupeem	-	Untrained

Table A12: Inconsistent untrained items for the phoneme detection task used across lists 1-4 in Chapter 4, Experiment 2.

Word	Picture	Inconsistent Type
tib regeem	axe	DetSuffGen
tib beveeem	rocking chair	DetSuffGen
ked zugool	goose	DetSuffGen
ked dwarlool	starfish	DetSuffGen
tib pirnool	swordfish	SuffOnGen
tib fecool	hippo	SuffOnGen
ked nazeem	barrell	SuffOnGen
ked frapeem	flag	SuffOnGen
tib malool	ball	DetOnGen
tib kelool	pram	DetOnGen
ked dilceem	hammerhead	DetOnGen
ked scwayeem	pigeon	DetOnGen
tib megool	-	PhonGen
tib mormool	-	PhonGen
ked harleem	-	PhonGen
ked nerpeem	-	PhonGen

DetSuffGen = Determiner & Suffix inconsistent item construction

SuffOnGen = Suffix Only inconsistent item construction

DetOneGen = Determiner Only inconsistent item construction

PhonGen = Phonological Form inconsistent item construction

Appendix B

B1. Pre-registered exclusion criteria for Chapter 4, Experiment 2

1. Participants who click the same smiley scale judgement across all four training tasks will be excluded for planned training task analysis.
2. Participants who click a picture on the same side of the screen across all four of the two alternative forced choice tasks during the training protocol will be excluded from the planned training task analysis.
3. Participants who click the same smiley scale judgement across all four of the word-picture matching generalisation tasks.
4. Participants who consistently fail to make a response within the allotted time for *20%* of responses within a task will be excluded from analysis for that task.
    1. A 20% no response rate is based on the no response rate found in previous studies in this research project. These previous studies have been conducted within a laboratory setting where participants were monitored for attention. Previous studies show the largest mean no response rate within a task was 20.83%, with most tasks showing a much lower rate if any at all. This indicated that a no response can occur when there is good attention or inattention isn't deliberate or prolonged. So, this needed to be factored into this exclusion criteria and used the mean highest no response rate as our exclusion bench mark.
5. Participants who report display or loading issues during the study.
6. Participants who report writing down presented words during the study.
7. Participants who fail to complete the study within a 3-hour time period.

B2. Participant Exclusions for Chapter 4, Experiment 2

Attempts

In total 89 people attempted the study. 12 withdrew before completing the training tasks and 13 failed the microphone and sound check, leaving 63 participants who took part in the experiment.

Complete Exclusion

Following the pre-registered exclusion criteria (see Appendix B1 or https://osf.io/xdvyg), 4 participants were completely excluded from analysis. 1 participant completed the study over the 3-hour time limit (they completed it over 24 hours). 2 participants reported technical and connectivity issue during the study. 1 participant only pressed one response key across all of the repetition and word-picture matching training tasks and all bar one of the word-picture matching generalisation tasks. Within the 'phonological form' word-picture matching generalisation task they pressed the same response option for 9 out of the 12 items and one other response option for the other 3 items. This left a total of 59 participants for the final analysis

Partial Exclusions

13 participants of the total 59 were excluded from analysis within certain tasks, based on the pre-registered exclusion criteria (see Appendix B1 or https://osf.io/xdvyg). The participants affected, along with which tasks they were excluded from and why can be found in the following table B2.

B3: Table detailing partial exclusions for Chapter 4, Experiment 2

Participant ID	Task excluded from	Reason
G14	RepWPM1	22.5% NRR
G23	RepWPM1,	20% NRR
	RepWPM2,	25% NRR
	RepWPM4	37.5% NRR
G35	RepWPM4	20% NRR
G40	RepWPM1	20% NRR
G45	RepWPM1,	42.5% NRR
	RepWPM2,	32.5% NRR
	RepWPM3	52.5% NRR
G48	RepWPM1	20% NRR
G49	All WPM Generalisation	Did not complete these tasks
	Phoneme Detection	
	Picture Naming	
	Debriefing Questionnaire	
G53	AFC1	21.88% NRR
	PhonGen	Did not complete these tasks
	Phoneme Detection	
	Picture Naming	
	Debriefing Questionnaire	
G54	RepWPM1	Only pressed 1 response option
G59	RepWPM1,	25% NRR
	RepWPM2,	25% NRR
	RepWPM3,	25% NRR
	RepWPM4	27.5% NRR
	SuffOnGen	25% NNR
G60	RepWPM1	37.5% NRR

RepWPM = Repetition and Word-Picture Matching training task

AFC = Two Alternative Forced Choice task

SuffOnGen = Suffix Only Word-Picture Matching Generalisation task

PhonGen = Phonological Form Word-Picture Matching Generalisation task

NRR = No Response Rate

References

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

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/S13428-019-01237-X/TABLES/8

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

Aslin, R. N. (2017). Statistical learning: a powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(1–2), 1373. https://doi.org/10.1002/wcs.1373

Aslin, R. N., & Newport, E. L. (2012). Statistical Learning: From Acquiring Specific Items to Forming General Rules. *Current Directions in Psychological Science*, *21*(3), 170–176. https://doi.org/10.1177/0963721412436806

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324. https://doi.org/10.1111/1467-9280.00063

Baker, C. L., & McCarthy, J. J. (1981). *The Logical Problem of Language Acquisition*. MIT press.

