Statistical learning of semantic and graphotactic regularities: Evidence from artificial orthography learning experiments

Nicole Sin Hang Law

Thesis submitted in fulfilment of the requirements for the Degree of
Doctor of Philosophy in Education (Applied Linguistics)

University of Oxford
University College

Supervisors: Professor Elizabeth Wonnacott and Professor Kate Nation

May 2025

Front Matter

Statement of Authorship

This thesis is presented as an integrated thesis and consists of three studies: a systematic review (Study 1) and four artificial orthography learning experiments (Study 2: Experiments 1 & 2; Study 3: Experiments 3 & 4). Each study was written in the format of article manuscripts with the intention for future publication. Throughout these studies, the use of the pronouns “we” and “our” reflects the multiple authorship of these studies. The following outlines the contribution of co-authors other than my DPhil supervisors, Prof Elizabeth Wonnacott and Prof Kate Nation:

Study 1, which contains a systematic review, was conducted in collaboration with Dr Johannes Schulz. His contribution involved conducting a dual-reviewer blind screening process to ensure reliability during the abstract and full-text screening stages.

Study 2, which included two artificial orthography learning experiments examining the simultaneous learning of semantic and graphotactic patterns, was conducted in collaboration with Dr Anna Samara, a collaborator of my supervisor. She provided advice on the initial conceptualisation of the experimental paradigm and stimuli.

Acknowledgement

I would like to express my deepest appreciation to my supervisors Prof Elizabeth Wonnacott and Prof Kate Nation. You truly formed the dream supervision team any DPhil student could wish for. Thank you, Liz, for guiding me through every step in designing my experiments, and for always challenging me to think critically about my data. Thank you, Kate, for showing me the importance of storytelling in academic writing. Your guidance, marked by your generous time and detailed feedback, has helped me become a more thoughtful writer.

This DPhil journey would not have been the same without the incredible friends I met in the Department of Education. Anki, Casey, Jisoo and Johannes – thank you for the board game nights, college balls and our getaways to the Cotswold and Cornwall. Your friendship has been a constant reminder that I was never alone in this journey. To Cindy, Ivan, and Kate K., thank you for sharing my love of good food and cosy pub evenings. These are memories I will cherish for years to come.

My sincere thanks also go to the ReadOxford and Nation-Wonnacott joint lab group for offering a safe and supportive space to share my half-baked research ideas. To my office mates, Jessie, Mohen, Nicky, Rainy, Sean, Tonia and Yifan, thank you for always giving your time to help me think through problems in my research. You made coming to Anna Watts something I look forward to every day. A special thank you also goes to my friend Stephen, back home in Hong Kong, for always answering my questions about statistical learning at odd hours during the day.

I am also thankful to University College, where I had the opportunity to work as a Junior Dean for the majority of my time at Oxford. Thank you for providing me with a community that truly felt like home. A special thank you to Alicia, whose wisdom and thoughtful advice over countless lunches and formal dinners in the hall meant more than I can express.

I feel incredibly lucky to have a family who has supported me in every academic endeavour I have pursued. I want to give special thanks to my siblings, Sin Yee, Sin Tung (Kathy) and Kai Yuen, whose presence and support have meant more to me than ever during the last four years. Kathy, I am especially grateful for the comfort of having you so close by during my DPhil. You welcomed me into your home whenever I needed it, and made sure that we celebrated every achievement and important festival together, even while far from home.

I cannot close this acknowledgement without expressing my love and deepest gratitude to my partner, Rory. From my master's degree through to this doctorate, you have believed in me more than I sometimes believed in myself. Thank you for your endless patience and kindness in hearing every idea and frustration I have had along the way. And to Anne and Shaun, thank you for including me in all the holiday celebrations and for your support throughout this journey.

Abstract

Written languages are complex. English orthography not only reflects phoneme-grapheme mappings, but it also includes other regularities related to form-meaning mappings and graphotactics. Pseudoword experiments have shown that children and adults are sensitive to these patterns, which can influence their spelling choices even when they have not been explicitly taught the patterns or are unable to verbalise them. From this perspective, spelling acquisition can be viewed as a process of statistical learning, in which individuals become increasingly sensitive to regularities and quasi-regularities in their writing system as they gain more experience with written language. Using the artificial orthography learning paradigm, the overarching aim of this thesis is to examine whether adult spellers can simultaneously learn semantic regularities (i.e., spelling patterns that carry meaningful information) and graphotactic regularities (i.e., spelling patterns concerning permissible grapheme combinations) in new writing systems through exposure, and how this evidence informs our understanding of statistical learning in reading and spelling acquisition. Three studies were conducted to address these questions.

Study 1 presents a systematic review that describes existing artificial orthography learning experiments and examines how this body of evidence informs our understanding of statistical learning in reading and spelling acquisition. Among other findings, this study showed that most experiments focused on orthography-phonology mappings with native English-speaking participants. Evidence from orthotactic learning experiments suggests that people can become sensitive to statistical regularities after brief periods of exposure. However, this paradigm faces several limitations, particularly concerning ecological validity and the comparability of findings across experiments.

Study 2 reports two artificial orthography learning experiments (Experiments 1 & 2) investigating whether native English-speaking adults could simultaneously learn both semantic (where possible spellings depend on grammatical word class) and graphotactic regularities (where possible spellings depend on earlier graphemes) in an artificial lexicon. Participants showed successful generalisation of semantic patterns at post-test, but there was no conclusive evidence of graphotactic learning. Learning was stronger when the semantic patterns were assigned to nouns and verbs compared to adjectives and adverbs. This learning was also associated with participants' ability to verbalise the patterns at post-test.

Study 3 includes two additional artificial orthography learning experiments examining whether participants could learn graphotactic patterns modelled on French noun pluralisation patterns and whether semantic cues facilitated this learning. The artificial orthography was presented in Latin alphabets (Experiment 3) and symbols (BACS-2 fonts; Experiment 4) to further examine whether unintended phonological cues impacted graphotactic learning. In both experiments, participants demonstrated graphotactic learning after brief exposure, although learning was stronger when Latin alphabets were used. Consistent with findings from Study 2, participants who could verbalise the patterns at post-test showed better generalisation. Contrary to predictions, there was limited evidence that semantic cues facilitated graphotactic learning.

Combining a systematic review of existing evidence with experimental work, this thesis provides new insights into the learning mechanisms underlying reading and spelling acquisition. This research addresses a knowledge gap in understanding how learners acquire regularities beyond orthography-phonology mappings in written language. It also holds

implications for theories of literacy development, spelling instruction and the use of artificial orthography in psycholinguistic research.

Table of Contents

List of Figures

List of Tables

Chapter 1. General Introduction

Learning to spell in English is often considered challenging. While some phonemes are consistently represented by a single grapheme (e.g., the /p/ sound in “pen”, “pot”, “pop”), other phonemes can correspond to several possible graphemes and are sensitive to contexts. For example, the /k/ sound is usually spelled as the letter “c” when followed by the vowels “a”, “o” or “u” (e.g., “carrot”), but it is more likely to be spelled as the letter “k” when followed by “e”, “i" or “y” (e.g., “kettle”). These inconsistencies in the mappings between sound and spelling have led some to describe the English writing system as “chaotic and unprincipled” (Dewey, 1971, p.4).

Apart from phoneme-grapheme mappings, research has shown that writing systems such as English and French contain other levels of consistency, which can provide useful cues to spellers (Deacon et al., 2008). One type of regularity is graphotactics[1], which concerns the occurrence, positioning and sequencing of graphemes. These patterns can be conditioned by phonology. For example, consonants tend to be extended following short vowels (e.g., “carrot” /kærət/ but not *carot). They can also be independent of phonology. For instance, consonant doublets in English can appear at the end of words but not at the beginning (e.g., “pill”, but not *ppil). Some vowels can form doublets in the word-medial position while others cannot (e.g., “meet”, but not *miit). In addition, English spellings reflect the relationship between spelling and meaning. Berg & Aronoff (2017) demonstrated in a corpus

[1] In this thesis, the terms “orthotactics” and “graphotactics” are both used to refer to regularities governing the occurrence, positioning, and sequencing of graphemes. The choice of terms varies across studies to align with their respective theoretical and empirical frameworks. Specifically, “orthotactics” was used in Study 1 because the review was framed within the connectionist ‘triangle’ model (Seidenberg & McClelland, 1989), where the relevant component is termed “orthography.” This term was chosen to better align with the model’s terminology. In contrast, “graphotactics” was used in my experimental work in Studies 2 and 3, as it reflects the more commonly used terminology in recent empirical research (e.g., Singh et al., 2021).

analysis that the spellings of some English suffixes provide consistent markers of lexical categories. For instance, the suffix spelling -ous as in “nervous” and “adventurous” is virtually always associated with an adjective but not with other lexical categories.

Given the relationships between spelling and sound, spelling and meaning as well as graphotactic constraints, it is clear that the English writing system is in fact rich with multiple types of regularity. To explain how such regularities are learned, the connectionist ‘triangle’ model (Harm & Seidenberg, 2004; Plaut et al., 1996; Seidenberg & McClelland, 1989) proposes that the mental representation of a lexicon consists of three structural elements including orthography (spelling), phonology (sound) and semantics (meaning). These elements are linked by weighted connections, which modulate the flow of activation. With each exposure to a word, these weights are adjusted incrementally and over time, this supports accurate reading and spelling as the model learns. From this perspective, spelling acquisition could be considered a statistical learning process in which individuals become increasingly sensitive to regularities and quasi-regularities in their writing system as they gain more exposure and experience with written language. It is important to note that the definition of statistical learning varies across the literature. In this thesis, statistical learning is broadly defined as the discovery of patterns in the input (Romberg & Saffran, 2010). This definition captures an individual’s ability to extract regularities from exposure without explicit instructions directly on the target patterns themselves, and it could encompass both supervised and unsupervised learning.

Experimental research using pseudoword spelling/choice tasks has demonstrated that both children and adults are indeed sensitive to orthography-phonology mappings (e.g., Schmalz et al., 2020; Treiman & Kessler, 2006), orthography-semantics mappings (e.g., Ulicheva et al., 2020, 2021) and graphotactic patterns (e.g., Cassar & Treiman, 1997; Hayes et al., 2006) in

their writing system (see the Introduction sections of Chapters 2-4 for detailed discussions). For example, Cassar and Treiman (1997) found that even first-grade children could tell apart permissible and impermissible consonant doublets in English to a certain extent (e.g., “baff” is correct but not *bbaf), despite not receiving explicit instruction on these patterns. These findings support the view that statistical learning processes underlie orthographic learning. However, using natural language patterns to examine this sensitivity has its limitations. Participants differ in their pre-existing knowledge about the orthography (Schmalz et al., 2021) as they may have varying levels of previous exposure to specific patterns and different types of language instructions in school. Furthermore, given the complexity of natural language and multiple correlated features, it is difficult to isolate and examine one particular feature, making it challenging to determine precisely what aspects of the patterns people are sensitive to.

To address these limitations, researchers have turned to the artificial orthography learning paradigm. In these experiments, participants are exposed to researcher-induced artificial orthographic patterns, and their knowledge of these patterns is then tested. This approach has considerable merit, as experimenters have complete control over the input statistics and level of exposure, which allows them to track learning and generalisations as a function of different manipulations. To date, this paradigm has been used to examine both reading and spelling acquisition, focusing on different orthographic patterns including orthography-phonology mappings (e.g., Taylor et al., 2011), orthography-semantic mappings (e.g., Rastle et al., 2021) and graphotactic patterns (e.g., Samara & Caravolas, 2014). The majority of studies have focused on orthography-phonology mappings in the context of reading acquisition. Much less is known about how learners acquire other types of regularities in the writing system, such as orthography-semantic mappings and graphotactic regularities, which

are key to spelling acquisition. As a result, our understanding of how statistical learning supports spelling acquisition remains limited. This thesis addresses these gaps by using the artificial orthography learning paradigm to investigate two central research questions:

(1) Can native English-speaking adults simultaneously learn semantic and graphotactic regularities in written language through exposure?
(2) How does evidence from artificial orthography learning experiments inform our understanding of statistical learning in reading and spelling acquisition?

Methodology

This thesis includes three studies to address the central research questions: a systematic review (Study 1) and four artificial orthography learning experiments with native English-speaking adults with different experimental manipulations (Study 2: Experiments 1 & 2 and Study 3: Experiments 3 & 4). The following section provides an overview of each study to highlight its contributions to this thesis.

Study 1: Systematic review on artificial orthography learning experiments

Study 1 (Chapter 2) presents a systematic review which was conducted to describe the body of literature that has used artificial orthography learning experiments to examine reading and spelling acquisition. While the central focus of this thesis is on spelling acquisition, reading is also considered in this review to provide a more comprehensive understanding of how this paradigm has been used to examine orthographic learning across both domains.

This review had two primary aims: (1) to document and describe original research investigating reading and spelling acquisition through the artificial orthography learning

paradigm and (2) to examine how this paradigm informs our understanding of statistical learning in reading and spelling acquisition. To address the first aim and following a systematic search of the literature, the review first outlined the key characteristics of the existing evidence base, and then examined the linguistic and non-linguistic external factors that have been studied for their effects on orthographic learning. For the second aim, it identified experiments that met the definition of statistical learning adopted in this thesis and examined the relationship between learning outcomes and individual differences in reading, spelling and other related abilities. Focusing specifically on orthotactic learning experiments, I also reviewed how the design of exposure and testing phases varied across experiments and reflected on the implications of this for learning.

A systematic search was conducted using a search string developed to capture two key concepts: artificial orthography and literacy development. The final search was conducted on six databases including APA PsycInfo, ProQuest Social Science Premium Collection (including Education and Linguistics), British Education Index EBSCO, Web of Science, Scopus and ProQuest Dissertation & Theses Global. Through this search, all experimental research published in English in which participants learned an artificial orthography[2] in written form and were tested on their knowledge of these artificial orthographic patterns afterwards was identified. The final sample included 114 unique experiments. The detailed methods for this systematic review were pre-registered on IDESR (https://idesr.org/article/IDESR000101).

[2] In this review, "artificial orthography" was used as a broad term to describe orthographic patterns that were induced by researchers. Different terms may be used across studies, including artificial orthography, artificial script and artificial lexicon.

Study 2: Simultaneous learning of semantic and graphotactic regularities

One key finding from Study 1 is that existing research in the artificial orthography learning paradigm focuses heavily on orthography-phonology mappings. Much less attention has been given to other regularities such as form-meaning mappings and graphotactic patterns, or to the question of learning multiple regularities at the same time. Therefore, two artificial orthography learning experiments were conducted in Study 2 (Chapter 3, Experiments 1 & 2) to investigate whether native English-speaking adults could simultaneously learn both semantic regularities (when possible spellings depend on grammatical word class) and graphotactic regularities (where possible spellings depend on earlier graphemes) in a new writing system.

In Experiment 1, participants were exposed to semantic and graphotactic patterns in an artificial lexicon created from a semi-artificial set of graphemes through a reading task. After the exposure phase, their knowledge of semantic and graphotactic patterns was assessed in a fill-in-the-blank generalisation task. Participants also completed a word category task designed to assess their ability to recall the intended word class of the artificial words. Following Singh et al. (2021), this experiment also included (1) an English spelling task to look for correlations in post-test performance and real-word spelling ability and (2) a post-experiment questionnaire designed to probe whether participants could verbalise the spelling rules and how that related to their performance in the post-tests. Experiment 2 replicated this experiment (Version 1) and extended it (Version 2) by counterbalancing the assignment of lexical categories to pattern types. In Experiment 1, artificial words that were nouns and verbs embedded semantic patterns whereas adjectives and adverbs embedded graphotactic patterns. In Experiment 2 (Version 2), the design was reversed: adjectives and adverbs embedded semantic patterns and nouns and verbs embedded graphotactic patterns.

Comparing results from the two versions allowed us to examine the effects of lexical categories on learning. Bayes Factor analysis was also introduced as the primary method of analysis in Experiment 2. The detailed methods and analysis plan for Experiment 2 were pre-registered on OSF (https://osf.io/j3ubk/).

Results from these experiments showed that native English-speaking adults were able to learn semantic patterns after a short exposure to the artificial lexicon without any explicit instruction on the regularities embedded in the artificial words. Contrary to previous research, there was no evidence of graphotactic learning. This suggests that learning two different types of regularity in an artificial writing system through short-term exposure is challenging. Given the absence of graphotactic learning in this study, Study 3 continued to examine graphotactic learning and investigate the impact of semantic cues on this learning.

Study 3: The impact of semantic and phonological cues on graphotactic learning

Study 3 (Chapter 4) comprises two artificial orthography learning experiments (Experiments 3 & 4) designed to investigate whether native English-speaking adults could learn graphotactic patterns from brief exposure to words embedding those patterns, and whether semantic cues facilitated this learning. The artificial orthography was modelled on the probabilistic nature of French noun pluralisation, where the plural suffix choice is governed by graphotactic constraints (e.g., most nouns take -s but nouns ending in -eau must take -x) and varies in frequency (with -s occurring more often than -x). Importantly, these plural suffixes are not usually pronounced and must be learned in their written form.

This study included three key manipulations. The first was a within-participant manipulation based on the frequency and consistency of mappings between noun endings and suffixes in the graphotactic constraints. Some mappings appeared more frequently than others in the exposure phase (e.g., the pairing of the noun ending -a with the suffix -x occurred 48 times, while the pairing of -i and -v appeared only 24 times). These mappings could also be consistent (e.g., the noun ending -a can only take the suffix -x) and inconsistent (e.g., the noun ending -o can take either -k or -d as its suffix). The second was a between-participants manipulation examining the impact of semantic cues on graphotactic learning. In each experiment, one group of participants was exposed to the graphotactic constraints embedded in artificial words with no reference (GR group), while another group saw referents alongside the written phrases, making it clear that the endings being learned were plural morphemes (SE group). The third was another between-subject manipulation exploring the impact of phonological cues on this learning. The artificial orthography was presented either in the Latin alphabet (Experiment 3) or via symbols (BACS-2 fonts; Experiment 4), with the latter designed to eliminate any unintended phonological cues.

Experimental procedures were similar to those used in Experiments 1 and 2. Participants were exposed to graphotactic patterns embedded in artificial singular-plural noun phrases written in the Latin alphabet (Experiment 3) or symbols (BACS-2 fonts; Experiment 4). After the exposure phase, participants completed a series of post-tests including a visual recognition task, an auditory recognition task (Experiment 3 only), a fill-in-the-blank generalisation task, a word meaning task and an awareness questionnaire. Both experiments were pre-registered on OSF (Experiment 3: https://osf.io/7qad4; Experiment 4: https://osf.io/yf39j).

In summary, this thesis combines a systematic review of existing evidence from the artificial orthography learning paradigm with findings from four learning experiments to provide insights into the learning mechanisms underlying reading and spelling acquisition. The next chapter presents the findings from the systematic review (Study 1), which informed the design of the subsequent experimental studies in this thesis.

Chapter 2. Systematic Review on Artificial Orthography Learning Experiments (Study 1)

Abstract

Artificial orthographic learning experiments have gained popularity in reading and spelling research as they allow researchers to isolate specific orthographic patterns or factors for investigation and examine learning mechanisms that are otherwise difficult to study in natural language contexts. This systematic review outlined the key characteristics of existing experiments, and the factors explored for their impacts on orthographic learning. It also synthesised findings to assess how this paradigm informs our understanding of statistical learning in reading and spelling acquisition. Our findings show that evidence from this paradigm remains limited, with most experiments focusing on orthography-phonology mappings in native English-speaking adults. Nonetheless, this paradigm offers valuable insights into factors influencing orthographic learning, including linguistic factors (e.g., consistency, frequency) as well as non-linguistic external factors (e.g., explicit instruction). In addition, orthotactic learning experiments show that individuals can quickly develop sensitivity to statistical regularities in novel written language after brief exposure. These findings support the view that statistical learning processes contribute to orthographic learning under real-world conditions. However, concerns about the validity and reliability of existing methods limit the strength of these conclusions. This review highlights the need for future research to examine orthographic patterns beyond orthography-phonology mappings and to adopt more consistent approaches to testing and measuring learning.

Introduction

Written language is complex as it contains regularities and quasi-regularities that operate at multiple levels. Therefore, learning spelling patterns in English goes beyond recognising that the word “cat” (/kæt/) can be broken down into three phonemes, each of which maps to a grapheme. It also includes understanding that the letter “c” is pronounced as /k/ before the vowels “a”, “o” and “u” (e.g., “cap”, “cot”, “cup”), but as /s/ before “e”, “i” and “y” (e.g., “cent”, “cider”, “cycle”). Additionally, English spelling patterns reflect the relationship between spelling and meaning. For example, the phonological sequence /əs/ is typically spelled as -ous in adjectival contexts (e.g., “nervous”, “adventurous”) but other spellings are used in non-adjectival contexts (e.g., “bonus”, “cactus”). Another level of consistency in English spelling concerns the orthotactic regularities that govern permissible letter sequence, patterns and positions in written words (Apel et al., 2006; Pacton et al., 2005). For example, a consonant is often doubled following a vowel spelled with a single letter but remains a single consonant after a vowel spelled with multiple letters (e.g., “ball” vs. *bal). Given the abundance of such patterns in writing systems, learning to read and spell can be considered a process where individuals become increasingly sensitive to the regularities and quasi-regularities of their writing systems as their experience with written language grows over time.

Statistical learning in reading and spelling acquisition

Seidenberg and McClelland’s (1989) connectionist model offers an explanation for how regularities and quasi-regularities are learned in written language. According to their triangle model, the mental representation of a lexicon consists of three structural elements including orthography (spelling), phonology (sound) and semantics (meaning) (see also Harm &

Seidenberg, 2004; Plaut et al., 1996). These elements are linked by weighted connections, often referred to as the hidden layers, which modulate the flow of activation. These weights are adjusted incrementally with each exposure to a word. Consider the English spelling pattern -ave, which is typically associated with the pronunciation /eɪv/, as in “save”, “pave” and “brave”. Through repeated exposure to these words, the model forms a stable mapping from the spelling -ave to the pronunciation /eɪv/. Consequently, when encountering novel words such as “mave”, it can generalise the likely pronunciation /meɪv/ based on the learned mapping. However, words such as “have” /hæv/ deviate from this pattern. As a high-frequency word, frequent exposure to the word “have” allows the model to adjust the weights in the connections to enable accurate pronunciation.

The connectionist model assumes that correct pronunciation, whether for regular or irregular words, depends on establishing an appropriate set of weights among orthography, phonology and semantics. From this perspective, learning to read and spell can be understood as a form of statistical learning (Seidenberg, 2005). That is, through repeated exposure, individuals extract statistical patterns in the relationships between orthography and phonology, orthography and semantics, as well as orthotactic regularities. This account is supported by evidence from pseudoword experiments (e.g., Treiman et al., 2021; Treiman & Kessler, 2006; Treiman & Wolter, 2018). For example, Treiman and Wolter (2018) used a pseudoword spelling task to examine consonant doubling in native English-speaking adults. They found that participants’ decisions about doubling consonants were impacted by both the quality (short vs. long) and spelling (single vs. multiple letters) of the preceding vowel. Importantly, a post-experiment questionnaire revealed that most participants were unaware of the relationship between consonant doubling and the quality/spelling of the preceding vowels,

suggesting that their sensitivity to these spelling patterns was likely developed through exposure to printed words.

Although some experimental research supports the view that learning to read and spell is a process of statistical learning, it is important to note that definitions of statistical learning vary in the current literature. Broadly speaking, statistical learning refers to the discovery of patterns in the input (Romberg and Saffra, 2010). However, some research emphasises that statistical learning occurs incidentally and that individuals do not have any conscious intention to learn (Turk-Browne et al., 2009). In this view, learning is a by-product of exposure, resulting in minimal explicit knowledge of the underlying statistical structure of the stimuli (e.g., Saffran et al., 1997; Turk-Browne et al., 2005). As the current review aims to provide a comprehensive overview of the evidence for statistical learning within the artificial orthography learning paradigm, we adopt a broad definition of the term. Specifically, we define statistical learning as the discovery of patterns in the input (Romberg & Saffran, 2010) to capture an individual's ability to extract regularities from exposure without explicit instructions directly on the target patterns themselves. This definition spans a range of learning approaches from supervised to unsupervised learning.

Examining reading and spelling acquisition through artificial orthographies

One approach to examining reading and spelling acquisition is the artificial orthography learning experiment. Typically, participants in these experiments are given an exposure phase during which they encounter novel orthographic patterns that do not exist in their native language. Participants are then tested on their knowledge of these patterns, and their ability to generalise to novel items written in the novel orthography. This paradigm has been used to examine different types of stimuli and the impacts of factors such as instruction methods

(e.g., Taylor et al., 2017), sleep deprivation (Tamminen et al., 2020) and non-invasive brain stimulation (e.g., Thakkar et al., 2020) on learning. Researchers have also explored the similarities and differences between typical and dyslexic readers when it comes to learning a novel orthography (e.g., Tong et al., 2020).

There are several key reasons why the artificial orthography learning paradigm is useful. First, it allows researchers to have complete control over the precise nature of the writing system, and to isolate a particular orthographic pattern or factor for investigation. Second, introducing novel orthographic patterns ensures that any between-condition differences in learning outcomes are not due to differences in prior knowledge. Finally, the paradigm provides an opportunity to capture learning and in doing so, to shed light on the mechanisms that underly reading and spelling acquisition. While longitudinal studies on reading and spelling acquisition are possible, they are logistically challenging. By focusing on a limited set of orthographic patterns or factors within the context of a single experiment, researchers can study acquisition in a miniature learning environment.

Importantly, however, while the artificial orthography learning paradigm has undoubtedly opened new possibilities, several issues have emerged. One major issue is that the evidence produced within this paradigm varied substantially in its scope and focus. This is partly due to the flexibility of the paradigm, as it allows experimenters to isolate their examination to a particular orthographic pattern. For example, while both Chetail (2017) and Singh et al. (2021) examined the learning of orthotactic regularities, the former study did so with bigram frequencies whereas the latter focused on consonant doubling. Although having different research foci is not inherently problematic, it complicates comparisons across experiments,

making it more challenging to draw broader conclusions about reading and spelling acquisition.

Another issue concerns how evidence from this paradigm contributes to our understanding of statistical learning in 'real' reading and spelling acquisition. Experimental setups within this paradigm are very different from experiencing spelling patterns in natural language contexts. Learning to read and spell takes many years throughout which we are exposed to rich statistics in our spoken and written language. How relevant is a lab-based experiment that takes an hour to complete to a lifelong mission of mastering an orthography? One way that existing research has attempted to demonstrate external validity of the paradigm is to examine the relationship between artificial orthography learning experiment outcomes and individual difference measure outcomes (e.g., reading and spelling ability in native language): if artificial orthography learning experiments do indeed tap into orthographic learning in natural language contexts, learning outcomes in these experiments should correlate with reading and spelling ability. However, only a small number of experiments have investigated this type of correlation, and this has not been systematically examined.

Further to the issue of external validity, another challenge in drawing conclusions about statistical learning stems from the variability of experimental designs. As mentioned earlier, definitions of statistical learning vary across the literature. For some, it refers broadly to the extraction of regularities from input, while others focus more narrowly on incidental learning conditions. These differing interpretations influence how learning conditions are designed for a particular set of experiments. Furthermore, there are no standardised methods to test and measure learning of the newly learned orthographic knowledge. Experimenters have used different tasks to measure learning, including legality judgement, fill-in-the-blank, and

wordlikeness tasks, and it is not clear whether these variations tap the same construct. Once again, variability in experimental design raises questions about the comparability of results, potentially affecting our understanding of statistical learning within this paradigm. Given the variability of the evidence base within the artificial orthography learning paradigm, a systematic synthesis of the literature is needed. To the best of our knowledge, Hirshorn and Fiez (2014) is the only review article that has examined the use of artificial orthographies to study reading acquisition. However, this review is somewhat outdated and its scope was limited, as it focused specifically on the impact of grain size (i.e., the unit of spoken language mapped onto a visual graph). Therefore, this systematic review set out to provide a more comprehensive assessment of all available evidence.

Our first aim was to document and describe original research investigating reading and spelling acquisition through artificial orthography learning experiments. To address this, we conducted a systematic search and summarised the characteristics of existing experiments in this paradigm and then categorised the evidence based on factors that may influence orthographic learning. All artificial orthography learning experiments were included for this aim, regardless of whether they met our criteria for statistical learning. Our second aim was to examine how evidence from artificial orthography learning experiments informs our understanding of statistical learning in reading and spelling acquisition. For this, we focused only on experiments that met our criteria for statistical learning, and investigated (a) whether learning outcomes observed in artificial orthography learning experiments relate to individual differences in other behavioural measures and (b) how variations in the experimental designs might impact our understanding of statistical learning from artificial orthographic learning experiments.

Method

The protocol for this systematic review followed the Preferred Reporting Items for Systematic Reviews (PRISMA-P; Page et al., 2021) and was pre-registered on the International Database of Education Systematic Reviews (IDESR) in September 2023. The registration number is IDESR000101 (https://idesr.org/article/IDESR000101).

Information sources and search strategy

We created a list of databases based on Li and Wang's (2023) systematic review of orthographic learning via self-teaching. We further refined this list in consultation with an experienced university librarian. Our final selection included six databases: APA PsycInfo, ProQuest Social Science Premium Collection (including Education and Linguistics), British Education Index (EBSCO), Web of Science, Scopus and ProQuest Dissertation & Theses Global (for grey literature). Together, these resources capture psychology, education, multidisciplinary research and the grey literature.

Table 2.1 shows the search terms and truncations derived from two concepts. Our search string was developed under two broad concepts: artificial orthography and literacy development. These were identified on the basis of eight index studies that we considered critical to include (highlighted in bold in the reference list). A pilot search included a wide range of labels to capture the concept of artificial orthography (e.g., "visual word*", "visual regular*" and bigram*), resulting in over 10,000 results on Web of Science. This is beyond the scope of manageability considering the timeframe and resources available for this review. Therefore, we removed some labels under the concept of artificial orthography on the basis that the search still returned all eight index studies on both Web of Science and Scopus. An

example of the final search string for ProQuest is: noft("artificial orthograph*" or "artificial script*" or "artificial lexic*" or "artificial language*" or "artificial writ*" or "artificial first language*" or "artificial second language*" or "artificial word*" or "novel orthograph*" or "novel word*" or "letter chunk*" or "graphotactic constraint*" or "graphotactic restrict*" or "graphotactic regular*" or "miniature language*") AND noft (literacy* OR spell* OR read* OR writ* OR semantic* OR meaning* OR morph* OR grapho* OR phon*). The search terms were applied to the title, abstract and keywords in each article.

Table 2.1. Search term truncations derived from the two main concepts

Concept	Truncations
Artificial orthography	artificial orthograph*, artificial script*, artificial lexic*, artificial language*, artificial writ*, artificial first language*, artificial second language*, artificial word*, novel orthograph*, novel word*, letter chunk*, graphotactic constraint*, graphotactic restrict*, graphotactic regular*, miniature language*
Literacy development	literacy*, spell*, read*, writ*, semantic*, meaning*, morph*, grapho*, phon*

Study inclusion and exclusion

The inclusion criteria were intentionally expansive in order to capture all original research that used artificial orthography learning paradigms to examine reading and spelling acquisition: (1) experimental research that involves participants learning an artificial orthography in written form, (2) research that uses behavioural measures to assess learning of the artificial orthographic patterns, (3) studies with a full reference list or sufficient information on references and (4) studies reported in English. We did not exclude studies based on their date of publication, population or publication status to ensure that we collected all available evidence on artificial orthography learning. It is important to note that we operationalised the concept of artificial orthography as orthographic patterns that are induced

by researchers. In other words, if participants were tested on these patterns without an exposure phase, they should perform at chance as their prior language knowledge should have little impact on performance. This concept may be captured under different terms such as “artificial script” and “artificial lexicon” in the literature.

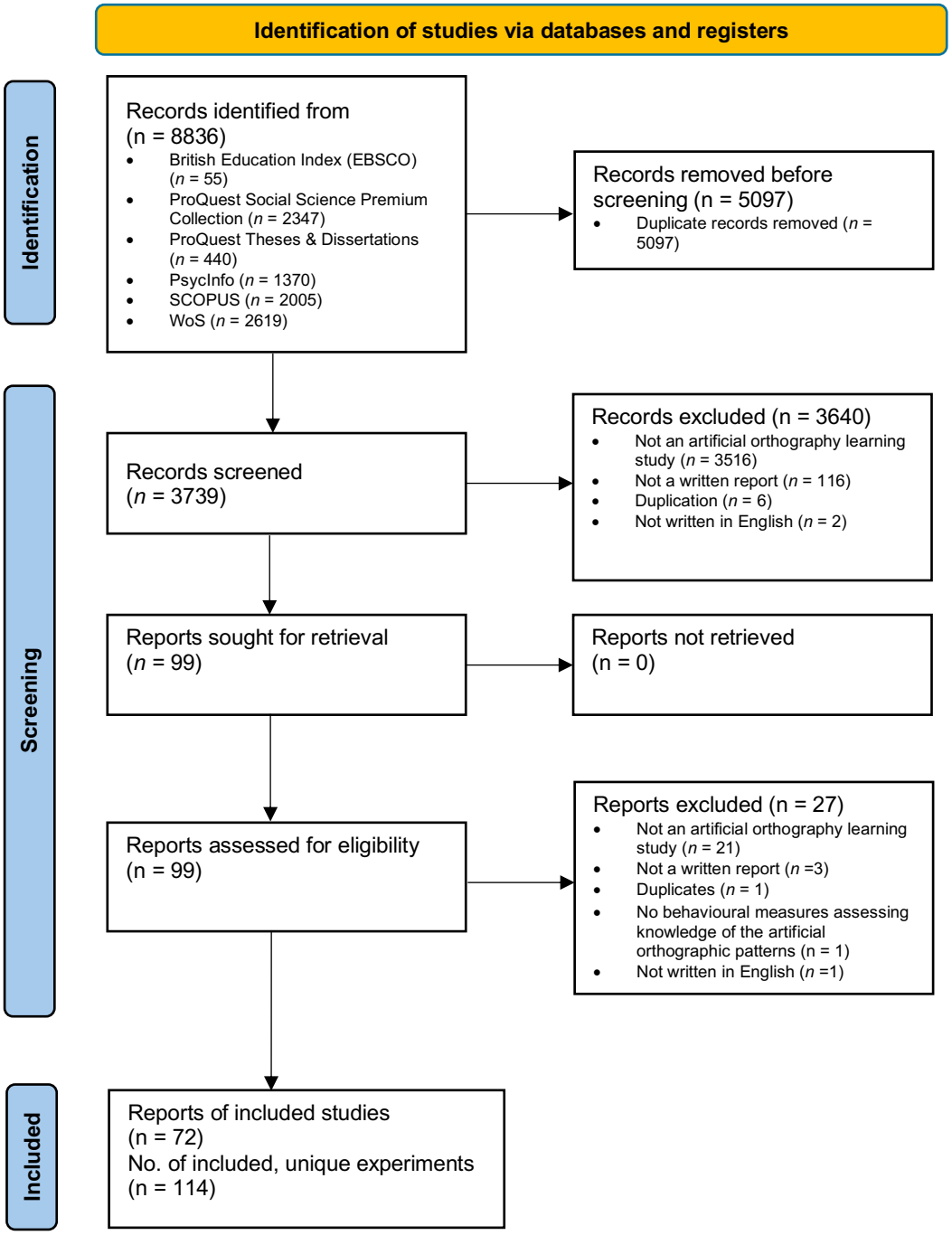
Figure 2.1 shows the flow diagram for the identification of studies through database searches. The final search identified a total of 8836 records. All were uploaded to Zotero for the first round of de-duplication. The remaining records were then uploaded to Rayyan software (Ouzzani et al., 2016) for a second round of de-duplication[3]. These steps removed 5097 records. The first author then screened the titles and abstracts of the remaining 3739 records using the inclusion and exclusion criteria stated above. The second author, who was blind to the first author’s decisions, randomly selected 10% of the abstracts and assessed them independently. The inter-rater reliability from the dual-reviewer blind screening process was strong (Cohen’s *kappa* = .71). Despite the high reliability, we noted that the seven records in conflict were all excluded by one reviewer but included by the other. To ensure no relevant studies were omitted, we decided to retain these seven records at this stage. The title and abstract screening process resulted in 3640 records for four reasons: (1) not an artificial orthography learning study (N = 3516), (2) not a written report (N = 116; e.g., datasets), (3) duplications not identified in the earlier stages (N = 6) and (4) not written in English (N = 2).

All 99 studies eligible for full-text screening were retrieved. During the full-text screening process, the first author conducted full-text screening by reading each study and following the

[3] We conducted the de-duplication process on both Zotero and Rayyan because there were over 8000 records from our final search. Zotero was useful in identifying most duplications, but we could not remove duplications that were labelled as different publication types on Zotero. These were identified and removed later in Rayyan.

inclusion/exclusion criteria. The second author randomly selected 10% of these 99 studies for independent assessment. There was good agreement (Cohen's *kappa* = .61) between the two raters in including/excluding the studies. After the full-text screening process, 27 studies were removed based on the following reasons: (1) not an artificial orthography learning study (N = 21), (2) not a written report (N = 3), (3) duplications not identified in the earlier stages (N = 1), (4) no behavioural measures assessing knowledge of the artificial orthographic patterns (N = 1) and (5) not written in English (N = 1). Within the remaining 72 studies, 118 individual experiments met the inclusion criteria. Four experiments (Experiments 3 to 5 in Singh, 2021 and Experiment 5 from Taylor, 2010) were further excluded due to duplication between theses and subsequently published journal articles also included in this review. The final sample, therefore, included 72 studies (marked with an asterisk in the reference list) and a total of 114 unique experiments.

Figure 2.1. PRISMA 2020 flow diagram



Data extraction

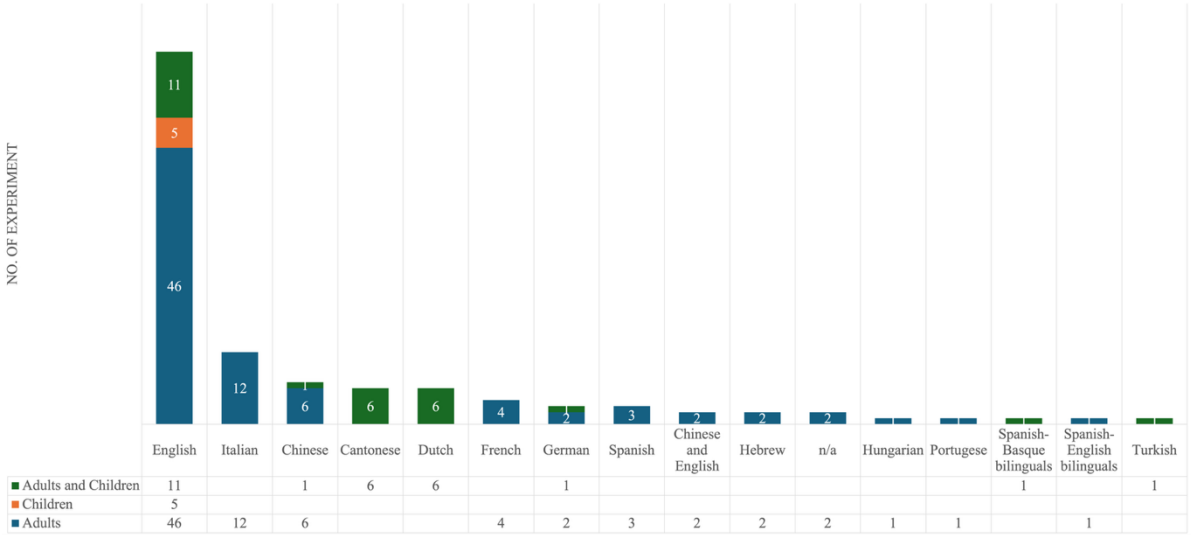
A data extraction form was created based on other published systematic reviews in applied linguistics (Chan et al., 2022; Chong & Reinders, 2025; Schulz et al., 2023) and suggestions from the Cochrane good practice data extraction form (Cochrane Effective Practice and Organisation of Care (EPOC), 2017). The data extraction form was piloted by the first and second authors using experiments reported by Singh et al. (2021), one of our index studies. Following this and for each included study, the first author extracted and entered information for these seven items: (1) general information (e.g., corresponding author's contact details, publication type and year of publication), (2) sample demographics (e.g., age, gender and L1 background), (3) study overview (e.g., research question, study duration and number of participants), (4) experiment details (e.g., details of the artificial orthography, description of task details and orders and descriptions of any other behavioural tasks capturing individual differences), (5) analyses (e.g., type and details of statistical analyses used), (6) outcome (e.g., outcome type, outcome definition and descriptive outcome) and (7) other information (e.g., key conclusions of the study, limitations noted by the authors and suggestions on future research directions).