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552. https://doi.org/10.1037/A0033268

Bates, E., & Goodman, J. C. (1997). On the Inseparability of Grammar and the Lexicon: Evidence from Acquisition, Aphasia and Real-time Processing. *Language and Cognitive Processes*, *12*(5–6), 507–584. https://doi.org/10.1080/016909697386628

Batterink, L. J., Paller, K. A., & Reber, P. J. (2019). Understanding the Neural Bases of Implicit and Statistical Learning. *Topics in Cognitive Science*, *11*(3), 482–503. https://doi.org/10.1111/tops.12420

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit Statistical learning allows implicit and explicit contributions to statistical learning. *Journal of Memory and Language*, *83*, 62–78. https://doi.org/10.1016/j.jml.2015.04.004

Berko, J. (1958). The Child's Learning of English Morphology. WORD, *14*(2–3), 150–177. https://doi.org/10.1080/00437956.1958.11659661

Bowerman, M. (1973). Structural Relationships in children's Utterances: Syntactic of Semantic? In *Cognitive Development and Acquisition of Language* (pp. 197–213). Academic Press, Inc. https://doi.org/10.1016/b978-0-12-505850-6.50015-3

Braine, M. D. S. (1987). What is learned in acquiring word classes—A step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language aquisition* (pp. 65–87). Lawrence

Erlbaum Associates, Inc.

Breitenstein, C., Zwitserlood, P., de Vries, M. H., Feldhues, C., Knecht, S., & Dobel, C. (2007). Five days versus a lifetime: Intense associative vocabulary training generates lexically integrated words. *Restorative Neurology and Neuroscience*, *25*(5,6), 493–500.

Brown, H., Smith, K., Samara, A., & Wonnacott, E. (2022). Semantic cues in language learning: an artificial language study with adult and child learners. *Language, Cognition and Neuroscience*, *37*(4), 509–531. https://doi.org/10.1080/23273798.2021.1995612

Channon, S., Shanks, D., Johnstone, T., Vakili, K., Chin, J., Sinclair, E., & Rebuschat, P. (2002). Is implicit learning spared in amnesia? Rule abstraction and item familiarity in artificial grammar learning. *Language Learning*, *40*(3), 2185–2197. https://doi.org/10.1016/S0028-3932(02)00037-4

Cheesman, J., & Merikle, P. M. (1984). Priming with and without awareness. *Perception & Psychophysics*, *36*(4), 387–395. https://doi.org/10.3758/BF03202793

Chetail, F., & Mathey, S. (2013). Interaction between phonemic abilities and syllable congruency effect in young readers*. *Journal of Child Language*, *40*(2), 492–508. https://doi.org/10.1017/S0305000911000493

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press.

Christiansen, M. H. (2019). Implicit Statistical Learning: A Tale of Two Literatures. *Topics in Cognitive Science*, *11*(3), 468–481. https://doi.org/10.1111/TOPS.12332

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to Segment Speech Using Multiple Cues: A Connectionist Model. *Language and Cognitive Processes*, *13*(2–3), 221–268. https://doi.org/10.1080/016909698386528

Christiansen, M. H., & Chater, N. (2015). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*(May). https://doi.org/10.1017/S0140525X1500031X

Cleeremans, A. (1997). Principles for Implcict Learning. In D. Berry (Ed.), *How implicit is implicit learning?* (Vol. 15). Oxford University Press.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1038/141613a0

Conway, C. M. (2020). How does the brain learn environmental structure? Ten core principles for understanding the neurocognitive mechanisms of statistical learning. *Neuroscience and Biobehavioral Reviews*, *112*(August 2019), 279–299. https://doi.org/10.1016/j.neubiorev.2020.01.032

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning Memory and Cognition*, *31*(1), 24–39. https://doi.org/10.1037/0278-7393.31.1.24

Corbett, G. G. (1991). *Gender*. Cambridge University Press.