Results

1. What are the characteristics of the artificial orthography learning experiments?

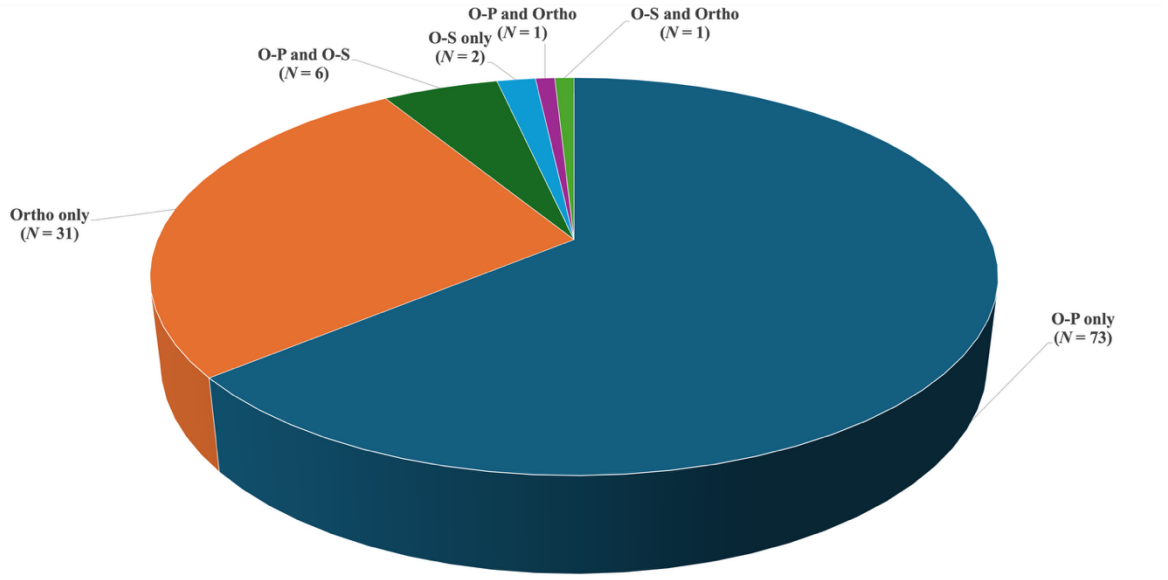
We begin by describing the characteristics of the 114 experiments selected for analysis, regardless of whether they met the criteria for statistical learning. Detailed descriptives for each individual experiment are provided in Appendix 2A.

Figure 2.2. Counts of experiments by participants' native language background and age groups



As illustrated in Figure 2.2, over half of the experiments (*N* = 62; 54%) collected data from participants with English as their native language. Most of the remaining experiments focused on speakers of other alphabetic languages including Italian, Dutch, French, German and Spanish. Only a small number (*N* = 15; 13%) included speakers of non-alphabetic languages, exclusively Mandarin Chinese or Cantonese. Most experiments were with adults (*N* = 82, over 70%); 24% were with both adult and child samples (N = 27), and five experiments (4%) used only a child sample. Finally, most of the experiments (90%) were with healthy adults or typically developing children; only 11 experiments (10%) included participants with spelling disabilities or dyslexia or with poor lexical skills (assessed by orthographic and phonological processing) or poor reading comprehension.

Figure 2.3. Types of artificial orthographic patterns that participants learned in the experiment



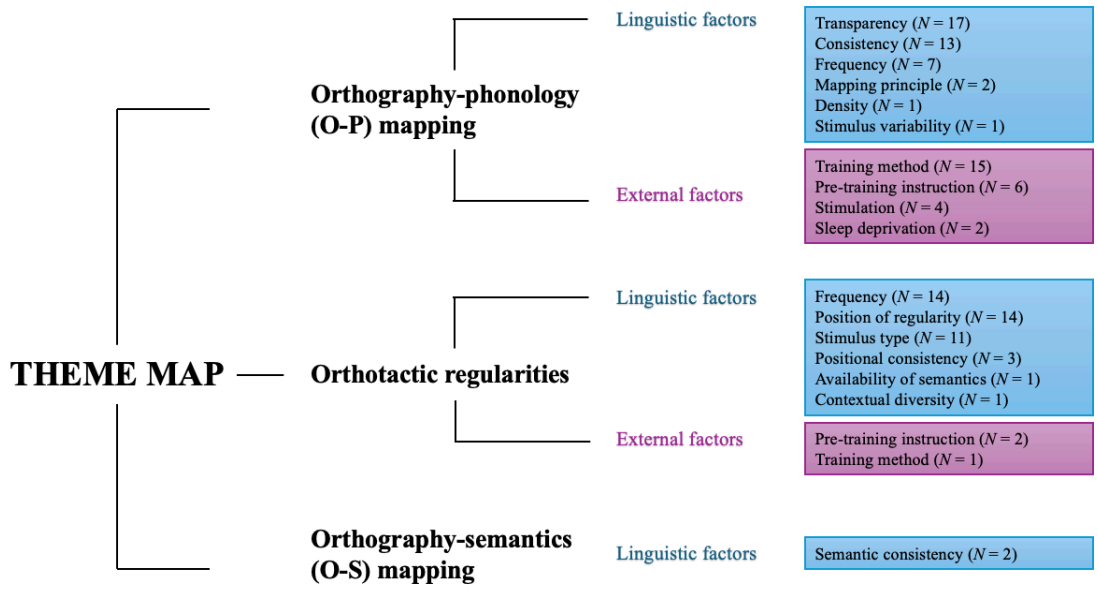
As introduced earlier, orthographic learning involves learning the statistical patterns underlying the relationships between orthography and phonology, orthography and semantics, as well as orthotactic regularities. Further details are provided in subsequent sections but for present purposes, we categorised each experiment based on the types of orthographic patterns participants were tasked with learning, as summarised in Figure 2.3. Most experiments included only one type of orthographic pattern, with the majority focusing on orthography-phonology (O-P) mappings (N = 73; 64%), followed by orthotactic regularities (Ortho; N = 31; 22%). Only two experiments focused exclusively on orthography-semantics (O-S) mappings. A small number of experiments included two types of pattern within the same experiment. This includes six experiments with both O-P and O-S mappings, one experiment with O-P mappings and orthotactic regularities, and one that included both O-S mappings and orthotactic regularities. An important finding is that participants in most experiments learned the artificial orthographic patterns in meaningless contexts. Only 16 experiments (14 on O-P mappings and two on orthotactic regularities)

required participants to also learn word meaning alongside the artificial orthographic patterns. Note that this differs from O-S mappings which concern how parts of the word correspond to aspects of its meaning.

2. What factors may impact orthographic learning?

Having classified the types of orthographic pattern examined in the 114 experiments, we next turned to examine the factors that may impact orthographic learning within this body of evidence. Two sets of factors, termed here linguistic factors and non-linguistic external factors, emerged across experiments within each type of orthographic pattern, as summarised in Figure 2.4. The following section is organised according to each type of orthographic patterns: orthography-phonology (O-P) mappings, orthotactic regularities and orthography-semantics (O-S) mappings.

Figure 2.4. Concept map of key factors studied for their impact on orthographic learning within the artificial orthography learning paradigm



2.1 Orthography-phonology (O-P) mappings: Linguistic factors

Among artificial orthographic learning experiments where participants were tasked with learning O-P mappings, research has examined six key factors related to linguistic input: transparency, consistency, frequency, mapping principles, density and stimulus variability.

2.1.1 Transparency

Transparency refers to the degree to which an orthography reflects phonology in a writing system (Katz & Frost, 1992). This concept is closely related to the consistency of an orthography because the more consistent the grapheme-phoneme mappings are, the more transparent the orthography is. In this review, we differentiate transparency from consistency and focus on experiments that compared the learning of transparent vs. non-transparent scripts within their experimental design. We identified seventeen experiments that contrasted learning transparent vs. non-transparent artificial orthographies.

Of these seventeen experiments, some (Dong et al., 2022; Mei et al., 2013, 2014, 2015; Wei et al., 2015; Xue et al., 2017) compared learning of transparent and non-transparent orthographies using artificial words written in Korean Hangul letters, with which participants had no prior experience. The key distinction between the two orthographies is that the grapheme-phoneme correspondences (GPCs) were systematic in the transparent orthography, but arbitrary in the non-transparent orthography (i.e., participants had to learn to read them as whole words). Using a between-participant design, Mei et al. (2014, 2015) found that both native English- and Chinese-speaking adult participants trained on the transparent orthography were better at reading trained words at post-test. Similar results were reported by Wei et al. (2015) who found a faster learning rate for those trained in a transparent orthography group. Dong et al. (2022) later adapted this experimental setup into a shorter,

within-participant design. Their findings aligned in showing that artificial words written in a transparent script were named more easily than those with arbitrary mappings. Collectively, these findings demonstrate a learning advantage for transparent orthographies over non-transparent ones.

However, some evidence suggests that this learning advantage may depend on the experimental design. For example, Mei et al. (2013) did not find any significant differences between groups of participants trained on transparent and non-transparent orthographies, even though they used the same training materials and procedures as later studies which did find a transparency advantage (Mei et al., 2014, 2015). One explanation for the discrepancy is the design of the naming task. In Mei et al. (2014, 2015), words in the naming task were presented for 1000 ms whereas the stimulus duration was 3000 ms in the earlier experiment. Additionally, the naming task used by Mei et al. (2014, 2015) included trained and untrained artificial words alongside English/Chinese real words and English pseudowords. These differences may have influenced the observed outcomes.

Another factor that may moderate the learning advantage for transparent orthographies is the similarity between the artificial orthographies and natural orthographies. In Yoncheva et al. (2015), participants were trained on both a transparent script (with consistent GPCs) and a whole-word script (with arbitrary GPCs) using letter-like figures created from line drawings. Following training, participants completed a reading verification task where they decided whether the visual form of the word matched the auditory input. Results showed that participants were more accurate and faster if the item had been trained under a whole-word script, where words were learned as logographs. The systematicity of GPCs in the transparent orthography did not help with their accuracy and speed in this task.

In addition to exploring behavioural differences in learning, several studies have also examined whether variations in orthographic transparency led to different levels of orthographic representation. Bitan et al. (2005) and Bitan and Karni (2003, 2004) trained participants to read artificial scripts made of symbols under three conditions: alphabetic training with whole-word instruction (i.e., training on a transparent script under implicit training), alphabetic training with letter instruction (i.e., training on a transparent script under explicit training) and arbitrary training (i.e., training on a non-transparent script with arbitrary grapheme-phoneme mappings). Instead of focusing on the naming accuracy of the trained items, these studies used transfer tests to examine participants' ability to generalise their learning across different levels (i.e., word, level, symbol or sequence structure). Although Bitan and Karni (2003) used a between-participant design while Bitan and Karni (2004) and Bitan et al. (2005) adopted a within-participant design, the findings were similar: arbitrary training (where participants learned a non-transparent script) results in whole-word knowledge, whereas explicit alphabetic training enabled participants to extract sub-lexical regularities.

2.1.2 Consistency

Consistency, defined as the predictability of a word's spelling-sound mapping, is another important factor that impacts O-P learning. Although consistency is closely related to transparency, it focuses more specifically on the variations in O-P mappings that operate at multiple levels within an orthography. For example, English includes both highly consistent mappings (e.g., the letter "m" is mapped onto /m/ across contexts) and context-conditioned mappings (e.g., the letter "c" can be pronounced as /k/ and /s/ depending on the following vowel context), and the former is easier to learn than the latter. Thirteen artificial orthography experiments have examined the effect of consistency on learning O-P mappings.

In Taylor (2010, Experiments 2-4) and Taylor et al. (2011, Experiments 1 & 2), adult participants learned to read consonant-vowel-consonant-structured artificial words where the grapheme-vowel mappings varied in consistency. In Taylor et al. (2011), for example, a consistent symbol was pronounced /u/ across all contexts whereas an inconsistent symbol was pronounced /əʊ/ when it followed the symbol pronounced /z/ but was pronounced /ɒ/ in all other contexts. When asked to read novel words written in the artificial orthography at post-test, participants were better able to read words that contained a consistent vowel than inconsistent vowels (see also McMillan et al., 2017, for a similar consistency effect with dyslexic readers). Schmalz et al. (2022) adapted the training paradigm to examine the consistency effect with German- and Italian-speaking adults. However, instead of only comparing consistent and context-conditioned mappings, they introduced an unpredictable condition in which a symbol was randomly mapped onto two vowel sounds with equal probabilities. Across three experiments, they found that by the end of training, participants were better at reading trained consistent items than both the trained context-conditioned and unpredictable items. However, contrary to Taylor et al., there was no consistency effect in the generalisation task with novel words, though their lenient scoring system (in which both plausible pronunciations for the context-conditioned mappings were coded as correct) may have reduced the sensitivity for detecting the consistency effect.

Other studies have also failed to observe the consistency effect. Acha et al. (2023) taught children to read artificial words written in symbols that contained regular consonants (i.e., letter “f” is always pronounced as /f/), context-dependent syllables (i.e., letter “c” pronounced /θ/ before “e” or “i” and pronounced /k/ otherwise) or inconsistent syllables (i.e., letters “b” and “v” both pronounced /b/). Post-test results indicated high accuracy in reading aloud and identifying the trained words across all word types with no significant differences. Similarly,

He and Tong (2017) exposed Chinese-speaking children to an artificial script which contained phonetic regularities with three levels of consistency: high (i.e., the phonetic radical corresponded to the same sound 100% of the time), moderate (80%) and low (60%). The results showed no consistency effect on learning, even though children demonstrated above-chance performance in recognition and generalisation (see also Tong et al., 2020, for similar findings with dyslexic readers).

2.1.3 Frequency

Frequency, indexed as the number of times people are exposed to a particular word or pattern, has been examined in seven artificial orthography experiments that focused on learning O-P mappings. In Bolger & Perfetti (2007), participants learned to read spoken English words written in Korean Hangul letters with consistent grapheme-phoneme mappings. They were assigned to one of the two training conditions: holistic training (where participants viewed whole words with their pronunciations in each trial) and componential training (where each component of the Hangul character was highlighted and paired with its phoneme before the whole word and its pronunciation were presented). Using a within-participant design, some artificial words appeared more frequently than others during the exposure phase. Overall, high-frequency items were recognised more quickly than low-frequency items. Participants in the whole-word training group were also more accurate in reading aloud high-frequency trained words.

Taylor (2010, Experiments 2-4) and Taylor et al. (2011, Experiments 1 & 2) also provide evidence for the frequency effect on learning O-P mappings, although the precise patterns varied across experiments and interacted with other factors. Generally, however, the frequency effect interacted with consistency: the frequency effect was only observed for

items that contained inconsistent vowels. McMillan et al. (2017) also observed this interaction in typical readers when using a similar artificial orthography and training paradigm. However, this was not observed in a dyslexic group. Further analyses showed that dyslexic participants were more impacted by the frequency effect than the consistency effect.

2.1.4 Other linguistic factors influencing the learning of O-P mappings

Apart from the above factors, a small number of experiments also examined the impacts of mapping principles, density, and stimulus variability on learning O-P mappings.

Hirshorn et al. (2016) and Martin, Hirshorn, et al. (2019) examined the impact of mapping principles, which concern the unit of spoken language mapped onto a visual graph, on learning O-P mappings. In Hirshorn et al. (2016), participants were assigned to learn either FaceFont (where each face picture maps onto an English phoneme) or Faceabary (where each face picture maps onto a syllable, with face identities representing the consonant component and facial expressions representing the vowel component). Training spanned across two weeks for the FaceFont group and three weeks for the Faceabary group, during which they were trained on phoneme-grapheme/syllable-grapheme mappings, as well as word-level and story-level reading. At post-test, participants read stories in their respective artificial orthography. Results revealed no significant difference between the two groups, suggesting that participants could develop the same extent of reading fluency in both orthographies once they have acquired the grapheme inventory.

To further examine how mapping principles might impact the learning of O-P mappings, Martin, Hirshorn, et al. (2019) provided extended training to a subset of participants who learned HouseFont (Martin, Durisko, et al., 2019; i.e., an orthography with one-to-one

grapheme-phoneme correspondences as FaceFont, but with house pictures) and the Faceabary group (Hirshorn et al., 2016). Over four weeks, participants continued to read stories in their respective orthographies outside the laboratory, and were tested on grapheme knowledge, word reading and passage reading across different time points and at post-test. The findings provided more nuanced evidence of differences in learning O-P mappings between the two mapping principles. The HouseFont group consistently outperformed the Faceabary group in grapheme knowledge and word reading over the course of extended training. In passage reading, although the Faceabary group initially lagged behind the HouseFont group after the initial training, they caught up during the extended training and the two groups showed comparable performance at post-test.

Lally et al. (2020) examined the effect of density (i.e., the number of words that can be created by changing a single letter of a given word) on learning O-P mappings. Participants were trained to read novel words written in an artificial orthography with either a dense or sparse script. The dense script included anagrams, where word pairs differ from each other by switching their initial and final consonants (an English pseudoword example would be "metap" and "petam") or the initial and middle consonant (e.g., "tidan" and "ditan"), while the sparse script did not. There was no reliable difference between the two groups in reading trained words, either during training or post-test. However, at post-test, participants who learned the dense script showed a substantial advantage in terms of generalisation, that is, reading untrained words aloud. They were also less likely to accept transposed-foil items (i.e., "mepat" for the target word "metap") relative to replaced-letter items (e.g., "mekav" for the trained word "metap") in a lexical decision type task, as compared to those who learned the sparse script. These findings suggest that learning a dense script not only resulted in better

knowledge of the underlying GPCs, but participants' processing of the orthography is also more fine-grained such that they were more sensitive to the order of the letters.

Adwan-Mansour and Bitan (2017) is the only experiment examining the effect of stimulus variability (i.e., the number of different words an individual letter appears in) on learning O-P mappings in an artificial orthography. Hebrew-speaking participants learned to read monosyllabic artificial words (e.g., "voj" and "pel") formed from various combinations of six letters ("j", "e", "l", "v", "o" and "p"). Each letter was represented by a pair of symbols in the artificial words. Participants were divided into three groups: a variable group and two non-variable groups. The variable group was trained on 24 unique artificial words, while two non-variable groups were each trained on 12 words only. Those in Group 1 saw the 12 artificial words once, while those in Group 2 saw them twice, thereby equating both the number of exposure per word and the total number of exposure to trained words with the variable group, respectively. By the end of training, all groups achieved over 75% accuracy in recognising whether a Latin-letter string was the correct transcription of a trained artificial word, indicating successful learning of the novel orthography overall. At post-test, the variable group was significantly more accurate in judging whether the artificial words were correctly transcribed into Latin-letter strings than the non-variable Group 2, and effect that was seen for both trained and novel words. Numerically, the variable group also outperformed the non-variable Group 1, however these analyses did not reach statistical significance. These findings suggest that variability in the number of unique training items increases the salience of the underlying GPCs, which allows participants to extract these regularities more easily.

2.2 Orthography-phonology (O-P) mappings: Non-linguistic external factors

In addition to the linguistic factors reviewed above, our review identified several non-linguistic external factors that have been related to learning O-P mappings in artificial orthographic learning experiments.

2.2.1 Training methods

Fifteen experiments have examined the effects of various training methods on learning O-P mappings. They can be grouped into five main types: those exploring effects of (A) explicit instructions on GPCs, (B) pre-exposure to semantic or phonological information, (C) training focus, (D) training order, and (E) handwriting and motor efforts.

2.2.1.A Explicit instructions on GPCs

A group of studies have investigated the effect of a short explicit training session on GPCs before whole-word training (within-participant design: Bitan & Karni, 2003; between-participants design: Bitan et al., 2005; Bitan & Booth, 2012; Bitan & Karni, 2004; Rastle et al., 2021). Generally speaking, experiments compare an explicit instruction condition (i.e., where participants are explicitly taught the symbol(s)-sound mappings first, and then experience words written in the artificial orthography) with a whole-word training condition (i.e., where participants are directly exposed to whole words without any instruction). The benefit of having a short session of explicit instruction is significant across experiments. Explicit training resulted in significantly better knowledge of the GPCs of the trained items (Bitan & Karni, 2003), better long-term retention of letter knowledge (Bitan & Karni, 2004), greater consolidation of letter knowledge between training sessions (Bitan & Booth, 2012) and better generalisation (Rastle et al., 2021).

Bolger and Perfetti (2007) examined the effect of explicit instruction of GPCs *during* the training phase, contrasting holistic and component training methods. In holistic training, participants saw only whole words and their corresponding pronunciations in each trial. For each trial in the component condition, participants were shown how each phoneme mapped onto each part of a Hangul character before seeing and hearing the pronunciation of the whole word. Both groups were able to learn the underlying GPCs, but component training resulted in better generalisation, greater retention of word form knowledge and overall performance that was less affected by frequency.

Aravena et al. (2013) reported similar effects in an experiment where Dutch-speaking children (dyslexic and non-dyslexia) learned mappings between Hebrew graphemes-Dutch phonemes. Participants were assigned to one of three learning conditions: explicit training (where participants were explicitly taught the GPCs), implicit associative learning (where participants learned GPCs by matching speech sounds to the orthographic form in a computer game environment), or a combination of both. Explicit training led to the highest performance in a timed word-reading task where they had to read high-frequency Dutch words written within the artificial orthography. Additionally, the explicit training group showed better learning of the ambiguous orthographic rule in which a grapheme represented either the short vowel /ɑ/ or the long vowel /a/, depending on the consonant that followed. Participants were more likely to apply this rule in the artificial word reading task correctly following explicit training than implicit training, and they were more likely to do so with combined training than with implicit training alone.

2.2.1.B Pre-exposure to semantic or phonological information

Taylor and colleagues provided participants with semantic or phonological information about the to-be-learned words before they were seen in written form. In Taylor (2010, Experiment 4), participants were assigned either to a semantic condition, or one termed lexical phonology. During the pre-exposure phase, both groups listened to and repeated the to-be-learned artificial words but only participants in the semantic group viewed pictures of novel objects associated with each word. Compared to a group that did not receive any pre-exposure (Taylor, 2010, Experiment 3), both pre-exposure groups were better able to read aloud the trained words, and this effect was driven by words that were inconsistent in terms of spelling-sound correspondence (see section 2.1.2). Neither lexical phonology nor semantic pre-exposure significantly improved accuracy in the old-new decision post-test. Both pre-exposure groups performed worse than the no-pre-exposure group in the generalisation task, suggesting that pre-exposure to lexical phonology or semantics may have resulted in stronger representations of the trained items, which in turn could have reduced participants' attention to spelling-sound mappings that are critical to generalisation.

Taylor et al. (2011, Experiment 2) used a within-participant design, where participants learned semantic information (definitions) for half of the to-be-learned artificial words in addition to lexical phonology, and only lexical phonology for the other half of items. By the end of training, words that had semantic pre-exposure were better read than those in the lexical phonology only condition, but only if the items contained low-frequency-inconsistent vowels. Items in the semantic condition were recognised more accurately in the old-new discrimination task, but there was no effect in terms of latency. Generalisation was also poorer after pre-exposure to both semantic information and lexical phonology, as compared to the no pre-exposure condition, especially in items with low-frequency-inconsistent vowels.

Overall, these findings suggest that pre-exposure to the semantic and/or lexical phonology information of the to-be-learned words has a limited facilitative effect on reading.

2.2.1.C Training focus

Taylor et al. (2017) examined the effect of training methods that emphasised orthography-phonology (O-P) and orthography-semantics (O-S) mappings on reading aloud and comprehension of written words. Over eight days, participants learned two artificial languages, performing both O-P tasks (e.g., reading aloud and rhyme judgement) and O-S tasks (e.g., saying the meaning aloud and semantic categorisation) for each language. In the O-P-focused language, O-P tasks were repeated three times each day while O-S tasks were performed once (vice versa for the O-S-focused language). Taylor et al.'s results suggest that O-P-focused training improved both the accuracy and speed in reading aloud whereas O-S-focused training did not enhance reading aloud performance during training. O-P-focused training also resulted in faster responses in the reading aloud generalisation post-test. More importantly, participants in both training methods had comparable performance in saying the meaning of the trained words throughout most of the training.

Zhao et al. (2018) compared learning an artificial orthography through lexical training alone versus a combination of lexical and sub-lexical training. The key difference is that lexical training featured learning tasks that focused on whole words (e.g., judging whether the pronunciation of the target character presented on the screen matched the spoken novel word), while sub-lexical training focused on smaller units of the words (e.g., judging whether the target character shared the same second syllable as a comparison item). Participants learned artificial characters that contained orthography-phonology and orthography-semantics mappings, and were assigned to either a consistent or inconsistent mapping

condition. In the consistent condition, phonetic radicals mapped to the same second syllable across characters (e.g., 小 consistently pronounced as /bi/ in 大小 /gUbi/, 中小 /pobi/), and the semantic radicals mapped onto the same semantic category (e.g., 土 referred to animals, with 土小 representing a dog and 土大 representing a cow). In the inconsistent condition, these mappings are arbitrary. Participants trained on the consistent mappings outperformed those trained on the inconsistent mappings at post-test, regardless of the type of training (lexical or sub-lexical) they received. This suggests that even with training tasks that focused on whole words, participants still became sensitive to the statistical sub-lexical regularities carried by the orthography-phonology and orthography-semantics mappings.

Finally, Gelzheiser (1991) compared two methods to teach symbol-syllable correspondences to children. In a specific correspondence condition, children were explicitly taught to pair a symbol with its spoken syllable (e.g., “jec”, “heg”) whereas children in the pattern detection training condition were taught to look for patterns within the stimuli, but not explicitly told what the patterns are. Post-test results showed that specific correspondence training was helpful to participants in reading trained symbol-syllable pairs, but was less effective in inducing a more general ability to decode patterns in novel stimuli. Pattern detection training may be more useful as it encourages participants to independently induce correspondences between symbols and sounds.

2.2.1.D Training order

Zhao and Rueckl (2012, Experiment 2) used a similar training paradigm to Zhao et al. (2018), except participants learned consistent O-P and O-S mappings only (see section 2.1.1.C). To examine the effect of learning order, participants were divided into two groups: one learned O-P mappings before O-S mappings, while the other group learned in the reverse order.

There was no effect of learning order during training, or in the sub-lexical post-tests. However, learning order did influence performance on the lexical post-tests, and this interacted with the position of the phonetic/semantic radicals in the character. Specifically, when the phonetic radicals were on the left of the character, those who were trained on O-P mappings later performed better in reading trained items at post-test. When the phonetic radicals were on the right side of the character, those who were trained on O-S mappings later were better at recalling the meanings of the trained characters.

2.2.1.E Handwriting and motor efforts

Two studies have investigated whether handwriting facilitates learning novel O-P mappings. Bhide (2018) used a semi-artificial orthography (modelled on the Devanagari script) in which complex aksharas are formed from simple aksharas (i.e., akin to forming a consonant blend from two consonants in Hindi graphs). Four training methods were employed, two of which involved reading and answering multiple-choice questions and the other two required copying (writing while viewing a model) or writing (writing from memory). There was evidence for better learning when the complex aksharas were learned through copying and writing, as compared to multiple-choice questions. Within the motor training methods (copying vs. writing), participants were better able to read the complex aksharas after writing than copying. In contrast to these findings, Catronas et al. (2020) found no effect of handwriting on learning. Participants learned to read a novel syllabic script, with one group receiving visual-only training (passive exposure to pseudoletters and sounds) and the other receiving visual-motor training (copying pseudoletters after seeing a demonstration and hearing the associated sounds). Contrary to their predictions, both groups read trained items well, and showed generalisation.

2.2.2 Pre-training instructions

Yoncheva et al. (2010) explored the effect of explicit instruction (see also Maurer et al., 2010). All participants saw the same stimuli and followed the same procedure during the exposure phase, but the type of pre-exposure instruction varied. The whole-word group was asked to associate an entire character with an auditory word whereas those in the grapheme-phoneme group were told to focus on associating parts of the character and phonemes in each spoken word. Following training, participants were tested with a reading verification task, choosing between two options that matched the auditory word. Results showed that the whole-word group was more accurate and faster in trained items. In contrast, the grapheme-phoneme group was more accurate with novel items while the whole-word group was at chance. A similar effect was observed by Verwimp et al. (2023) in an experiment with Dutch-speaking children who learned to map Hebrew graphemes with Dutch phonemes. They compared the effects of goal-directed instructions (where children were told to learn as many symbols as possible) and implicit instructions (where children were told to play a computer game, the goal of which would become clear at the end) on learning. There was a clear learning advantage in the goal-directed condition. Children were more accurate in judging whether an artificial letter matched a speech sound, and they were also better at naming the letters out loud at post-test.

Byrne and Carroll (1989, Experiment 2) found no advantage for explicit instruction in an experiment where participants received either explicit instructions (where they were told that systematic associations existed between symbols and sounds and were asked to try and figure them out) or implicit instructions (where they were told to learn the sound of each symbol). Neither group performed above chance in the generalisation task, making the null effect of instruction type difficult to interpret. However, it is important to note that the stimuli in this

experiment were complex. Each symbol comprised two parts where the upper portion represented the place and manner of articulation whereas the lower portion represented voicing. In other words, participants needed to learn associations between phonetic features and the components of a symbol. This may explain the lack of learning as well as there being no effect of instruction type.

Several studies have investigated how grain size focus during initial instructions affects learning O-P mappings, by comparing instructions that focus on letters (small grain) versus whole words (large grain). Brennan and Booth (2015) and Brennan and Kiskin (2020) explored this using artificial symbols (3 days of training) and Russian Cyrillic (2 days of training), respectively. Participants were assigned to one of the two instruction conditions, where their attention was directed to letters (the smallest grain size) or whole words (the largest grain size) before whole-word training. Following each day of training, participants completed a test where they determined whether the visual and auditory words matched. At post-test, they completed further matching tasks that assessed their knowledge of letter-phoneme and rime-rhyme mappings. In the rime-rhyme matching task, for example, participants had to decide whether the novel artificial written word “smoad” matched the auditory word /smɔd/ (correct trial) or /smif/ (foil trial), for which the two auditory words only differed in the rime.

In both studies, large- and small-grain groups achieved similar accuracy in matching trained words. As for test items, the large-grain group showed higher accuracy in the rime-rhyme matching task than the small-grain group, though this was only found in a subset of test items in Brennan and Kiskin’s (2020) experiment. On the other hand, higher accuracy in the letter-phoneme matching task for the small-grain group was only found by Brennan and Kiskin

(2020). Additionally, Brennan and Booth (2015) noted slower RTs across all tasks for the small-grain group, whereas Brennan and Kiskin (2020) found no RT differences. Overall, these findings suggest that initial grain size instructions impact sensitivity to rime-rhyme and letter-phoneme mappings, though differences between the experiments highlight the potential role of visual familiarity of the orthography (symbols vs. Russian Cyrillic) and training duration in shaping these effects.

2.2.3 Other non-linguistic external factors influencing the learning of O-P mappings

In addition to the factors discussed above, a small number of experiments have examined the impact of non-invasive brain stimulation and sleep deprivation on learning O-P mappings.

Non-invasive brain stimulation has been shown to facilitate language learning (e.g., Flöel et al., 2008). Four experiments have examined whether such techniques (transcranial direct current stimulation [tDCS]; McMillan et al., 2017; Xue et al., 2017; Younger & Booth, 2018) and transcutaneous auricular vagus nerve stimulation (taVNS; Thakkar et al., 2020) can enhance the learning of O-P mappings within an artificial orthography. McMillan et al. (2017) trained native English-speaking dyslexic and typical readers to read English pseudowords written in Hungarian Runes and Georgian Mkhedruli. During training, half of each group received active stimulation on the left lateralisation of the temporoparietal cortex. Contrary to predictions, tDCS improved performance only for dyslexic readers. Dyslexic participants with active stimulation performed better on the low-frequency inconsistent trained items in the immediate post-test and retained more of these items a week later relative to individuals in the sham group (i.e., those who did not receive active stimulation). Additionally, dyslexic readers with active stimulation also showed an interaction effect between consistency and frequency in the word reading tasks for both trained and untrained

items; this effect was absent in the dyslexic group who received sham stimulation. A similar effect was reported in Younger and Booth (2018) where participants learned English words written in a Klingon-like script. Their results showed that stimulation improved acquisition rates and untrained word reading accuracy, but this effect was limited to participants with lower reading skills based on standardised reading measures. As for long-term retention (post-test performance four weeks after training), stimulation had a positive impact on participants of all skill levels.

The effect of stimulation has also been observed in experiments where participants were trained on natural orthographies, rather than artificial or ancient scripts. Xue et al. (2017) taught participants Korean Hangul that was either written with consistent GPCs (transparent script) or arbitrary GPCs (non-transparent script). One group of participants received tDCS on the left temporoparietal cortex (LTPC; target site) whereas the other group received it on the visual cortex (control site). Their results showed that stimulation on the target site facilitated reading of untrained words in the transparent script, whereas there were no effects on reading trained words in the non-transparent script. However, this was only observed in RT but not in accuracy. A similar pattern was observed in the post-test where the stimulation effect was only found in RT. On the other hand, Thakkar et al. (2020) examined the effect of taVNS on learning to read Hebrew letters. Their results showed that both the control and stimulation groups performed close-to-ceiling when reading Hebrew consonant-vowel combinations with no significant difference between the two groups. However, the stimulation group outperformed the control group in the rapid automatised naming task and decoding task.

Tamminen et al. (2020) is the only study that has used artificial orthography to explore the effects of sleep deprivation on the process of learning to read. In their study, participants learned GPCs from English pseudowords written in Hungarian runes after (Experiment 1) or before (Experiment 2) a night of sleep deprivation. Following two nights of recovery sleep and again after 10 days, participants were tested on various tasks with both trained and novel words. Contrary to predictions, Tamminen and colleagues found that participants in the control and sleep deprived groups had comparable performance in reading both trained and novel words. The only differences observed were that when tested on phoneme knowledge, the control group gave faster responses than the sleep deprived groups in Experiment 1 whereas they were more accurate than the sleep deprived groups in Experiment 2. These findings suggest that the absence of sleep before or after learning had a limited impact on the acquisition or generalisation of the newly learned GPCs.

Having reviewed factors that influence O-P learning in artificial orthography experiments, we now consider the learning of orthotactic regularities. These regularities concern permissible letter sequence, patterns and positions in written words; for example, English consonants are often doubled after a single-letter vowel but not after a vowel spelled with multiple letters (e.g., “carrot” but not *carot).

2.3. Orthotactic regularities: Linguistic factors

Five subtypes of orthotactic regularities have been examined across 33 experiments on learning orthotactic regularities from artificial scripts: graphotactic constraints (Samara et al., 2019, Experiments 1 & 2; Samara & Caravolas, 2014; Singh et al., 2021, Experiments 1-3; Singh, 2021, Experiments 1, 2, 6 & 7), bigram or trigram frequency (Chetail, 2017, Experiments 1a, 1b & 2; Chetail & Sauval, 2022; Fernández-López & Perea, 2023; Ise et al.,

2012; Vidal et al., 2021, Experiments 1-3), morpheme or morphological structure (Lelonkiewicz et al., 2020, Experiments 1 & 2; Lelonkiewicz et al., 2023, Experiments 1a, 1b, 2a, 2b & 3; Wu et al., 2011, Experiments 1-3), positional regularities in Chinese characters (He & Tong, 2017, Experiment 1; Tong et al., 2020, Experiment 1; Tong et al., 2023) and letter inventory (Laine et al., 2014). These experiments have examined several linguistic factors.

2.3.1 Frequency

Fourteen experiments examined the effect of frequency on learning orthotactic regularities, with nine experiments focusing on bigrams/trigrams. In Chetail (2017, Experiments 1a & 1b), participants saw 320 artificial words, each comprising symbol strings with frequently occurring bigrams. Results from a wordlikeness task (where they had to decide which one of two options was more like the ones from the exposure phase) showed that participants developed sensitivity to frequently occurring bigrams and their positions (see also Chetail, 2017, Experiment 2; Chetail & Sauval, 2021; Fernández-López & Perea, 2023, for similar findings). Similar sensitivity to bigram frequencies has also been observed with other types of stimuli (Ise et al., 2012; Vidal et al., 2021, Experiments 1-3).

Wu et al. (2011, Experiments 1-3) demonstrated a robust frequency effect using both within- and between-subject designs. In their experiments, participants learned two types of frequency in artificial characters created from Tibetan letters: co-occurrence frequency (i.e., how frequently the sub-components of a visual character appear together) and cue of position frequency (i.e., the frequency of a certain component occurring in a given position within a word). Three groups of participants saw the stimuli in an exposure phase once, twice or four times, respectively. Wu and colleagues found that participants became sensitive to the co-

occurrence and position frequency after exposure, and that this sensitivity increased with increasing exposure.

Several experiments have manipulated frequency by varying exposure levels between participants. In Samara and Caravolas (2014), for example, participants were assigned to either a short or long exposure condition (9 vs. 18 repetitions per artificial word) and learning was tested via a legality judgement task. While there was evidence of learning, there was no significant effect of frequency. Laine et al. (2014) investigated frequency effects in relation to learning orthographic surface features. In their experiment, participants were exposed to artificial words that contained an orthographic surface feature (in that all of the artificial words were created from the same six consonants) and syllabic features (in that all of the artificial words had a CVCVCV structure). Critically, participants were divided into three groups where Group 1 saw two training items once each, Group 2 saw 20 training items once each and Group 3 saw 20 training items three times each. In line with Samara and Caravolas' (2014) findings, there was no effect of frequency on legality judgement performance in trials related to the orthographic surface feature.

2.3.2 Position of regularity

Orthotactic regularities can appear in different positions within a string. Some evidence suggests a string-initial processing advantage, with orthotactic regularities being easier to process at the initial than internal positions. Fourteen experiments investigated this type of effect. For example, Chetail (2017, Experiments 1a, 1b & 2) found better performance in a wordlikeness task for items where the bigrams had been experienced in external (initial and final) than internal positions during the exposure phase (see also Fernández-López & Perea, 2023). In addition, when making same-different judgments, participants were more able to

identify transposed-letter pairs in the initial position (Fernández-López & Perea, 2023). This evidence suggests that people are more sensitive to bigrams at the initial position of a string.

Further to bigram frequencies, Lelonkiewicz et al. (2020, Experiments 1 & 2) observed a position effect using a between-participants design to examine sensitivity to affixes (akin to the English prefix re- in “rewrite” and suffix -ful in “fruitful”). Specifically, regardless of whether participants were trained on affix-like chunks in the string-final (Experiment 1) or string-initial (Experiment 2) positions during exposure, they were more accurate at identifying trained affixes when these appeared in the string-initial position at post-test. On the other hand, in a wordlikeness task where they had to judge whether a novel item belonged to the familiarisation language, participants who were trained on affixes in the string-initial position were more sensitive to whether the affixes matched the position seen during exposure. Lelonkiewicz et al. (2023) mostly replicated the position effect using shapes (Experiments 1a & 1b) and the Latin alphabet (Experiments 2a & 2b) as stimuli and affix detection and wordlikeness to assess learning. However, when participants were trained on the affix in the string-final position, sensitivity to the affix position was only seen for the Latin alphabet experiment, and not with shapes.