Corbett, G. G. (2013). The Expression of Gender. In *The Expression of Gender*. De Gruyter.

https://doi.org/10.1515/9783110307337

Culbertson, J., Gagliardi, A., & Smith, K. (2017). Competition between phonological and semantic cues in noun class learning. *Journal of Memory and Language*, *92*, 343–358. https://doi.org/10.1016/J.JML.2016.08.001

Culbertson, J., Jarvinen, H., Haggarty, F., & Smith, K. (2019). Children's sensitivity to phonological and semantic cues during noun class learning: Evidence for a phonological bias. *Language*, *95*(2), 268–293. https://doi.org/10.1353/lan.2019.0031

Cutler, A. (1994). Segmentation problems, rhythmic solutions. *Lingua*, *92*(C), 81–104. https://doi.org/10.1016/0024-3841(94)90338-7

Davies, C., Andrés-Roqueta, C., & Norbury, C. F. (2016). Referring expressions and structural language abilities in children with specific language impairment: A pragmatic tolerance account. *Journal of Experimental Child Psychology*, *144*, 98–113. https://doi.org/10.1016/j.jecp.2015.11.011

Dell, G. S., & Newman, J. E. (1980). Detecting phonemes in fluent speech. *Journal of Verbal Learning and Verbal Behavior*, *19*(5), 608–623. https://doi.org/10.1016/S0022-5371(80)90661-1

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

Dienes, Z. (2007). Subjective measures of unconscious knowledge. *Progress in Brain Research*, *168*, 1–45. https://doi.org/10.1016/S0079-6123(07)68005-4

Dienes, Z. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), *Behavioral Methods in Consciousness Research* (pp. 199–220). Oxford University Press. https://doi.org/10.1093/ACPROF:OSO/9780199688890.003.0012

Dienes, Z., Altmann, G. T. M., Kwan, L., & Goode, A. (1995). Unconscious Knowledge of Artificial Grammars Is Applied Strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1322–1338. https://doi.org/10.1037/0278-7393.21.5.1322

Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic Bulletin and Review*, *4*(1), 3–23. https://doi.org/10.3758/BF03210769

Dienes, Z., & Perner, J. (1996). Implicit knowledge in people and connectionist networks. In G. D. M. Underwood (Ed.), *Implicit Cognition* (pp. 227–255). Oxford University Press.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, *22*(5), 735–808. https://doi.org/10.1017/S0140525X99522188

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: distinguishing structural knowledge and judgment knowledge. *Psychological Research 2005 69:5*, *69*(5), 338–351. https://doi.org/10.1007/S00426-004-0208-3

Dienes, Z., & Seth, A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition*, *19*(2), 674–681. https://doi.org/10.1016/j.concog.2009.09.009

Dobel, C., Junghöfer, M., Breitenstein, C., Klauke, B., Knecht, S., Pantev, C., & Zwitserlood, P. (2010). New names for known things: On the association of novel word forms with existing semantic information. *Journal of Cognitive Neuroscience*, *22*(6), 1251–1261. https://doi.org/10.1162/jocn.2009.21297

Donaldson, W. (1992). Measuring Recognition Memory. *Journal of Experimental Psychology: General*, *121*(3), 275–277. https://doi.org/10.1037/0096-3445.121.3.275

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1998). *Rethinking innateness: A connectionist perspective on development*. MIT press.

Evans, S., & Azzopardi, P. (2007). Evaluation of a "bias-free" measure of awareness. *Spatial Vision*, *20*(1–2), 61–77. https://doi.org/10.1163/156856807779369742

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(32), 12203–12208. https://doi.org/10.1073/pnas.0602173103

Ferman, S., & Karni, A. (2010). No Childhood Advantage in the Acquisition of Skill in Using an Artificial Language Rule. *PLoS ONE*, *5*(10), 13648. https://doi.org/10.1371/journal.pone.0013648

Ferman, S., Olshtain, E., Schechtman, E., & Karni, A. (2009). The acquisition of a linguistic skill by adults: Procedural and declarative memory interact in the learning of an artificial morphological rule. *Journal of Neurolinguistics*, *22*(4), 384–412. https://doi.org/10.1016/J.JNEUROLING.2008.12.002

Fernald, A., & Marchman, V. A. (2006). Language Learning in Infancy. In *Handbook of Psycholinguistics* (pp. 1027–1071). Academic Press. https://doi.org/10.1016/B978-012369374-7/50028-6

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

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906. https://psycnet.apa.org/fulltext/1980-09388-001.pdf?auth_token=a62672b5506f64ae2f1edfb4c41085db4db49949

Flavell, J. H. (1999). Cognitive Developement: Children's Knowledge About the Mind. *Annual Review of Psychology*, *50*(1), 21–45. www.annualreviews.org

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. In *Behavior Research Methods, Instruments, and Computers* (Vol. 35, Issue 1, pp.

116–124). Psychonomic Society Inc. https://doi.org/10.3758/BF03195503

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do Procedures for Verbal Reporting of Thinking Have to Be Reactive? A Meta-Analysis and Recommendations for Best Reporting Methods. *Psychological Bulletin*, *137*(2), 316–344. https://doi.org/10.1037/A0021663

Franco, A., Cleeremans, A., & Destrebecqz, A. (2011). Statistical learning of two artificial languages presented successively: How conscious? *Frontiers in Psychology*, *2*(SEP), 1–12. https://doi.org/10.3389/fpsyg.2011.00229

Franco, A., Cleeremans, A., & Destrebecqz, A. (2016). Objective and subjective measures of cross-situational learning. *Acta Psychologica*, *165*, 16–23. https://doi.org/10.1016/j.actpsy.2016.02.001

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

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

Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2019). Mark My Words: High Frequency Marker Words Impact Early Stages of Language Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000683

Gathercole, V. C. M., & Hoff, E. (2007). Input and the Acquisition of Language: Three Questions. In *Blackwell Handbook of Language Development* (pp. 107–127). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470757833.CH6

Gentner, D. (1982). Why Nouns Are Learned before Verbs: Linguistic Relativity Versus Natural Partitioning. *Centre for the Study of Reading Technical Report; No.257*. https://eric.ed.gov/?id=ED219724

Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek & R. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 544–564). Oxford University Press.