In contrast to these findings, Samara et al. (2019, Experiments 1 & 2) did not observe a string-initial learning advantage. Instead, they found sensitivity to graphotactic patterns at string-initial and string-final positions, but they further noted that evidence for a position effect (evaluated using Bayes Factors) was inconclusive. Additionally, He and Tong (2017, Experiment 1) and Tong et al. (2020, Experiment 1) found mixed findings for a position effect with Chinese-speaking children who were exposed to positional regularities in Chinese characters that were structured either as left-right or top-bottom characters. While He and

Tong (2017, Experiment 1) did not see a position effect with typical readers, Tong et al. (2020, Experiment 1) found that dyslexic participants were impacted by the character structure such that the left-right characters were recognised more poorly than the top-bottom ones.

2.3.3 Stimulus type

Eleven experiments examined the effect of stimulus type on learning orthotactic regularities. In Lelonkiewicz et al.'s (2023) study, participants learned affix-like chunks through passive exposure in scripts that comprised either abstract shapes (Experiments 1a and 1b) or Latin alphabets (Experiments 2a and 2b). Across both script types, people learned the chunks. However, there was a learning advantage for the Latin alphabet stimuli suggesting that the availability of linguistic information such as phonology can facilitate the learning of visual regularities. Vidal et al. (2021) compared sensitivity to bigram frequencies across three types of stimuli including symbols (BACS fonts; Experiment 1), 3-D objects (Experiment 2) and circular gratings (Experiment 3). Results from the legality judgement task showed that participants were sensitive to the bigram frequencies across all three stimulus types. More importantly, there was no difference in the magnitude of the effects, leading the authors to conclude that sensitivity to the co-occurrence feature was developed to the same extent across materials.

Ise et al. (2012) compared the learning of frequent letter chunks using pronounceable and unpronounceable stimuli. In their experiment, participants learned frequent letter chunks embedded in five-letter pseudowords that were either arranged in a pronounceable CVCVC structure (e.g., XABOZ) or unpronounceable CCCCC structure (e.g., FTGCZ). Following exposure, participants were given a legality judgement task where they had to decide which

string belonged to the secret language. Performance differed between good and poor spellers, as determined by standardised measures of spelling ability. For the trained items, good spellers performed better with pronounceable strings than with unpronounceable strings, whereas the opposite was observed for poor spellers. For the untrained items, the same pattern was observed for good spellers, but there was no difference between the pronounceable and unpronounceable strings for poor spellers.

Samara and Caravolas (2014) compared the learning of positional and contextual graphotactic constraints. In their experiments, artificial words had a CVC structure using two sets of consonants. For the positional constraints, one set of consonants appeared only in the onset position whereas the other set was reserved for the coda position. For the contextual constraints, the medial vowel determined which set of consonants can appear in the onset and coda positions of the artificial words. In other words, contextual constraints (first-order constraint) were more difficult than the positional constraints (zero-order). Results showed a main effect of constraint types suggesting that participants had learned the positional constraints more readily than the contextual constraints.

However, this effect was not consistently observed by Singh et al. (2021, Experiments 1 & 2). In their experiment, participants learned two types of graphotactic patterns from similar CVC structured artificial words. In the simpler patterns, the medial vowel determined whether the second consonant would be a singlet or doublet. In the complex pattern, both vowels could take a consonant singlet or doublet in their second consonant positions; whether they were a singlet or doublet depended on the consonant letter. Using Bayes analyses, there was no conclusive evidence to suggest that children performed better when learning the simpler patterns. For adults, there was conclusive evidence for better generalisation in the

simpler pattern in the fill-in-the-blank task, but the evidence was inconclusive in the legality judgement task.

2.3.4 Other linguistic factors influencing the learning of orthotactic regularities

A small number of experiments have investigated the impact of positional consistency, semantic cues and contextual diversity in learning orthotactic regularities. Starting with the effect of positional consistency on learning Chinese radicals, He and Tong (2017, Experiment 1), Tong et al. (2020, Experiment 1) and Tong et al. (2023) created artificial characters comprised of radicals that varied in positional consistency. For example, in left-right-structured artificial characters, a radical with high positional consistency (100%) was in one position (e.g. left) across all four characters whereas a radical with moderate positional consistency (75%) was in one position (e.g. left) in three out of four characters and the other position in the other character. This manipulation was designed to mimic the positional regularity of radicals in Chinese characters. For instance, the radical 亻 is highly consistent in its positional regularity because it only appears on the left of a character (e.g., 休 "to rest", 伸 "to stretch"). Across the learning experiments, participants recognised more characters at post-test if they contained a radical that had high positional consistency in the exposure phase. This positional consistency effect was found in both typical and dyslexic readers (Tong et al., 2020) and across children in Grades 1 to 3 (Tong et al., 2023). However, there was no evidence for the effect on generalisation to new characters (He & Tong, 2017, Experiment 1; Tong et al., 2020, Experiment 1).

Lelonkiewicz et al. (2023, Experiment 3) examined whether semantic information impacts affix learning. Participants saw alphabetic letter strings containing four affixes in the word-final position. Half of the participants were assigned to the semantic group, in which each

string was preceded by a picture presenting one of four semantic categories (i.e., clothing, food, mammals, or musical instruments) during exposure. Each semantic category was consistently associated with a specific affix. In contrast, participants in the no-semantic group only saw noise squares before seeing the orthographic strings. At post-test, participants in the semantic group were more sensitive to both affix presence and position than those in the no-semantic group, as assessed by a wordlikeness task. In addition, participants who were better at associating affixes to their semantic categories also had greater sensitivity towards the presence and position of affixes. These findings suggest that semantic information can facilitate the learning of orthotactic regularities that map to affix presence and position.

Contextual diversity refers to the number of contexts in which the words occurred, independent of their frequency. Chetail and Sauval (2022) examined whether contextual diversity is important for new bigrams to become salient orthographic regularities, as compared to repeated exposure in a fixed context. Over two months, native French-speaking participants completed daily 20-minute sessions at home where they saw artificial words that contained bigrams that are illegal in French. They used a 2x2 within-participant design that crossed frequency and contextual diversity. Participants saw the bigrams either 4 times or 16 times a day (low- vs. high-frequency) and the bigrams were either embedded in the same two artificial words or distributed in 8 different words creating a contrast between low and high contextual diversity. Learning was assessed via a wordlikeness test in which participants judged which one of two options looked more familiar to the words they had encountered during exposure. Both frequency and contextual diversity influenced learning, with more learning for high frequency and high diversity items. There was also an interaction between these two effects such that the effect of contextual diversity was driven by high-frequency bigrams.

2.4 Orthotactic regularities: Non-linguistic external factor

While a substantial number of experiments have explored the effects of training methods and pre-training instructions on O-P learning in artificial orthographies, as reviewed in sections 2.2.1 and 2.2.2, only three experiments have investigated these factors with respect to orthotactic regularities. Singh et al. (2021, Experiments 1 & 3) examined the effect of explicit instruction on graphotactic learning. Using a one-back paradigm, participants in both experiments saw strings written in the artificial language and were asked to press a button when a stimulus repeated consecutively. However, those in Experiment 3 were explicitly told that the graphotactic patterns governing the artificial words before the exposure phase. Comparing across experiments, there was substantial evidence (evaluated with Bayes Factors) that those who received explicit instructions showed better learning.

Singh (2021, Experiment 6) examined graphotactic learning in a more natural setting where written words convey meaning. In a word learning paradigm, participants were exposed to graphotactic patterns while learning associations between objects and written words. Overall, participants learned both the graphotactic patterns and the object-word associations. In comparison, Singh et al. (2021, Experiment 4) exposed participants to these graphotactic patterns in meaningless words through a one-back task. There was no conclusive evidence that participants performed less well when learning took place in a word learning paradigm, suggesting that graphotactic patterns can be learned regardless of whether words have meaning. Singh (2021, Experiment 7) replicated the word learning paradigm from Experiment 6 and tested it further by explicitly informing a new group of participants of the graphotactic patterns prior to training. Consistent with earlier findings, participants who were given explicit instructions showed more graphotactic learning at post-test. At the same time, however, they performed less well on a word-object associations task. This suggests that

intentionally focusing on graphotactic learning might reduce attention from other aspects of learning.

2.5 Orthography-semantics (O-S) mappings: Linguistics factors

Only two experiments have specifically focused on learning orthography-semantic mappings using an artificial orthography. Both He and Tong (2017, Experiment 3) and Tong et al. (2020, Experiment 3) examined the effect of semantic consistency on learning Chinese radicals. Their experiments are like those on phonetic and positional regularity (as described in sections 2.1.2 and 2.3.2) but focused instead on semantic regularity reflected in Chinese characters. For example, radicals such as 木 are highly consistent in indicating the concept of wood such as 樹 (“tree”), while radicals such as 弓 (“bow”) are lower in semantic consistency, as only 30% of the characters (e.g., 弦 “string”) that contain this radical share its meaning (Chen & Weekes, 2004). The two experiments yielded different results. He and Tong (2017) observed a semantic consistency effect in generalisation, where participants were more accurate at generalising from items that contained radicals of high- and moderate-semantic consistency items during the exposure phase. There was no effect of semantic consistency in the recognition task. Tong et al. (2020, Experiment 3) reported semantic consistency effects in both tasks. Participants were again more accurate at generalising radicals with higher semantic consistency to novel items. Surprisingly, however, they were also better at recognising artificial characters that contained radicals with low semantic consistency.

3. How does evidence from artificial orthographic learning experiments inform our understanding of statistical learning in reading and spelling acquisition?

Having reviewed the nature and content of the literature, we now turn to consider what this evidence base tells us about reading and spelling acquisition as a statistical learning process. To explore this, we first examined the relationship between learning outcomes in artificial orthography experiments and performance on individual difference measures across all of the experiments that met our criteria for statistical learning. We then examined how differences in experimental design may impact learning and from this, our understanding of statistical learning from this body of evidence. As noted in the introduction, definitions of statistical learning vary. In this review, we adopted a broad definition to capture an individual's ability to extract regularities from exposure without any explicit instructions on the target patterns.

3.1 Relationship between learning outcomes and individual differences in other behavioural measures

Of the 114 experiments included in our review, 26 examined the relationship between learning within the artificial orthography paradigm and individual differences in reading, spelling or another domain, and of these, 19 experiments met our criteria for tapping statistical learning[4]. These experiments examined either O-P mappings or orthotactic regularities. As summarised in Table 2.2, the individual difference measures can be broadly categorised into five domains: reading, spelling, phonological awareness, cognitive abilities and others.

[4] Four experiments (E009, E013, E014 and E015) contain learning conditions that do not meet our criteria as a statistical learning experiment according to our definitions. Data from these learning conditions are not considered in this section.

Table 2.2. Individual difference measures used in artificial orthography learning experiments

Individual differences domain	Individual differences measures	Experiment ID	Count of experiments
Reading	Real letter/word reading	E003, E004, E005, E009, E046, E066*, E071, E072*, E073*, E076*, E077*, E094	12
	Pseudoword reading	E004, E005, E013, E046, E071, E073*	6
Spelling	Spelling to dictation	E005, E040*, E072*, E073*, E076*, E077*	6
	Spelling recognition	E004, E005	2
Phonological awareness	Elision	E014, E015	2
	Phoneme deletion	E005, E046	2
	Blending words	E015	1
	Segmenting words	E014	1
Cognitive abilities	Non-verbal IQ	E001, E003, E006, E054	4
	Phonological working memory	E001, E006, E054	3
	Rapid naming	E005, E046	2
	Inhibitory control	E054	1
Others	English reading proficiency (self-reported)	E054	1
	Orthographic knowledge	E046	1
	Pair association learning	E071	1
	Vocabulary size	E054	1

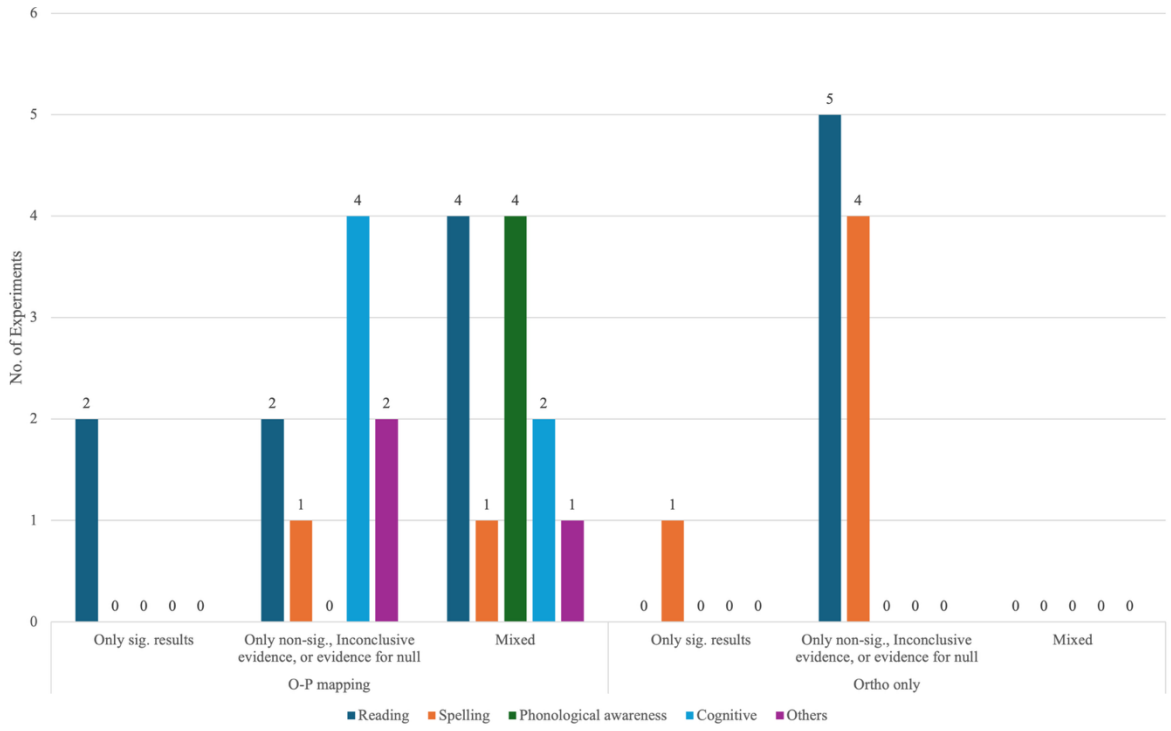
Note. Experiments that examined orthotactic regularities are marked with asterisk (*), otherwise they examined O-P mappings.

Addressing the nature of the relationship between individual differences, such as reading ability, and learning outcomes in an artificial orthography learning experiment is challenging for several reasons. First, learning outcomes are measured quite differently in different experiments. For example, Singh et al. (2021) used fill-in-the-blank and legality judgement tasks to measure learning, whereas Schmalz et al. (2021) indexed learning by calculating the number of training blocks to reach 70% reading accuracy. Second, some experiments used multiple tasks to assess the same ability (e.g., using word and pseudoword reading tasks to assess reading ability), which could lead to mixed findings, if the tasks are correlated differently with learning outcomes. Third, analysis methods vary. Some experiments used

simple correlations whereas others used regression models to control for confounds and other factors. Potentially, this might affect the sensitivity to detect relationships. Finally, publication bias may have affected the evidence base as positive evidence is more likely to be reported in the literature. While more recent experiments have embraced pre-registration (e.g., Schmalz et al., 2022), this was not the norm a few years ago.

Given these challenges, we counted the number of experiments that reported consistent findings for (1) significant results, (2) non-significant, inconclusive evidence or evidence for the null and (3) mixed findings for a correlation between artificial orthography learning outcomes and individual difference measures in each category. Figure 2.5 summarises the count of experiments in each category. This data is further divided between experiments focusing on O-P mapping and orthotactic regularities.

Figure 2.5. Bar chart showing the number of experiments reporting a relationship between artificial orthography learning outcomes and individual difference measures by their results.



As shown in Figure 2.5, only three experiments consistently reported a significant correlation between learning outcomes and all of the individual differences measures they considered. For the two experiments examining O-P mappings, Aravena et al. (2013) observed a significant positive correlation between reading rate in native language and learning in the artificial orthography in a sample of dyslexic readers, and Verwimp et al. (2023) reported similar findings. Ise et al. (2012) is the only experiment to investigate sensitivity to orthotactic regularities in an artificial script in relation to individual differences. They found a consistent and significant relationship between people's spelling ability and their performance on a legality judgement task that probed learning from the artificial orthography.

In contrast to these positive findings, eight experiments reported only non-significant results using frequentist statistics: seven on O-P mapping and one on orthotactic regularity. For O-P

mapping, these non-significant patterns ranged across several domains including reading (Bitan & Booth, 2012), spelling (Aravena et al., 2016) and cognitive abilities such as non-verbal IQ (Acha et al., 2023; Aravena et al., 2013; Bartolotti & Marian, 2019), rapid serial naming (Law et al., 2018) and phonological working memory (Acha et al., 2023; Bartolotti & Marian, 2019). Additionally, non-significant patterns were found for vocabulary (Marian et al., 2021) and self-reported language proficiency (Marian et al., 2021). In the one experiment focused on the learning of orthotactic regularities, Samara et al. (2019) found no significant correlation between this learning and individual differences in reading.

Instead of frequentist statistics, five experiments used Bayes Factor as their primary analysis method, which allows researchers to determine whether their data provides conclusive evidence for the alternative hypothesis, the null hypothesis, or inconclusive evidence when neither hypothesis is supported. Schmalz et al. (2021) found a mix of inconclusive evidence and evidence for the null in correlations between learning O-P mappings and both reading ability and paired association learning. Similarly, in experiments on graphotactic learning, Singh (2021, Experiments 1 & 2) and Singh et al. (2021, Experiments 1 & 2) reported inconclusive evidence for most correlations in their study, with evidence for the null in some cases.

Seven experiments on learning O-P mappings reported a mix of significant and non-significant findings across different categories of individual difference measures. We offer three reasons to explain these mixed findings. First, some experiments had more than one learning outcome and it might be that some are more sensitive or reliable than others, and that this impacts the extent to which individual differences are detected. This is particularly apparent in Aravena et al. (2016, 2018) and Law et al. (2018) where there were three learning

outcomes – accuracy in reading artificial words, and both the speed and accuracy of performing a letter-speech sound identification task for items presented in the artificial orthography. While Law et al. (2018) observed significant correlations between individual differences in reading the artificial orthography and a range of reading-related abilities, none of these abilities correlated with learning as indexed by performance on the letter-speech sound task. Similar mixed patterns were described by Aravena et al. (2018). Taken together, these findings demonstrate that different learning outcomes may vary in sensitivity (see also Bolger & Perfetti, 2007).

Second, even within the same category of individual difference measures, the specific abilities being assessed can vary, potentially leading to different patterns of correlations. For example, Marian et al. (2021, Experiment 1) examined a range of cognitive abilities including non-verbal IQ, inhibitory control and phonological working memory (digit span and nonword repetition). Only non-verbal IQ was related to learning the artificial orthography as assessed by word recognition and production measures. Non-verbal IQ also correlated with learning rate during training, as did phonological working memory when measured by digit span, but not when indexed by nonword repetition. This shows that even though all these measures are related to cognitive abilities, only some aspects of this ability are significantly related to learning the artificial orthography.

Third, patterns of individual differences may depend on the nature of the exposure phase of the experiment – as discussed in earlier sections, experiments vary considerably in terms of a range of linguistic and non-linguistic features, and it is possible that some are more amenable to learning in ways that covary with people's reading and reading-related abilities. Brennan and Booth (2015) trained a group of participants to read an artificial orthography through

whole-word exposure (i.e., with no explicit instruction on GPCs). Performance on the word segmentation task, but not the elision task, accounted for unique variance in cross-modal matching accuracy (a task where participants determined whether the visual and auditory word, rime or letter presentations of the artificial orthography matched). Since participants in this group were instructed to focus on whole words during exposure, the authors argued that participants who had better word segmentation ability (meaning that they were already good at breaking up larger structures) were at an advantage and thus achieved higher scores on the matching task. The skill tapped by the elision task, which involves synthesising smaller phonological units into larger ones, was likely less relevant for participants who learned the stimuli at the whole-word level during exposure. This is to say that the correlations between the individual difference measures and artificial orthography learning outcomes may vary depending on how participants were trained on the orthographic patterns.