Gerken, L. (2007). Acquiring Linguistic Structure. In *Blackwell Handbook of Language Development* (pp. 173–190). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470757833.CH9

Goldberg, A. E. (1999). The Emergence of the Semantics of Argument Structure Constructions. In B. MacWhinney (Ed.), *The Emergence of Language* (pp. 215–230). Psychology Press. https://doi.org/10.4324/9781410602367-12

Goldberg, A. E. (2005). Constructions at Work: The Nature of Generalization in Language. *Constructions at Work: The Nature of Generalization in Language*, 1–288. https://doi.org/10.1093/ACPROF:OSO/9780199268511.001.0001

Goldberg, A. E. (2009). The nature of generalization in language. *Cognitive Linguistics*, *20*(1), 93–127. https://doi.org/10.1515/COGL.2009.005

Gómez, R. L. (2002). Variability and Detection of Invariant Structure. *Psychological Science*, *13*(5), 431–436. https://doi.org/https://doi.org/10.1111/1467-9280.00476

Gómez, R. L. (2017). Do infants retain the statistics of a statistical learning experience? Insights from a developmental cognitive neuroscience perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160054. https://doi.org/10.1098/rstb.2016.0054

Gómez, R. L. ., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135. https://doi.org/10.1016/S0010-0277(99)00003-7

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and acquisition. *J. Cogn. Neurosci*, *99*(5), 70. https://doi.org/https://doi.org/10.1016/S1364-6613(00)01467-4

Gómez, R. L., & Maye, J. (2009). The Developmental Trajectory of Nonadjacent Dependency Learning The Developmental Trajectory of Nonadjacent Dependency Learning. *Online*, *7*(918975534), 183–206. https://doi.org/10.1207/s15327078in0702

Grunow, H., Spaulding, T. J., Gómez, R. L., & Plante, E. (2006). The effects of variation on learning word order rules by adults with and without language-based learning disabilities. *Journal of Communication Disorders*, *39*(2), 158–170. https://doi.org/10.1016/j.jcomdis.2005.11.004

Hall, J., Owen van Horne, A., & Farmer, T. (2018). Distributional learning aids linguistic category formation in school-age children. *Journal of Child Language*, *45*(3), 717–735. https://doi.org/10.1017/s0305000917000435

Hebb, D. O. (1961). Brain mechanisms and learning. In *Distinctive features of learning in the higher animal*. Oxford: Blackwell.

Henderson, L., Devine, K., Weighall, A., & Gaskell, G. (2015). When the daffodat flew to the intergalactic zoo: Off-line consolidation is critical for word learning from stories. *Developmental Psychology*, *51*(3), 406–417. https://doi.org/10.1037/a0038786

Henderson, L. M., & James, E. (2018). Consolidating new words from repetitive versus multiple stories: Prior knowledge matters. *Journal of Experimental Child Psychology*, *166*, 465–484. https://doi.org/10.1016/j.jecp.2017.09.017

Henderson, L. M., Weighall, A. R., Brown, H., & Gaskell, M. G. (2012). Consolidation of vocabulary is associated with sleep in children. *Developmental Science*, *15*(5), 674–687. https://doi.org/10.1111/j.1467-7687.2012.01172.x

Hopman, E. W. M., & MacDonald, M. C. (2018). Production Practice During Language Learning Improves Comprehension. *Psychological Science*, *29*(6), 961–971. https://doi.org/10.1177/0956797618754486

Hsu, H. J., & Bishop, D. V. M. (2010). Grammatical difficulties in children with specific language impairment: Is learning deficient? *Human Development*, *53*(5), 264–277. https://doi.org/10.1159/000321289

Icht, M., & Mama, Y. (2015). The production effect in memory: A prominent mnemonic in children. *Journal of Child Language*, *42*(5), 1102–1124. https://doi.org/10.1017/S0305000914000713

Isbilen, E. S., & Christiansen, M. H. (2020). Chunk-Based Memory Constraints on the Cultural Evolution of Language. *Topics in Cognitive Science*, *12*(2), 713–726. https://doi.org/10.1111/tops.12376

Isbilen, E. S., Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2018). Bridging artificial and natural language learning : Comparing processing- and reflection-based measures of learning. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 1856–1861.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically Induced Chunking Recall: A Memory-Based Approach to Statistical Learning. *Cognitive Science*, *44*(7), e12848. https://doi.org/10.1111/cogs.12848

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In F. Craik & H. L. Roediger III (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 391–422). Taylor & Francis.

James, E., Gaskell, M. G., Weighall, A., & Henderson, L. (2017). Consolidation of vocabulary during sleep: The rich get richer? *Neuroscience and Biobehavioral Reviews*, *77*, 1–13. https://doi.org/10.1016/j.neubiorev.2017.01.054

Jusczyk, P. W., & Aslin, R. N. (1995). Infants′ Detection of the Sound Patterns of Words in Fluent Speech. *Cognitive Psychology*, *29*(1), 1–23. https://doi.org/10.1006/COGP.1995.1010

Karmiloff-Smith, A. (1986). From meta-processes to conscious access: Evidence from children's metalinguistic and repair data. *Cognition*, *23*(2), 95–147. https://doi.org/10.1016/0010-0277(86)90040-5

Katsos, N., & Bishop, D. V. M. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67–81. https://doi.org/10.1016/j.cognition.2011.02.015

Katsos, N., & Smith, N. (2010). Pragmatic Tolerance or a Speaker-Comprehender Asymmetry in the Acquisition of Informativeness? *Proceedings of the 34Th Annual Boston University Conference on Language Development, Vols 1 and 2, January 2010*, 221–232.

Kelley, C. M., & Jacoby, L. L. (2000). Recollection and familiarity: Process-dissociation. In E. Tulving & F. Craik (Eds.), *The Oxford handbook of memory*. Oxford Univeristy Press.

Kelley, C. M., & Lindsay, D. S. (1996). Conscious and Unconscious Forms of Memory. *Memory*, 31–63. https://doi.org/10.1016/b978-012102570-0/50004-5

Kelly, M. H. (1992). Using sound to solve syntactic problems: The Role of phonology in grammatical category assignments. *Psychological Review*, *99*(2), 349–364. https://doi.org/10.1037/0033-295X.99.2.349

Krogh, L., Vlach, H. A., & Johnson, S. P. (2013). Statistical learning across development: Flexible yet constrained. *Frontiers in Psychology*, *3*(JAN), 1–11. https://doi.org/10.3389/fpsyg.2012.00598

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*(3), 294–340. https://doi.org/10.1006/ccog.2000.0494

Lachman, J. L., Lachman, R., & Thronesbery, C. (1979). Metamemory through the adult life span. *Developmental Psychology*, *15*(5), 543–551. https://doi.org/https://doi.org/10.1037/0012-1649.15.5.543

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2017). Statistical Learning in Specific Language Impairment: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *60*(12), 3474–3486. https://doi.org/10.1044/2017_jslhr-l-16-0439

Lany, J. (2014a). Judging words by their covers and the company they keep: Probabilistic cues support word learning. *Child Development*, *85*(4), 1727–1739. https://doi.org/10.1111/cdev.12199

Lany, J. (2014b). *Twelve-Month-Old Infants Benefit From Prior Experience in Statistical Learning NIH Public Access*. *19*(July), 1247–1252. https://doi.org/10.1111/j.1467-9280.2008.02233.x

Lany, J., & Saffran, J. R. (2010). From statistics to meaning: Infants' acquisition of lexical categories. *Psychological Science*, *21*(2), 284–291. https://doi.org/10.1177/0956797609358570

Lany, J., & Saffran, J. R. (2011). Interactions between statistical and semantic information in infant language development. *Developmental Science*, *14*(5), 1207–1219. https://doi.org/10.1111/j.1467-7687.2011.01073.x

Lany, J., & Saffran, J. R. (2013). *Comprehensive Developmental Neuroscience: Neural Circuit Development and Function in the Brain*. *3*. https://doi.org/10.1016/B978-0-12-397267-5.00034-0

Lew-Williams, C., & Fernald, A. (2007). Young Children Learning Spanish Make Rapid Use of Grammatical Gender in Spoken Word Recognition. *PSychological Science*, *18*(3), 193–198.

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language*, *63*(4), 447–464. https://doi.org/10.1016/j.jml.2010.07.003

Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*(1), 187–219. https://doi.org/10.1017/S0305000996002930

Lindsay, D. S., & Kelley, C. M. (1996). Creating Illusions of Familiarity in a Cued Recall

Remember/Know Paradigm. *Journal of Memory and Language*, *35*(2), 197–211. https://doi.org/10.1006/jmla.1996.0011

Lockl, K., & Schneider, W. (2002). Developmental trends in children's feeling-of-knowing judgements. *International Journal of Behavioral Development*, *26*(4), 327–333. https://doi.org/10.1080/01650250143000210

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703. https://doi.org/10.1037/0033-295x.101.4.676

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/J.CONCOG.2011.09.021

Marchman, V. A., & Bates, E. (1994). Continuity in lexical and morphological development: a test of the critical mass hypothesis. *Journal of Child Language*, *21*(2), 339–366. https://doi.org/10.1017/S0305000900009302

Mehler, J., Dommergues, J., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, *20*, 298–305. https://doi.org/https://doi.org/10.1016/S0022-5371(81)90450-3

Merikle, P. M. (1994). On the futility of attempting to demonstrate null awareness. *Behaviour and Brain Sciences*, *17*(3), 412.