In summary, only a small number of experiments that met our criteria for statistical learning also examined individual differences and most of these focused on O-P mappings. While there is some evidence for positive correlations, most experiments report either non-significant results, inconclusive evidence, or evidence for the null or mixed findings. It is not possible to discern a particular pattern across studies. Moreover, the lack of clear and stable relationships questions the extent to which artificial orthography learning experiments reflect the product of orthographic learning in natural language contexts.

3.2 The impact of experimental design variability on our understanding of statistical learning

This systematic review makes clear that experimental designs that use the artificial orthography vary hugely. This makes it impractical to conduct a meta-analysis to meaningfully compare the magnitude of learning effects across experiments or to decide what

factors might moderate learning effects. Instead, we chose to conduct a qualitative comparison of the evidence base instead. As sensitivity to orthotactic regularities has received less attention in previous review articles, we focus our discussion on this type of orthographic learning. To do this, we next describe the main characteristics of the orthotactic learning experiments that met our definition of statistical learning. From this evidence base, we then consider (1) the conditions under which learning occurred, (2) the methods used to assess newly acquired knowledge and (3) the approaches used to measure the learning effects. By examining these aspects, we aim to provide insights into how learning effects were produced in each experiment, and how this might impact our understanding of statistical learning from this body of evidence.

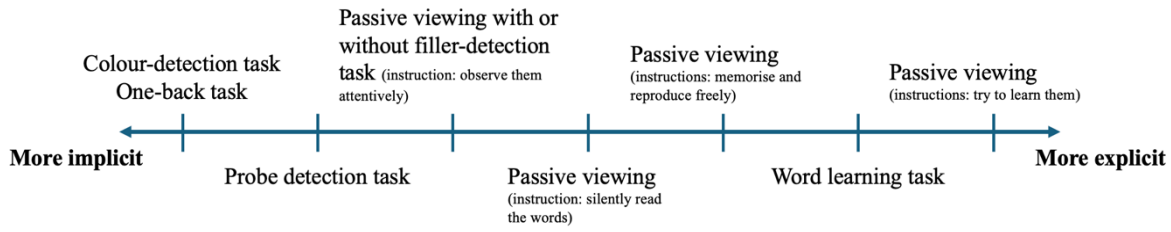
Of the 31 orthotactic learning experiments that met our definition of statistical learning, the majority comprised only one (71%; $N = 22$) or two training sessions (22%; N = 7) with the exposure phase often only lasting a few minutes[5]. The total duration of the entire experiment ranged from 30 to 60 minutes. Only two experiments lasted for an extended period: Chetail (2017, Experiment 2) extended over four days and Chetail and Sauval (2023) involved daily sessions over a two-month period. In addition, nearly all experiments assessed learning immediately after the end of the exposure phase, with no longer-term follow-up tests. The one exception is Chetail and Sauval's (2023) experiment which included six test points, distributed before, during and after the training period. Despite the brevity of most of the experiments, all reported significant learning of the specific types of orthotactic regularities they examined.

[5] We recognise that reporting the total duration of the exposure phase is more meaningful however most studies did not explicitly report this information. We report the total duration of the experiment to show that even with the testing phase, most experiments are very short.

3.2.1 The continuum of explicitness in learning: From implicit to explicit learning

Experiments vary in how stimuli are experienced during the learning phase, and the tasks used can lead to learning processes that span a continuum from more implicit to more explicit. Figure 2.6 shows an overview of the tasks used in the exposure phase in the 31 experiments under consideration, organised along an explicitness continuum based on their task demand and instructions given to participants. Following Hulstijn's (2005) definitions, we differentiate between two key dimensions of learning. The first is incidental vs. intentional learning, which refers to the learning condition and reflects whether participants were informed beforehand that they would be tested on the retention of a particular type of information. The second is implicit vs. explicit learning, which refers to the learning process and reflects whether participants had a conscious intention to find out whether the input information contains regularities. Although incidental learning often leads to implicit learning, and intentional learning is often associated with explicit learning, the two dimensions are not synonymous (Hulstijn et al., 2003).

Figure 2.6. Explicitness scale of tasks used in exposure phase in statistical learning experiments on orthotactic regularities.



On the left end of this scale are cover tasks, where learning occurs under more incidental conditions. Typically, participants are engaged in an irrelevant activity as they experience the orthographic patterns. For example, in Samara et al. (2019, Experiments 1 & 2) and Singh (2021, Experiment 2), participants pressed a corresponding key when a stimulus changed

colour (see also Samara & Caravolas, 2014, for another version of the colour-detection task). Six experiments used a one-back task (Singh et al., 2021, Experiments 1 & 2; Singh, 2021, Experiment 1; He & Tong, 2017, Experiment 1; Tong et al., 2020, Experiment 1; Tong et al., 2023) in which participants press a key when a stimulus repeated consecutively. In a similar vein, Chetail and Sauval (2023) used a probe detection task which exposed participants to target bigrams incorporated into a computer game that participants completed daily for three months. They were shown a probe (which could be a letter, a bigram or an open bigram) and asked to click on the targets containing the probe as quickly as possible. Although this required that some attention was paid to the linguistic stimuli, there was no instruction about the targets or a requirement to learn them.

Moving to the centre-right of Figure 2.6, passive viewing can result in learning processes that vary in their explicitness depending on the instructions given to participants. For example, Lelonkiewicz et al. (2020, Experiments 1 & 2; 2023, Experiments 1a, 1b, 2a, 2b & 3) simply asked participants to observe the stimuli attentively. Chetail (2017, Experiments 1a & 1b) and Fernández-López and Perea (2023) did so too, but as an additional attention check, participants pressed a button when they saw strings of Latin alphabets. The learning that occurs under these conditions is relatively implicit, as participants were not asked explicitly to look for the target structure of the input, or to engage with the stimuli. Passive viewing tasks can be adapted to promote more explicit learning. For example, Ise et al. (2012) asked participants to memorise the artificial words and reproduce them after the word disappeared from the screen while Laine et al. (2014) asked participants to read the artificial words silently. In Vidal et al. (2021, Experiments 1-3), participants were explicitly asked to pay attention to the stimuli and to try to learn them. Word learning tasks also appear on the right of Figure 2.6. For example, participants in Singh's study (2021, Experiment 6) engaged in a

task where the main goal was to learn to associate artificial words with novel objects. The target graphotactic patterns were embedded in the artificial words. In other words, participants were actively engaged in a learning task while being implicitly exposed to the target orthotactic regularity.

To summarise, learning statistical orthotactic regularities can take place under conditions that vary in two aspects: (1) participants' level of engagement with the stimuli and (2) the type of instructions they are given before exposure. These observed learning effects suggest that people can extract statistical regularities whether by passively viewing them or engaging in learning unrelated information in the orthography (e.g., object-word associations). They can also become sensitive to these regularities both when explicitly asked to learn something or when there is no specific intention to do so. Importantly, the duration of these tasks tends to be quite short, suggesting that statistical learning can take place rapidly and with ease (Schapiro & Turk-Browne, 2015).

3.2.2 Variations in learning measures: From recognition to generalisation

Just as the 31 experiments varied in terms of exposure characteristics, they also varied in terms of the type of test used to assess learning outcomes, as summarised in Table 2.3. We identified two types of post-test: those that tap into generalisation of the newly learned knowledge and those that assess how well the learned items are recognised and processed.

Table 2.3. Methods of testing and measuring learning effects in the artificial orthographic learning experiments on orthotactic regularities

Task	Type	Learning effects	Experiment ID	Count of experiment
Legality judgement	G	• Endorsement rate (i.e,. proportion of yes responses)	E044, E047, E048, E049, E050, E051, E052, E053, E066, E067	10
		• Accuracy	E040, E072, E073, E074, E076, E077	6
		• d-prime score	E065, E095, E096, E097	4
		• Reaction time	E044, E065	2
Wordlikeness	G	• Selection rate of critical items over control items	E025, E026, E027, E028, E030	5
		• Percentage in selecting items with high frequency or co-occurrence, or old characters	E104, E105, E106	3
		• Accuracy	E036	1
		• Positional preference	E091	1
		• Reaction time	E036	1
Fill-in-the-blank	G	• Accuracy	E072, E074, E076, E077	4
Detection	P	• d-prime score	E047, E048, E049, E050, E051, E052, E053	7
		• Error rates	E025, E026, E027	3
		• Reaction time	E025, E026, E027	3
Recognition	R	• Reaction time	E036, E044, E090	3
		• Accuracy	E036, E090	2
		• d-prime score	E044, E091	2
Same-different	P	• Accuracy	E030	1
		• Reaction time	E030	1

Generalisation has been most commonly tested using legality judgement tasks ($N = 20$) in which participants see novel items written in the artificial script and are asked to decide (typically yes/no) whether they belong to the new language. Six of these 20 experiments (Ise et al., 2012; Singh, 2021, Experiments 1, 2 & 6; Singh et al., 2021, Experiments 1 & 2) measured proportion correct and noted whether this was above chance. Ten experiments

examined endorsement rate (i.e., the proportion of 'yes' responses), specifically investigating the effects of various experimental manipulations on performance including item legality (Samara et al., 2019, Experiments 1 & 2), letter inventory (Laine et al., 2014) and string type (Lelonkiewicz et al., 2020, Experiments 1 & 2; Lelonkiewicz et al., 2023, Experiments 1a, 1b, 2a, 2b & 3). Four experiments used d-prime to index sensitivity and account for response bias in endorsement rates; these experiments asked whether d-prime scores differed between items that contained bigrams from exposure items and those that did not (Vidal et al., Experiments 1-3) and whether the scores were above chance (Samara & Caravolas, 2014). Finally, two experiments (Laine et al., 2014; Samara & Caravolas, 2014) assessed generalisation by measuring reaction time to endorse a given item in the task.

Ten experiments used a wordlikeness task in which participants choose which one of the presented novel items is more similar to the training items, or more familiar to them. This task is conceptually similar to legality judgement but instead of yes/no response, participants select one option from several alternatives. Arguably, this requires more specific knowledge to be extracted from the exposure phase. Five experiments examined whether selection rates for critical items over control items were above chance (Chetail, 2017, Experiments 1a, 1b & 2; Chetail & Sauval, 2022; Fernández-López & Perea, 2023) and whether performance was influenced by bigram frequency or amount of exposure (Chetail & Sauval, 2022). One experiment measured learning based on above-chance accuracy and whether consistency impacted their accuracy and reaction time (He & Tong, 2017, Experiment 1). Four other experiments used a more tailored approach, examining the effect of consistency on participants' positional preference (Tong et al., 2020, Experiment 1) and whether participants were more likely to select high-frequency or consistent items, or characters that they have encountered during exposure (Wu et al., 2011, Experiments 1-3). A final measure of

generalisation seen in the literature is the fill-in-the-blank task, used in four experiments (Singh, 2021, Experiments 1 & 6; Singh et al., Experiments 1 & 2). Here, participants are shown a novel word string (written in the now familiar artificial script) with a missing vowel, and asked to choose between one of two vowel options to fill the missing gap. Above chance performance is taken as evidence of generalisation, and therefore of learning.

Turning to the recognition and processing of the experienced words themselves, four experiments used a simple recognition task where participants judged whether an item had been a trained item and present in the exposure phase. Learning was measured by testing whether the d-prime scores (Laine et al., 2014; Tong et al., 2020, Experiment 1) or mean accuracy was above chance (He & Tong, 2017, Experiment 1; Tong et al., 2023). Additional analyses also examined whether positional consistency influenced the d-prime scores (Tong et al., 2020, Experiment 1), and the effect of training (trained vs. untrained items; Laine et al., 2014) and positional consistency (He & Tong, 2017, Experiment 1; Tong et al., 2023) on reaction time.

Ten experiments used a letter detection (Chetail, 2017, Experiment 1a, 1b & 2) or an affix detection task (Lelonkiewicz et al., 2020, Experiments 1 & 2; Lelonkiewicz et al., 2023, Experiments 1a, 1b, 2a, 2b & 3) where participants judged whether an artificial word contained a target letter or affix. Learning effects were examined through the frequency effects on the error rates and reaction time (letter detection), and by comparing d-prime scores across string positions (affix detection). Finally, Fernández-López and Perea (2023) used the same-different task where participants were shown two items sequentially and asked to judge whether they were the same or different. Learning was measured by testing the effect of bigram frequency on accuracy and reaction time.

To summarise, experiments examining statistical learning of orthotactic regularities have assessed learning in terms of whether people can recognise the items from the exposure phase, and whether they can generalise this knowledge to novel items. While these tasks generally require participants to make judgements on items, task demands vary. That is, some tasks may ask participants to choose between yes and no responses, while others require them to choose the correct items among several foil items. In addition, the indices used to measure learning effects are different across experiments, with some being more sensitive (e.g., d-prime sensitivity score that accounts for response bias) than others (e.g., whether the overall performance is above-chance). Overall, approaches to measuring orthotactic learning are not consistent, and this makes it challenging for researchers to directly compare results across experiments.

General Discussion

The artificial orthography learning paradigm has gained popularity in recent years for good reason. By introducing a novel orthography, researchers can isolate and manipulate a particular regularity or feature while controlling for others, and for differences in prior knowledge. This paradigm also provides a unique opportunity to examine the learning itself in a miniature environment, as opposed to tracking learning in natural language where learning unfolds across multiple episodes over long periods of time. The paradigm has been used to explore a range of specific questions, and with many variations in terms of methodology. Not surprisingly, therefore, the evidence base is complex and difficult to evaluate.

To make progress, we took a systematic approach to reviewing evidence from existing artificial orthography learning experiments. Specifically, we summarised the characteristics

of these experiments and identified the factors that may impact orthographic learning. As noted in the introduction, reading and spelling acquisition can be considered a feat of statistical learning in which repeated exposure shapes the connections between the three elements of a mental lexicon: orthography (spelling), phonology (pronunciations) and semantics (meaning) (Seidenberg & McClelland, 1989). Therefore, another aim of this review was to examine how evidence from this paradigm informs our understanding of statistical learning in reading and spelling acquisition. In the following section, we discuss the key findings for each research aim and consider the broader implication of artificial orthography learning for orthographic learning in natural language contexts.

Evidence gaps in artificial orthography learning experiments

By summarising the characteristics of existing artificial orthography learning experiments, our systematic review reveals three key trends in the literature. The most prominent of these is the strong focus on examining O-P mappings. This finding is perhaps not surprising given the focus on alphabetic languages in reading research (Share, 2008). However, while we acknowledge the critical role of phoneme-grapheme mappings, there are other regularities to be considered too. One type of regularity is provided by form-meaning mappings. In a corpus analysis, Berg and Aronoff (2017) showed that the spelling of some English suffixes provides a more consistent marker of lexical category than the corresponding phonological forms (see also Heyer, 2021; Treiman et al., 2021; Ulicheva et al., 2020). For instance, the phonological sequence /əs/ is found equally often in adjectives (e.g., “nervous” /nɜ:vəs/) and non-adjectives (e.g., “cactus” /kæktəs/) yet the written suffix -ous is predominantly used in adjectives, as in “hazardous” and “dangerous”. Similarly, people are sensitive to orthotactic patterns in their own writing system, and these regularities can be independent of phonology. Using pseudoword experiments, Cassar and Treiman (1997) showed that children as young as first

grade could already identify which consonants (e.g., “baff” vs. *bahh) and vowels (e.g., “sook” vs. *saak) are allowed to be doublets. The predominant focus of O-P mappings in this paradigm likely reflects a broader research trend that emphasises O-P mappings and perhaps overlooks other types of regularities in a writing system. Thus, future research should extend this paradigm to examine orthographic patterns beyond O-P mappings in a writing system.

A second key trend is that most data come from English-speaking participants, with only a minority of experiments sampling people with a non-English (or non-alphabetic) language background. This finding is expected, as much of the literacy development research is based on evidence from alphabetic languages (Share, 2008). However, this limits our understanding of orthographic learning from this paradigm as it largely reflects individuals who already heavily rely on phonological information when reading and spelling in their native language. Little consideration is given to non-alphabetic languages such as Chinese, which emphasises form-to-meaning mappings. In addition, most of the artificial orthography learning experiments recruited only adults as participants. While this sampling bias is not surprising, as recruiting adult participants is much easier, experiments comparing adults and children point to potential differences in learning effects between the two groups. For example, Singh et al. (2021) found that, at the end of the learning experiment, adults were more able to verbalise the orthotactic regularities in a post-test awareness questionnaire and they also outperformed children in generalising graphotactic patterns. Given that adults have had much more reading experience than children, they are more likely to actively search for patterns in the input and develop explicit awareness of these patterns. Future research should extend this paradigm to examine orthographic learning with both adults and children, as well as participants whose native language is non-alphabetic.

A third key trend is that to date, most artificial orthographic learning experiments asked participants to learn novel orthographic patterns in meaningless words. This raises concerns about external validity, as orthographic learning in natural language occurs in a meaningful context. A few experiments have directly compared orthographic learning with and without word meanings. For instance, Singh (2021) asked whether graphotactic learning was better in participants who were exposed to the patterns in meaningless words than those exposed to the patterns in a meaningful context. The results were inconclusive. As a result, it remains unclear how orthographic learning in meaningless contexts compares to learning in meaningful contexts. Future research should investigate how the presence of semantic information impacts orthographic learning in this paradigm.

Influence of linguistic and external factors on orthographic learning

As part of our first research aim, we systematically documented experiments using this paradigm based on factors that may influence orthographic learning. Across O-P mappings, O-S mappings and orthotactic regularities, we identified several common, fundamental factors that impact learning and two types of factor were identified: those based on linguistic features of an orthography and others based on non-linguistic external influences. In terms of linguistic factors, consistency influences learning. Not only does it influence the learning of O-P mappings in alphabetic languages (e.g., Taylor et al., 2011), but its effects also extend to O-S mappings and positional regularities in non-alphabetic languages such as Chinese (e.g., He & Tong, 2017). Similarly, frequency effects are ubiquitous, impacting both O-P mapping and orthotactic regularities. Sensitivity to distributional features such as consistency and frequency allows people to develop nuanced, context-conditioned knowledge of various orthographic patterns in the writing system and to use this knowledge in reading and spelling.

External factors such as pre-training instructions and training methods also influence the learning of both O-P mappings and orthotactic regularities, though there are substantially more experiments examining these effects in O-P mappings. Overall, explicit instructions given prior to training and explicit training both impact learning. Even though orthographic learning can take place under incidental learning conditions, people who receive explicit instructions or training are more likely to perform better at post-tests. However, the evidence from the paradigm is largely based on a small subset of highly consistent orthographic patterns. Since consistency influences how readily people learn phoneme-grapheme mapping (e.g., Aro & Wimmer, 2003) and reading accuracy (e.g., Ellis et al., 2004), it remains unclear how the consistency or complexity of orthographic patterns may interact with the effectiveness of explicit instruction or training.

While both explicit instruction and explicit training lead to better learning outcomes, it is important to consider the nuanced differences between them. In our review, we specifically contrasted the concept of explicit instruction given before training (e.g., informing participants about the presence of patterns in the stimuli to encourage them to approach the task with the intention of identifying those patterns) with explicit training (e.g., people are specifically shown how each symbol maps onto a phoneme in each trial). Both involve explicit learning, as participants are intentional with learning target patterns. However, explicit pre-training instructions do not guarantee explicit awareness, as participants may still struggle to recognise the target patterns despite being aware of their existence. Explicit awareness may also develop at different time points over the course of an experiment, making it difficult to precisely examine its impact on learning. In contrast, explicit training provides more direction instruction and thus participants are more likely to become explicitly aware of patterns. Given that this distinction is not currently clear within the artificial

orthography learning literature, future research should explore how the explicit learning process induced by explicit instructions and explicit training impacts orthographic learning.