Merikle, P. M., & Reingold, E. M. (1992). Measuring unconscious perceptual processes. In R. F. Bornstein & T. S. Pittman (Eds.), *Perception without awareness: Cognitive, clinical, and social perspectives* (pp. 55–80). Guildford Press.

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition*, *30*(5), 678–686. https://doi.org/10.3758/BF03196424

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117. https://doi.org/10.1016/S0010-0277(03)00140-9

Mirkovic, J., Forrest, S., & Gaskell, M. G. (2011). Semantic Regularities in Grammatical Categories : Learning Grammatical Gender in an Artificial Language. *33rd Annual Conference of the Cognitive Science Society*, 324–329.

Mirkovic, J., & Gaskell, M. G. (2016). Does sleep improve your grammar? Preferential consolidation of arbitrary components of new linguistic knowledge. *PLoS ONE*, *11*(4), 1–26. https://doi.org/10.1371/journal.pone.0152489

Mirković, J., MacDonald, M. C., & Seidenberg, M. S. (2005). Where does gender come from? Evidence from a complex inflectional system. *Language and Cognitive Processes*, *20*(1–2), 139–167. https://doi.org/10.1080/01690960444000205

Mirković, J., Seidenberg, M. S., & Joanisse, M. F. (2011). Rules Versus Statistics: Insights From a Highly Inflected Language. *Cognitive Science*, *35*(4), 638–681. https://doi.org/10.1111/j.1551-6709.2011.01174.x

Mirkovic, J., Yee, M., Kennedy, M., & Hayiou-Thomas, M. E. (2021). *Cognitive mechanisms underpinning statistical learning of grammatical regularities in child and adult learners: Evidence for a multicomponential view*. https://doi.org/10.31234/OSF.IO/Y8XWM

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010a). On-line individual differences in statistical learning predict language processing. *Frontiers in Psychology*, *0*(SEP), 31. https://doi.org/10.3389/FPSYG.2010.00031/BIBTEX

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010b). Sequential Expectations: The Role of Prediction-Based Learning in Language. *Topics in Cognitive Science*, *2*(1), 138–153. https://doi.org/10.1111/J.1756-8765.2009.01072.X

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*(2), 143–182. https://doi.org/10.1016/j.cognition.2004.09.001

Monaghan, P., Schoetensack, C., & Rebuschat, P. (2019). A Single Paradigm for Implicit and Statistical Learning. *Topics in Cognitive Science*, *11*(3), 536–554. https://doi.org/10.1111/tops.12439

Moroshkina, N. V., Ivanchei, I. I., Karpov, A. D., & Ovchinnikova, I. (2019). The verbalization effect on implicit learning. *Implicit Learning*, 189–207. https://doi.org/10.4324/9781315628905-9

Morton, J., & Long, J. (1976). Effect of word transitional probability on phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, *15*(1), 43–51. https://doi.org/10.1016/S0022-5371(76)90005-0

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *37*(1), 1–19. https://doi.org/10.1017/S0140525X12003214

Newport, E. L. (1990). Maturational Constraints on Language Development. *Studies in Second Language Acquisition*, *12*(3), 251–285. https://doi.org/10.1017/s0272263100009165

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. C. (2015). Measuring consciousness with confidence ratings. In M. Overgaard (Ed.), *Behavioral Methods in Consciousness Research* (pp. 159–180). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199688890.003.0010

Overgaard, M., Timmermans, B., Sandberg, K., & Cleeremans, A. (2010). Optimizing subjective measures of consciousness. *Consciousness and Cognition*, *19*(2), 682–684. https://doi.org/10.1016/j.concog.2009.12.018

Page, M. P. A., & Norris, D. (2009). A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word forms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3737–3753.

https://doi.org/10.1098/rstb.2009.0173

Pallier, C. (2002). Computing discriminability and bias with the R software. *URL Http://Www. Pallier. Org/Ressources/Aprime/Aprime. …*, 1–6. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.218.7769&rep=rep1&type=pdf

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical Learning in a Natural Language by 8-Month-Old Infants. *Child Development*, *80*(3), 674–685. https://doi.org/https://doi.org/10.1111/j.1467-8624.2009.01290.x

Perruchet, P. (2019). What Mechanisms Underlie Implicit Statistical Learning? Transitional Probabilities Versus Chunks in Language Learning. *Topics in Cognitive Science*, *11*(3), 520–535. https://doi.org/10.1111/TOPS.12403

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

Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*(3), 217–283. https://doi.org/10.1016/0010-0277(79)90001-5

Pinker, S. (1984). *Language learnability and language development*. Harvard University Press.