Issues with external validity

The artificial orthography learning paradigm faces challenges with external validity, as learning in an experimental set up differs considerably from natural language learning contexts. It remains unclear the extent to which artificial orthographic learning reflects how individuals develop sensitivity to statistical regularities within their own writing systems. To address this, some experiments have investigated whether learning outcomes from artificial orthography paradigms are related to individual differences in reading, spelling and other related abilities, the rationale being that a correlation between the two would support the idea that learning in the artificial experiment can inform our understanding of learning to read and spell. We reviewed the evidence to determine whether this type of relationship exists. Our synthesis reveals no clear pattern, with mixed evidence and some evidence for there being no relationship. However, this lack of correlation may not necessarily reflect an issue with external validity and before abandoning the endeavour, it is important to consider whether other factors might account for the lack of relationships in the evidence base to date.

One possibility might be limitations in sample size. Experiments in our review varied in sample size from 20 to 120 participants, with an average of 60 participants in each experiment. It is possible that many lacked power analyses to test correlations reliably. This speculation is supported by experiments which used Bayes Factors as their primary analysis method (Schmalz et al., 2021; Singh et al., 2021). Unlike frequentist statistics, Bayes Factors allow researchers to distinguish whether the data provides conclusive evidence for or against the alternative hypothesis, or if the evidence is inconclusive. For most of the correlations

reported by Schmalz et al. (2021) and Singh et al. (2021), the evidence was inconclusive, meaning that the data supports neither the alternative nor the null hypothesis. This raises the possibility that the observed correlations might be in the expected directions, but fail to reach significance because of insufficient statistical power.

Another possible factor contributing to the lack of a relationship with individual differences concerns reliability. Nunnally and Bernstein (1994) pointed out that task reliability is affected by the number of trials, as fewer trials increase measurement error and reduce measurement reliability. In the current literature using the artificial orthography learning paradigm, there is no consensus on the optimal number of items needed to assess learning. For example, Samara et al. (2019) and Singh et al. (2021) both used legality judgement tasks to test similar graphotactic patterns, yet the former study included 16 test items (8 legal and 8 illegal unseen items) while the latter included 32 items (16 legal and 16 illegal unseen items). Neither study provided justifications for their item count or addressed the reliability of their tasks. Furthermore, both studies also included multiple versions of the stimuli to control for item-specific effects on learning, potentially further reducing measurement reliability. These examples highlight a broader issue regarding the lack of psychometrically robust task design in the current artificial orthography learning paradigm. Future research in this paradigm should draw on more recent work that addresses improvements in task reliability for measuring statistical learning (Siegelman, Bogaerts, & Frost, 2017).

Insights into statistical learning through artificial orthography learning

Across various strands of research, statistical learning is conceptualised in different ways. While some researchers assume this process to be completely incidental and implicit (e.g., Turk-Browne et al., 2009), others adopt a broader definition that emphasises how individuals

develop sensitivity to statistical regularities through exposure (e.g., Romberg & Saffran, 2010). This variation in conceptualising statistical learning is also evident within the artificial orthography learning paradigm, where researchers have considerable flexibility in experimental designs. To further understand statistical learning within this paradigm, we took a close look at features of experimental design, focusing specifically on experiments that investigated the learning of orthotactic regularities. Our analysis identified two key findings that relate to statistical learning.

First, the extent to which participants are intentional with learning varied depending on the instructions and task designs. Some experiments used incidental learning conditions such as a one-back task, while others explicitly asked participants to learn the stimuli without revealing the target orthographic patterns. These variations suggest that artificial orthography learning operates along a continuum from more implicit to more explicit processes. However, even in learning conditions designed to be incidental, it is debatable whether they fully eliminate participants' tendency to intentionally search for patterns during the exposure phase. This speculation is supported by the observation in Singh et al. (2021, Experiment 1) in which 20% of the adult participants were able to verbalise the patterns after incidental exposure. Thus, while evidence from these experiments supports the notion that orthographic learning is a form of statistical learning, it does so in a broader sense by highlighting that people can discover and extract patterns from the input, and does not necessarily assume the statistical learning process to be completely implicit or incidental.

Second, our synthesis showed considerable variations in how learning is assessed. Particularly, there are substantial differences in the statistical analysis methods used to measure the learning effects. Some methods (e.g., comparing d-prime sensitivity scores with

chance) are more sensitive than others (e.g., comparing mean accuracy with chance). This leads to the possibility that variation in learning across experiments may reflect differences in measurement. Additionally, some tasks such as the fill-in-the-blank may require more precise orthographic knowledge than others such as legality judgement. While we acknowledge that some task designs are better suited to testing specific orthographic patterns or populations, the lack of common methods for testing and measuring learning limits our ability to synthesise findings and fully evaluate evidence for statistical learning in orthographic learning.

Limitations

Before closing, it is important to note the limitations of this review. The first relates to the literature search and selection process. As our final search across six databases resulted in over 3000 entries after de-duplication, we did not conduct snowball sampling to extend our search. Therefore, it is possible that some relevant artificial orthography learning experiments were missed from the review. However, with 114 unique experiments included in our final sample, we believe our review provides a meaningful overview of the research within this paradigm.

A second limitation concerns how our research questions were addressed in order to keep the scope of the review manageable. Given the large number of experiments, we focused on highlighting key ideas when summarising the factors that may impact orthographic learning, rather than discussing every individual experiment that has examined them. In addition, we limited our detailed analysis of the variability in experimental designs to those experiments that focused on the learning of orthotactic regularities. These regularities have received less

attention than O-P mappings, reviewed elsewhere (e.g., Hirshorn & Fiez, 2014), yet there is an evidence base that is sizeable enough to synthesise (cf. O-S mappings).

Finally, in addressing the research question on individual differences, we were limited to a descriptive approach due to the significant variability across experiments, both in terms of the artificial learning experiment itself and the measures used to capture individual differences. This variability prevented us from conducting a quantitative comparison of effect sizes, especially given that publication bias may have led to non-significant results not being reported. Nevertheless, our observations provide insight into the challenges inherent in this approach, and highlight the need for factors such as reliability to be taken seriously.

Conclusion

This systematic review aimed to document and describe evidence from the artificial orthography learning paradigm and examine how this informs our understanding of statistical learning in reading and spelling acquisition. Our review shows that the current body of evidence remains relatively limited, with most artificial orthographic learning experiments focusing on orthography-phonology mappings in native English-speaking adults. Nevertheless, the paradigm provides valuable insights into how linguistic features such as consistency and frequency, as well as external influences such as explicit instruction shape learning. Findings from orthotactic learning experiments show that individuals can develop sensitivity to statistical regularities after brief exposure to artificial orthographic patterns, supporting the notion that orthographic learning is a form of statistical learning. However, concerns around validity and reliability of existing methods limit the strength of these conclusions. We highlight the need for future research to address these challenges through harmonising the methods for testing and measuring learning.

Chapter 3. Simultaneous Learning of Semantic and Graphotactic Regularities (Study 2)

Abstract

Orthographies such as English represent various levels of regularity such that spelling patterns are conditioned to varying extents on phonological, graphotactic and semantic cues. Previous research has shown that people learn graphotactic regularities (i.e., spelling rules governing possible combinations of graphemes) implicitly through exposure, and that this knowledge generalises to new forms. Using an artificial orthography learning task, we investigated whether adults could simultaneously learn both graphotactic regularities and semantic regularities (where possible spellings depend on grammatical word class). Results across two experiments showed that adults learned spelling patterns conditioned on semantics, especially when semantic information was more salient with nouns and verbs rather than adjectives and adverbs. This learning was associated with the ability to verbalise the patterns at post-test. In contrast to earlier work, there was no evidence of graphotactic learning. This suggests that learning the two types of nonphonological regularity simultaneously may not be possible given the short exposure to the artificial lexicon in our paradigm. These findings are discussed in terms of a statistical learning account of orthographic learning.

Introduction

English words are infamous for being difficult to spell. Although English is an alphabetic writing system, the relationship between phonemes (sounds) and graphemes (spelling) is not transparent, and there are multiple ways to spell a particular sound. Importantly, however, phoneme-grapheme mappings are not the only systematic patterns found in English. Consider graphotactic regularities, which concern the occurrence, positioning and sequencing of graphemes. For example, a word is more likely to end with a consonant doublet than start with one (e.g., “hill” vs. *hhil and “staff” vs. *sstaf). These regularities are partially related to phonology, but not always. A third type of pattern concerns morphological regularities. These connect meaning and orthographic form and can be unmarked in phonological form. For example, the letter sequence -ous is virtually always associated with an English adjective, yet its phonological counterpart /əs/ occurs with equal probabilities in adjectives and non-adjectives (Berg & Aronoff, 2017). Taken together, the availability of phoneme-grapheme mappings, graphotactic patterns and morphological cues means that the English writing system is rich with multiple types of regularity. From this perspective, learning to spell can be considered an endeavour in statistical learning with children becoming increasingly sensitive to regularities and quasi-regularities in their writing system, as their experience with written language grows.

Evidence in support of this perspective comes from artificial orthography experiments. In natural language, multiple cues are correlated but by using an artificial orthography, it is possible for language experience to be manipulated and controlled precisely, and for the effects on learning to be tracked. Previous studies have used this approach to investigate sensitivity to phoneme-grapheme mappings (e.g., Taylor et al., 2011), graphotactic constraints (e.g., Samara & Caravolas, 2014) and morphological regularities (e.g., Rastle et

al., 2021). Lacking, however, is an understanding of how learners deal with multiple cues in concert. In this study, we extended the use of the artificial orthographic learning paradigm to examine the simultaneous learning of two different types of nonphonological cue that concern graphotactics and form-to-meaning mappings, respectively. Additionally, we explored whether some participants become explicitly aware of these regularities without instruction and asked whether this awareness affects test performance.

Sensitivity to graphotactic patterns in written language

Broadly speaking, graphotactic regularities refer to the formal rules that govern permissible letter sequences and positions in a given orthography (Ferreiro & Teberosky, 1982; Lehtonen & Bryant, 2005; Venezky, 1967). As noted above, English words never start with double consonants such as “hh”, whereas French words must end with a vowel when preceded by a consonant doublet (e.g., “femme”, but not *femm). These formal rules are not always explicitly taught in school, yet children are sensitive to them from the early stages of reading acquisition (e.g., Pacton et al., 2001, 2013, 2014; Sobaco et al., 2015) and this knowledge can impact on how they spell words, beyond their knowledge of phoneme-grapheme mappings.

Graphotactic knowledge is typically measured using a variation of a pseudoword choice task. For example, Cassar & Treiman (1997) showed children (Kindergarten to 9th Grade) and adults (undergraduate students) pairs of pseudoword (e.g., “meer” vs. “miir” and “pess” vs. “ppes”) and asked them to select the one that was more like a real English word. The researchers hypothesised that if participants chose “meer” over “miir” and “pess” over “ppes”, this would suggest that they were aware of which letters in which position serve as doublets in English. Their results supported this hypothesis, and even first graders were able to differentiate between permissible and impermissible consonant doublets to a certain extent.

People also make decisions about spelling patterns based on other contextual cues, including the preceding vowel. This conditioning can be characterised as graphotactic in terms of vowel length (consonant doubling is more likely when preceding vowel is spelled with a single letter) and/or phonological in terms of vowel quality (consonant doubling typically occurs after short vowels such as /ɪ/, /ɛ/, /æ/, /ɑ/, /ʌ/, and /ʊ/). Using a pseudoword choice task, Hayes et al. (2006) found that English-speaking 7-11-year-olds and adults preferred consonant doublets when the preceding vowel was spelled with one letter (e.g., choosing "vaff" over "vaf"), and consonant singlets when the preceding vowel was spelled with more than one letter (e.g., choosing "vaif" over "vaiff"). In a production task, however, when they were asked to spell pseudowords to dictation, there was clear evidence that participants doubled the consonant spelling more often after short vowels than long vowels (e.g., spelling /θʌl/ as "thull" and /θul/ as "thool"). These findings are consistent with both the spelling of an English vowel and its sound influencing how people spell pseudowords that contain that vowel. Hayes et al. (2006) also found that grapheme choice depends on the presence of other graphemes, and that this could override patterns that are otherwise conditioned by phonological constraints (i.e., vowel length). Alongside work showing that people are sensitive to medial consonant doubling (Treiman & Boland, 2017), these findings demonstrate that graphotactic context influences spelling (see also Treiman & Kessler, 2016).

Sensitivity to meaning regularities in written language

Mappings between meaning and written form can provide systematic cues to aid spelling. Consider for example English morphology. Through a corpus analysis, Berg and Aronoff (2017) found that the spelling of English suffixes, specifically -ous, -al, -ic, and -y, provides a more consistent marker of lexical category than the corresponding phonological forms. For instance, as noted above, the phonological sequence /əs/ is found equally often in adjectives

(e.g., “nervous” /nɜ:vəs/) and non-adjectives (e.g., “cactus” /kæktəs/). In contrast, the written suffix -ous is predominantly used in adjectives, as in “hazardous” and “dangerous”. Similarly, the suffix spelling -ic provides a more reliable cue that a word is an adjective (as in “basic”, “historic” and “classic”) than the phonological sequence /ɨk/. Building on this work, Ulicheva et al. (2020) assessed the diagnosticity and specificity of all written derivational suffixes in English. *Diagnosticity* measures the extent to which a suffix spelling predicts a particular lexical category (e.g., -ous is almost entirely found in adjectives but not verbs). *Specificity* refers to the degree to which a particular spelling is the preferred choice for representing a given phonological sequence and specific lexical category. Diagnosticity and specificity were both high, confirming that suffix spellings are salient indicators for lexical category in English.

Given this consistency between meaning and orthographic form, an important question is whether people are sensitive to these regularities when they spell words. To address this, Ulicheva et al. (2020) asked English-speaking adults to classify and spell pseudowords. In the classification task, participants were shown pseudowords that contained English suffixes with high diagnosticity (e.g., adjectives: “cevable”, “dolous”, “tumish”; nouns: “tobness”, “jumer”, “nadence”) and were asked to categorise them as nouns or adjectives. They found that participants could accurately categorise pseudowords based on their suffixes. Moreover, diagnosticity influenced performance, that is, participants were more likely to categorise words with more adjective-biasing suffixes as adjectives. To assess spelling, participants were shown sentences that contained a missing word and were given the auditory form of the pseudowords and asked to type their spelling. An example sentence which contains a missing adjective is “The strong allegations proved < _/ɔɪtləs/_ > once the investigation was complete.” The spelling attempts tended to be appropriate for the anticipated lexical category

(e.g., -ous for the phonological sequence /əs/ in adjectives) and performance was graded depending on the specificity of the suffix spellings. However, it is important to note that Ulicheva et al. (2020) did not observe close-to-ceiling production of the -ous spelling for the phonological sequence /əs/; this would be expected if production mirrored the patterns observed in the corpus statistics (see also Heyer, 2021; Treiman et al., 2021). Heyer (2021) attributed this to interactions between phonological and nonphonological information. That is, to accurately produce the -ous spelling for an adjectival pseudoword ending with /əs/, spellers must first correctly segment the phonological sequence /əs/ from the auditory input and identify the missing word in the sentence context as an adjective, not a noun. This suggests that multiple cues and sources of information influence spelling choices.

Evidence from artificial orthography learning studies

Natural language is complex and it is difficult to isolate and assess the influence of a particular feature or cue. Studies using an artificial orthography offer a solution to this and in doing so, they provide an opportunity to test hypotheses about how learners track the occurrence and co-occurrences of letters and letter combinations as they experience a new writing system, and from this extract orthographic patterns. As the writing system is novel, the experimenter has complete control of the input statistics and learning and generalisation can be tracked as a function of different manipulations. Samara and Caravolas (2014) used this approach to examine the learning of graphotactic regularities. They created artificial words with a C_1VC_2 (i.e., consonant-vowel-consonant) structure. The words were designed with either positional constraints (i.e., some consonants will always appear in the C_1 position but never in the C_2 position, and vice versa) or contextual constraints (i.e., the legal position of consonants is based on the medial vowel). After a short exposure phase, participants were tested with a "surprise" generalisation test where they had to decide whether unseen words

written in the newly learned orthography “went well with” the words they had seen in the exposure phase. Children (aged 6-8 years) and adults performed above chance and showed sensitivity to both positional and contextual constraints. Notably, the extent of learning was affected by the complexity of patterns in the orthography. These findings show that people can learn graphotactic patterns from a short amount of exposure, consistent with a statistical learning account of spelling development.

Although there was no overt spoken language in their experiment, Singh et al. (2021) questioned whether the apparent graphotactic learning reported by Samara and Caravolas (2014) might in fact be underpinned by phonotactic learning. The items were pronounceable pseudowords where each grapheme corresponded to a distinct phoneme, leaving open the possibility that learners might be learning constraints based on phoneme position rather than letter position. To address this concern, Singh et al. (2021) capitalised on the phenomenon that English consonant singlets and doublets share the same pronunciation (e.g., “duf*”* and “duff” are both pronounced as /dʌf/) to create experiments where participants had the opportunity to learn an association between the medial vowel (“u” or “e”) and consonant doubling. For example, in their Experiment 1, if the medial vowel is “e”, the following consonants must be a doublet whereas if the medial vowel is “u”, the following consonant must be a singlet. They found that English-speaking children and adults learned the constraints on the occurrence of “f” versus “ff” and generalised this knowledge when spelling novel words. This cannot be explained by phonotactic learning since both corresponded to the same phoneme, and as shown by Samara and Caravolas (2014), the extent of learning depended on the complexity of the pattern.

An alternative way to examine pure graphotactic learning is to use novel symbols that are not pre-associated with any sounds (Chetail, 2017; Fernández-López & Perea, 2023). Chetail (2017) created 320 artificial strings using Phoenician Moabite, an artificial script that provides no phonological information, at least to their French-speaking participants. Participants were exposed to these strings which contained either two (Experiment 1a) or four bigrams (Experiment 1b). At test, participants completed a wordlikeness task, in which they were asked to choose which one of two letter strings was more like those in the exposure phase. They also completed a letter detection task in which they had to decide whether a particular letter had been in the strings in the exposure phase. Participants were above chance at choosing the target items in the wordlikeness task across several conditions, indicating sensitivity to the positions of letter clusters and letter co-occurrences, and in the letter detection task, participants were better able to spot letters that had higher frequency in the exposure phase. These findings demonstrate that even when an artificial script contains no phonological information, participants still learn graphotactic constraints.

Turning to learners' sensitivity to mappings between form and meaning, Lelonkiewicz et al. (2023, Experiment 3) used an incidental learning paradigm and found that adults learned the associations between novel affixes and semantic categories. Participants were exposed to 80 different letter strings, each comprising a string-initial affix and a random letter sequence. Before seeing each letter string, participants in the semantic category group were shown a picture representing an object from one of four semantic categories (clothes, food, mammals, or musical instruments). Each semantic category was associated with a specific affix in the letter sequence (i.e., "krv", "isq", "admw" and "cdhs"). To assess generalisation, participants were shown a novel picture from one of the four semantic categories and asked to indicate

which of the four novel strings matched the novel picture. Performance was above chance, suggesting that the association between affixes and semantic categories had been learned.

The studies reviewed thus far were designed to tap incidental learning and implicit statistical learning, yet it is possible that explicit learning mechanisms were deployed, at least for some participants. Consistent with this, Singh et al. (2021, Experiment 1) found that approximately 30% of adult participants could describe some of the patterns in the language, according to a post-experiment questionnaire. Notably, while 'unaware' participants showed above-chance generalisation performance, those who were 'aware' performed significantly better. To directly investigate the impact of explicit instruction, Rastle et al. (2021) exposed English-speaking adults to symbol-phoneme and symbol-meaning (semantic category) regularities in two artificial languages in a 10-day training programme. An explicit instruction group received direct instruction on both types of regularity before training, while an incidental learning group – which they call the discovery-learning group – received no instruction. After training, both groups were able to read aloud the trained items, and provide their meanings. However, those in the explicit instruction group showed better generalisation, both for symbol-phoneme and symbol-meaning mappings. In summary, learning experiments using novel stimuli (be they unfamiliar symbols or artificial words comprised of familiar letters) show that sensitivity to grapheme-phoneme mappings, form-meaning regularities and graphotactic patterns can be acquired via incidental exposure, to some extent. However, learning appears to be significantly stronger if participants are aware of the patterns (as indexed by their ability to verbalise them) and when they are given explicit guidance on the "rules" ahead of learning.