Pinker, S. (1989). *Learnability and cognition: The acquistion of argument structure*. Harvard University Press.

Pipijin, K., & Schaeken, W. (2012). Children and pragmatic implicatures: A test of the pragmatic tolerance hypothesis with different tasks. *Building Bridges Across Cognitive Sciences Around the World. Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, 2186–2191.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*(1–12), 125–126. https://doi.org/10.3758/bf03342823

Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*(1), 28–52. https://doi.org/10.1191/0265532204lt273oa

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339–1362. https://doi.org/10.1080/02724980244000099

Raviv, L., & Arnon, I. (2017). The developmental trajectory of children's auditory and visual statistical learning abilities: modality-based differences in the effect of age. *Developmental Science*, *21*(4), 1–13. https://doi.org/10.1111/desc.12593

Reber, A. S. (1969). Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology*, *81*(1), 115–119. https://doi.org/10.1037/h0027454

Reber, A. S. (1976). Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(1), 88–94. https://doi.org/10.1037/0278-7393.2.1.88

Reber, A. S. (1989). Implicit Learning and Tacit Knowledge. *Journal of Experimental Psychology: General*, *118*(3), 219–235. https://doi.org/10.1037/0096-3445.118.3.219

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. Oxford Univeristy Press.

Reber, A. S., Kassin, S. M., Lewis, S., & Cantor, G. (1980). On the relationship between implicit and explicit modes in the learning of a complex rule structure. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 492–502. https://doi.org/10.1037/0278-7393.6.5.492

Reber, P. J. (2013). The neural basis of implicit learning and memory: A review of neuropsychological and neuroimaging research. *Neuropsychologia*, *51*(10), 2026–2042. https://doi.org/https://doi.org/10.1016/j.neuropsychologia.2013.06.019

Reber, P. J., Batterink, L. J., Thompson, K. R., & Reuveni, B. (2019). Implicit Learning: History & Applications. In A. Cleeremans, V. Allakhverdov, & M. Kuvaldina (Eds.), *Implicit learning* (pp. 16–27). Routledge.

Reber, P. J., & Squire, L. R. (1994). *Parallel Brain Systems for Learning with and without Awareness*.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*(3), 595–626. https://doi.org/10.1111/lang.12010

Redington, M., Nick, C., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*(4), 425–469. https://doi.org/https://doi.org/10.1207/s15516709cog2204_2

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

Rossion, B., & Pourtois, G. (2001). Revisiting Snodgrass and Vanderwart's object database: Color and texture improve object recognition. *Journal of Vision*, *1*(3), 413–413. https://doi.org/10.1167/1.3.413

Rubin, P., Turvey, M. T., & Van Gelder, P. (1976). Initial phonemes are detected faster in spoken words than in spoken nonwords. *Perception & Psychophysics*, *19*(5), 394–398. https://doi.org/10.3758/BF03199398

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., & Kirkham, N. Z. (2018). Infant Statistical Learning. *Ssrn*. https://doi.org/10.1146/annurev-psych-122216-011805

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621. https://doi.org/10.1006/jmla.1996.0032

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and Learning) out of the Corner of Your Ear. *Psychological Science*, *8*(2), 101–105. https://doi.org/10.1111/j.1467-9280.1997.tb00690.x

Saffran, J. R., & Thiessen, E. D. (2007). Domain-General Learning Capacities. In *Blackwell Handbook of Language Development* (pp. 68–86). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470757833.ch4

Sanchez, D. J., & Reber, P. J. (2013). Explicit pre-training instruction does not improve implicit perceptual-motor sequence learning. *Cognition*, *126*(3), 341–351. https://doi.org/10.1016/j.cognition.2012.11.006

Sandberg, K., Timmermans, B., Overgaard, M., & Cleeremans, A. (2010). Measuring consciousness: Is one measure better than the other? *Consciousness and Cognition*, *19*(4), 1069–1078. https://doi.org/10.1016/j.concog.2009.12.013

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038. https://doi.org/10.1016/J.JML.2019.104038

Schmitt, N. (2014). Size and Depth of Vocabulary Knowledge: What the Research Shows. *Language Learning*, *64*(4), 913–951. https://doi.org/10.1111/LANG.12077

Schwab, J. F., Lew-Williams, C., & Goldberg, A. E. (2018). When regularization gets it wrong: children over-simplify language input only in production. *Journal of Child Language*, *45*(5), 1054–1072. https://doi.org/10.1017/s0305000918000041

Seidenberg, M. S. (1997). Language Acquisition and Use: Learning and Applying Probabilistic Constraints. *Science*, *275*(March), 1599–1603. https://doi.org/10.1126/science.275.5306.1599

Seidenberg, M. S., & MacDonald, M. C. (2018). The impact of language experience on language and reading a statistical learning approach. *Topics in Language Disorders*, *38*(1), 66–83. https://doi.org/10.1097/TLD.0000000000000144

Smalle, E. H. M., Bogaerts, L., Simonis, M., Duyck, W., Page, M. P. A., Edwards, M. G., & Szmalec, A. (2016). Can chunk size differences explain developmental changes in lexical learning? *Frontiers in Psychology*, *6*, 1925. https://doi.org/10.3389/fpsyg.2015.01925

Smalle, E. H. M., Page, M. P. A., Duyck, W., Edwards, M., & Szmalec, A. (2018). Children retain implicitly learned phonological sequences better than adults: a longitudinal study. *Developmental Science*, *21*(5), e12634. https://doi.org/10.1111/DESC.12634

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, *20*(4), 1787–1792. https://doi.org/10.1016/J.CONCOG.2010.12.011

Tamminen, J., Davis, M. H., Merkx, M., & Rastle, K. (2012a). The role of memory consolidation in generalisation of new linguistic information. *Cognition*, *125*(1), 107–112. https://doi.org/10.1016/J.COGNITION.2012.06.014

Tamminen, J., Davis, M. H., Merkx, M., & Rastle, K. (2012b). The role of memory consolidation in generalisation of new linguistic information. *Cognition*, *125*(1), 107–112. https://doi.org/10.1016/j.cognition.2012.06.014

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology*, *79*, 1–39. https://doi.org/10.1016/j.cogpsych.2015.03.003

Team, R. C. (2020). *R: A language and environment for statistical computing.* http://www.r-project.org/index.html

Timmermans, B., & Cleeremans, A. (2015). How can we measure awareness? An overview of current methods. In *Behavioral Methods in Consciousness Research* (pp. 21–46). https://doi.org/10.1093/acprof:oso/9780199688890.003.0003

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, *4*(4), 156–163. https://doi.org/10.1016/S1364-6613(00)01462-5

Tunney, R. J. (2005). Sources of confidence judgments in implicit cognition. *Psychonomic Bulletin and Review*, *12*(2), 367–373. https://doi.org/10.3758/BF03196386

Tunney, R. J., & Shanks, D. R. (2003). Subjective measures of awareness and implicit cognition. *Memory and Cognition*, *31*(7), 1060–1071. https://doi.org/10.3758/BF03196127

Visser, M., Krahmer, E., & Swerts, M. (2014). Children's Expression of Uncertainty in Collaborative and Competitive Contexts. *Language and Speech*, *57*(1), 86–107. https://doi.org/10.1177/0023830913479117

von Koss Torkildsen, J., Dailey, N. S., Aguilar, J. M., Gómez, R., & Plante, E. (2012). Exemplar Variability Facilitates Rapid Learning of an Otherwise Unlearnable Grammar by Individuals With Language-Based Learning Disability. *Journal of Speech, Language, and Hearing Research*, *56*(2), 618–629. https://doi.org/10.1044/1092-4388(2012/11-0125)

Vujović, M., Ramscar, M., & Wonnacott, E. (2021). Language learning as uncertainty reduction: The role of prediction error in linguistic generalization and item-learning. *Journal of Memory and Language*, *119*(February). https://doi.org/10.1016/j.jml.2021.104231

Wierzchoń, M., Asanowicz, D., Paulewicz, B., & Cleeremans, A. (2012). Subjective measures of consciousness in artificial grammar learning task. *Consciousness and Cognition*, *21*(3), 1141–1153. https://doi.org/10.1016/j.concog.2012.05.012

Williams, J. N. (2005). LEARNING WITHOUT AWARENESS. *Studies in Second Language Acquisition*, *27*(2), 269–304. https://doi.org/10.1017/S0272263105050138

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, *65*(1), 1–14. https://doi.org/10.1016/j.jml.2011.03.001

Wonnacott, E., Boyd, J. K., Thomson, J., & Goldberg, A. E. (2012). Input effects on the acquisition of a novel phrasal construction in 5year olds. *Journal of Memory and Language*, *66*(3), 458–478. https://doi.org/10.1016/j.jml.2011.11.004

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*(3), 165–209. https://doi.org/https://doi.org/10.1016/j.cogpsych.2007.04.002

Woodward, J. Z., & Aslin, R. N. (1990). Segmentation cues in maternal speech to infants. *7th Biennial Meeting of the International Conference on Infants Studies*.

Zamuner, T. S., Morin-Lessard, E., Strahm, S., & Page, M. P. A. (2016). Spoken word recognition of novel words, either produced or only heard during learning. *Journal of Memory and Language*, *89*, 55–67. https://doi.org/10.1016/j.jml.2015.10.003

Zhang, J., & Mueller, S. T. (2005). A note on ROC analysis and non-parametric estimate of sensitivity. *Psychometrika*, *70*(1), 203–212. https://doi.org/10.1007/s11336-003-1119-8

Zubin, D. A., & Köpcke, K. M. (1981). Gender: A less than arbitrary grammatical category. In *Papers from the Seventeenth Regional Meeting* (Issue 17, pp. 439–449).