Another important issue to consider is the extent to which these artificial orthographic learning experiments reflect "real" learning. Previous work has shown that individuals with stronger reading and spelling abilities (as measured by standardised assessments) tend to be more sensitive to orthographic regularities in experimental tasks (Treiman et al., 2021; Treiman & Boland, 2017; Treiman & Kessler, 2006). This fits with the view that statistical learning is associated with learning to spell. Less consistent with this view are findings reported by Singh et al. (2021). In their artificial orthography study, there was no correlation between spelling ability and graphotactic learning in the implicit learning condition, although there was a correlation in the explicit learning condition. Similarly, Schmalz et al. (2021) also reported no reliable correlation between learning grapheme-phoneme mappings in an artificial orthography and individual differences in reading ability. Given this mixed pattern of results, another goal of our study was to examine whether the ability to learn statistical patterns in an artificial orthography is associated with variation in spelling ability.

In this study, we aimed to extend previous research to examine the simultaneous learning of two types of nonphonological patterns – semantic (mappings between visual form and meaning, instantiated in this study as word class category) and graphotactic (which concerns grapheme combinations) – in an artificial writing system. In Experiment 1, semantic patterns were formed by the last symbol of a word indicating its lexical category, nouns vs. verbs. Graphotactic patterns were formed between the medial vowel predicting the word-final symbol; stimuli in this condition were adjectives and adverbs but these word classes were not marked in the orthography. Experiment 2 built on the findings of Experiment 1 and was pre-registered (https://osf.io/j3ubk). It had two versions: Version 1 sought to replicate Experiment 1, and Version 2 reversed the pattern assignments (i.e., adjectives and adverbs had semantic patterns while nouns and verbs had graphotactic patterns), allowing us to examine the effect

of lexical category on learning. Within each experiment, we also investigated whether explicit awareness influenced learning outcomes. Finally, we probed individual differences by assessing spelling ability for English words and examined whether this was associated with patterns of learning in each experiment. Findings relating to spelling ability and its relationship with learning outcomes will be presented for both experiments after Experiment 2, with implications discussed in the General Discussion.

Experiment 1

This experiment examined whether adults can simultaneously learn semantic patterns (form-to-meaning mappings assigned to nouns and verbs) and graphotactic patterns (the medial vowel predicted the word-final symbols regardless of word class) in an artificial lexicon created from a semi-artificial set of graphemes. Participants first completed a self-paced reading task during the exposure phase, and learning was tested by a fill-in-the-blank task and word category task. At the end of the experiment, we also investigated whether any of the participants could verbalise the spelling rules embedded in the language, and if so, how this related to learning and task performance.

Method

Participants

Thirty-seven native English-speaking adults (26 female; mean age = 28.65; *SD* = 12.07) were recruited, all with self-reported normal or corrected-to-normal vision and no known neurological and learning impairments[6] (see Appendix 3A for additional information on

[6] All information for participants' backgrounds including their additional language profiles is available on OSF (https://osf.io/j3ubk).

participants' language backgrounds). Twenty participants were recruited through the online participant recruitment platform Prolific and 17 were recruited through the University's research participation scheme. Participants were either paid £6 or earned 3 course credits. Most of the participants ($N = 30$) had obtained or were currently pursuing an undergraduate or postgraduate degree at the time of the study; the other seven had either completed secondary school or entered other forms of tertiary education.

Materials

As summarised in Table 3.1, 144 CVC artificial words were created using four English vowels ("A", "E", "O", and "U") and 20 symbols (16 for C_1 and 4 for C_2 positions, taken from Taylor et al., 2011)[7]. For ease of description, we refer to the 16 C_1 symbols as "T", "P", "N", "S", "D", "M", "L", and "G" (used in the exposure phase only) and "B", "C", "F", "J", "K", "R", "Q" and "H" (used in testing phase only) while the C_2 symbols are referred to as "X", "Y", "Z" and "W" (see Appendix 3B for the full list of symbols). Out of the 144 words, 48 were used in the exposure phase to generate two lists of 32 words for counterbalance purposes shown in Appendix 3C.[8] Each word was assigned a meaning that corresponded to one of the four word classes (i.e., noun, verb, adjective and adverb). There were eight words of each word class type on each list.

[7] Our experiment used a semi-artificial script which incorporated novel symbols and English vowels. The rationale was to make the learning process less burdensome by having fewer unfamiliar symbols so that we could better evaluate emerging sensitivity to both types of pattern.

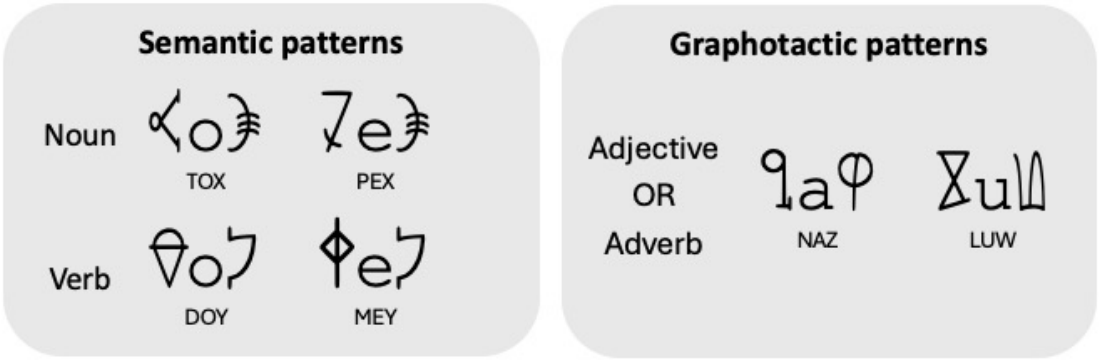
[8] Since there is more variability in the combination of medial vowels and second consonant in the semantic pattern (i.e., "OX", "OY", "EX", and "EY"), there is no overlap in the artificial words with semantic patterns across the two counterbalance lists. However, there are only two graphotactic patterns AZ and UW in our experiment. Only 16 artificial words with graphotactic patterns could be generated and they were all used in both lists but assigned to different word meanings.

Table 3.1. Artificial lexicon creation matrix for the exposure phase in Experiments 1 and 2.

Pattern	Lexical category		First consonant (C_1)	Medial vowel	Second consonant (C_2)
	Expt 1 and Expt 2 (V1)	Expt 2 (V2)			
Semantic	Noun	Adjective	T, P, N, S, D, M, L, G	O	X
	Noun	Adjective		E	X
	Verb	Adverb		O	Y
	Verb	Adverb		E	Y
Graphotactic	Adjective	Noun		A	Z
	Adverb	Verb		A	Z
	Adjective	Noun		U	W
	Adverb	Verb		U	W

Note. Only a subset of C_1 was used to create words in each lexical category in Experiment 1. To increase the variability of items, they had equal probabilities of occurrence in each lexical category in Experiment 2.

Figure 3.1. Examples of artificial words and sentences used in the exposure phase of Experiment 1.



Pattern	Example sentence
Semantic	Lily went to the store because she needed a new winter ᚲoᚷ.
Graphotactic	Ivy has a very ᚹaᚠ work schedule on Thursday because of the new project.

To create semantic and graphotactic patterns, we manipulated the combinations of the final consonant and (i) the medial vowel and (ii) the lexical category. Figure 3.1 shows the examples of artificial words used in the exposure phase of Experiment 1. For words that were semantically conditioned, the medial vowel was either “O” or “E” while the second consonant must be “X” for nouns and “Y” for verbs, respectively, regardless of the medial

vowel. For words that were graphotactically conditioned, the medial vowel could be either "A" or "U" while the second consonant must be "Z" following "A" and "W" following "U", regardless of lexical category. If participants were successful in learning the semantic patterns, they should be able to associate nouns with "X", and verbs with "Y" regardless of whether the medial vowel was "O" or "E". If they were successful in learning the graphotactic cues, they should be able to associate "A" with "Z", and "U" with "W", regardless of whether the word was an adjective or an adverb.

The other 96 words were used for the fill-in-the-blank task in the test phase, with two lists of 48 item pairs (one correct answer and one foil for each pair) shown in Appendix 3D. For counterbalancing purposes, participants were assigned to one of the two lists in the exposure phase and the corresponding list for the test phase.

A total of 208 sentences were created for the exposure phase (160 sentences; 8 words × 4 word classes × 5 repetitions) and the fill-in-the-blank task (48 sentences; 16 sentences × 3 trial types).

Procedure

The experiment was run online using Gorilla.sc (Anwyl-Irvine et al., 2020) and the link was distributed to participants either through Prolific (www.prolific.co) or via the University's Research Participation Scheme. Participants completed the experiment on their own computer or laptop. The experiment took approximately 45 minutes to complete.

Exposure phase

This comprised self-paced reading of 160 sentences (see Fig. 3.1 for example sentences). Participants were told that they would be reading sentences that contained alien words and that their task was to learn these words. The sentences were presented in 5 blocks with 32 sentences in each block (i.e., each artificial word appeared once in each block). Each sentence remained on the screen for at least 2000 ms before participants could press the "next" button to move on to the next sentence.

Test phase

Fill-in-the-blank task. This tested whether participants had learned the semantic and graphotactic patterns. We also looked at whether they showed a preference to use either pattern type when pitted against each other. In each trial, participants were presented with an English sentence that contained a missing word. Two artificial words were displayed beneath the sentence, and they were instructed to click on the option which they thought best fitted the sentence. As shown in Tables 3.2 and 3.3, there were three types of trial – semantic, graphotactic and preference. In the semantic trials ($N = 16$), the sentence contained either a missing noun or verb. The target word conformed to the semantic rule in Table 3.1 while the foil was identical but with the wrong C_2 (e.g., "BOX" (correct) and "BOY" (foil) for a missing noun trial). In the graphotactic trials ($N = 16$), the sentence contained either an adjective or adverb. The target word conformed to the graphotactic rule in Table 3.1 while the foil differed only in the C_2 that does not match the vowel (e.g., "BAZ" vs. "BAW"). In the preference trials ($N = 16$), neither option conformed to the rule, C_2 went with either the semantic or graphotactic patterns. An example stimuli pair for a noun context is "KAX" ("X" went with nouns but "A" was never a vowel for noun) and "KAZ" ("Z" went with "A" but "AZ" was never a combination for noun). Participants had to choose which one best fit the

sentence frames such as “Sarah tripped over a big _____ as she was not paying attention to where she was walking.”

Table 3.2. Artificial lexicon creation matrix for semantic and graphotactic trials in the fill-in-the-blank task in Experiments 1 and 2.

Trial Type	Lexical category		First consonant (C_1)	Medial vowel	Correct Ending	Foil Ending
	Expt 1 and Expt 2 (V1)	Expt 2 (V2)				
Semantic	Noun	Adjective	B, C, F, J, K, R, Q, H	O	X	Y
	Noun	Adjective		E	X	Y
	Verb	Adverb		O	Y	X
	Verb	Adverb		E	Y	X
Graphotactic	Adjective	Noun		A	Z	W
	Adverb	Verb		A	Z	W
	Adjective	Noun		U	W	Z
	Adverb	Verb		U	W	Z

Note. Only a subset of C_1 was used for creating words in each lexical category in Experiment 1. To increase variability of items, they had equal probabilities of occurrence in each lexical category in Experiment 2.

Table 3.3. Artificial lexicon creation matrix for preference trials in the fill-in-the-blank task in Experiments 1 and 2.

Trial type	Lexical category		First Consonant (C_1)	Medial Vowel	Option 1 (go with semantic pattern)	Option 2 (go with graphotactic pattern)
	Expt 1 and Expt 2 (V1)	Expt 2 (V2)				
Preference	Noun	Adjective	Experiment 1: B, C, F, J, K, R, Q, H Experiment 2: 1, 2, 3, 4, 5, 6, 7, 8	A	X	Z
	Noun	Adjective		U	X	W
	Verb	Adverb		A	Y	Z
	Verb	Adverb		U	Y	W

Note. Only a subset of C_1 was used for creating words in each lexical category in Experiment 1. To increase variability of items, they had equal probabilities of occurrence in each lexical category in Experiment 2. In Experiment 1, we used “B”, “C”, “F”, “J”, “K”, “R”, “Q” and “H” as first consonants when creating artificial lexicons for preference trials. However, this may be problematic as some items repeated in the graphotactic trials. Therefore, new symbols, referred to as “1”, “2”, “3”, “4”, “5”, “6”, “7” and “8”, were used in Experiment 2. Note that sentences for preference trials are different in version 1 and 2 in Experiment 2 because the missing words in each trial were of different word classes.

Word category task. This examined the extent to which participants inferred the class of the items seen in the exposure phase (independent of whether they learned that a spelling is conditioned on this). Participants were shown the 32 words from the exposure phase one at a time and were asked to provide an English word with the same/closest meaning. These were later scored as to whether the word provided matched the intended word class. Responses were scored as 1 if it was an English word in the same word class as the target. The total score was 32, with a maximum of 8 for each of the four word classes.

Awareness questionnaire. Following the test phase, participants completed an awareness questionnaire. This asked whether they (i) were aware of the purpose of the experiment, (ii) could explicitly verbalise any patterns and rules in the words, and (iii) used any strategies when choosing the words to fit into the sentence in the fill-in-the-blank task.

Test of English spelling. This was adopted from Andrews & Hersch (2010) and provided an estimate of spelling knowledge. Participants were shown a list of 88 words and told that half of the words were spelled incorrectly. They were instructed to check off all the words that they thought were spelled incorrectly. The incorrect items were formed by changing one to three letters of the word (e.g., “psycology” for “psychology”).

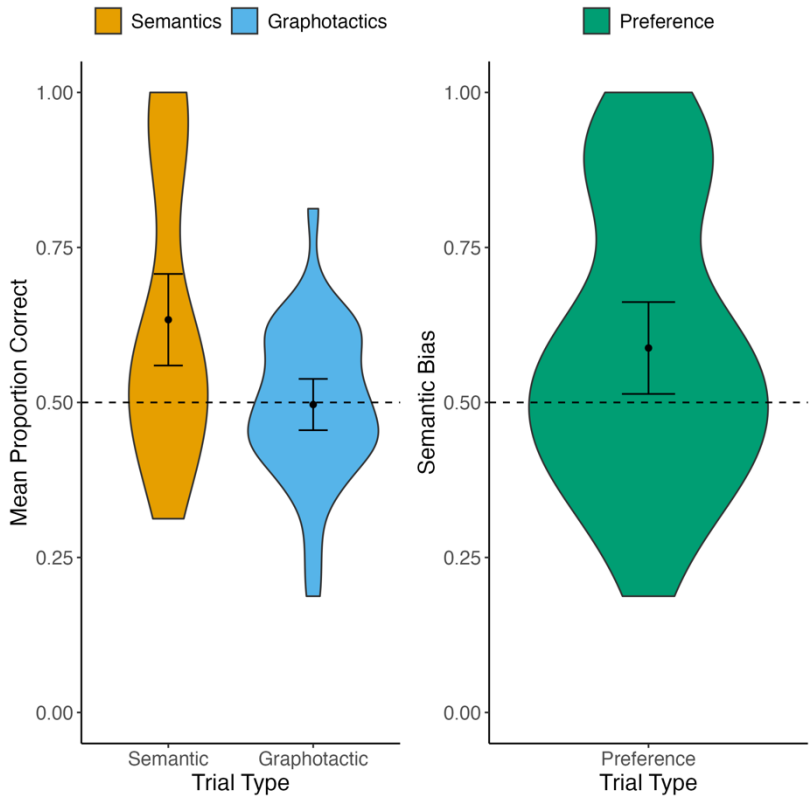
Results

Our results are organised as follows. We first present overall performance in the fill-in-the-blank task. We then examine whether participants had explicit awareness of the patterns and re-analysed their performance in the fill-in-the-blank task based on awareness status. Finally, we consider the relationship between performance on the fill-in-the-blank task and knowledge of word category. All data and analysis scripts are available on https://osf.io/jkzyt/.

Fill-in-the-blank overall performance

Descriptive statistics are provided in Table 3.4. Figure 3.2 (left) shows the mean proportion of correct responses in the semantic and graphotactic trials. In the preference trials (Figure 3.2, right), there is no "correct response" since participants were forced to choose between an item which is semantically correct but graphotactically incorrect, or vice versa. To generate a semantic bias index, we coded responses that went with semantic patterns as 1, and those with graphotactic patterns as 0 (so that an index over 0.5 suggests semantic bias, below 0.5 graphotactic bias). Statistical analyses used logistic mixed effect models (for details of model structure see Appendix 3E). For the semantic trials, participants were above chance (comparing to intercept of log odds 0 – i.e., 50%), which suggests successful learning of semantic patterns ($\beta = 0.73$, $SE = 0.21$, $z = 3.49$, $p < .001$). There was no evidence of learning for graphotactic trials ($\beta = -0.01$, $SE = 0.08$, $z = -0.16$, $p = 0.87$). In the preference trials, the semantic bias index was over 0.5 ($\beta = 0.44$, $SE = 0.18$, $z = 2.47$, $p < .05$), suggesting a bias to follow the semantic patterns.

Figure 3.2. Mean proportion of correct responses (semantic and graphotactic trials) and preference index (preference trials) in the fill-in-the-blank task in Experiment 1 (chance = .5).



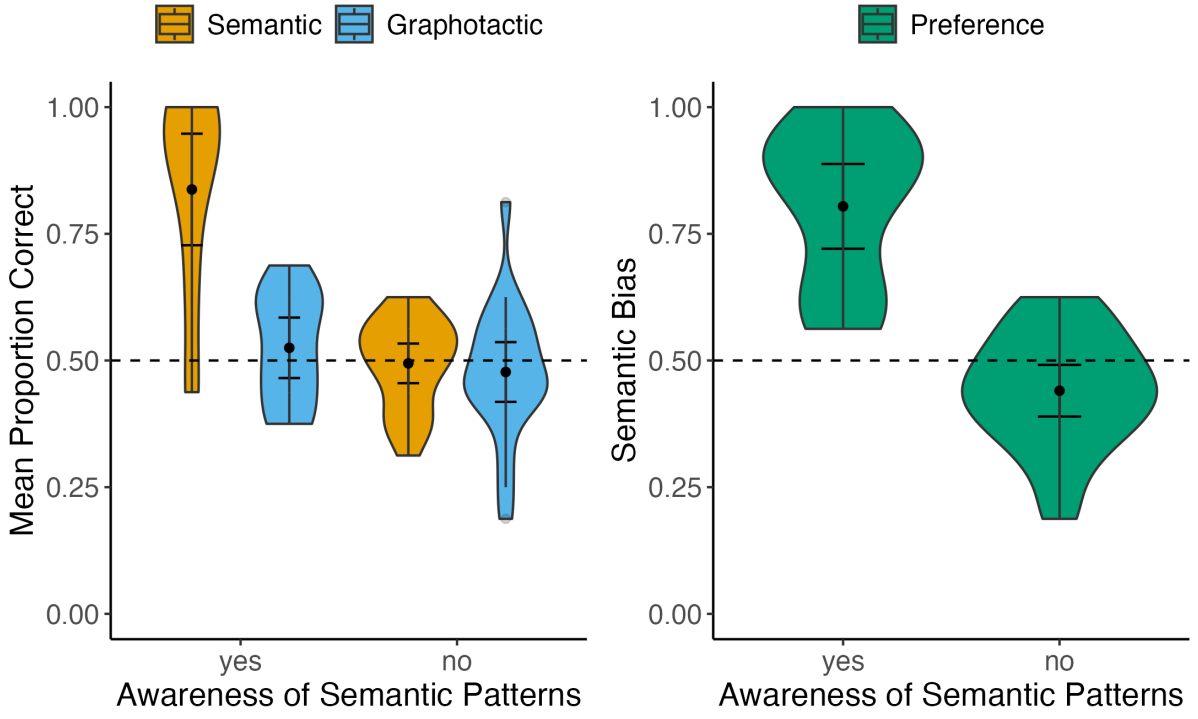
Note. Error bars show 95% confidence intervals.

Performance by awareness status

We coded the data from the awareness questionnaire as to whether each participant was aware of semantic and graphotactic patterns. Participants were considered as semantically aware (coded as 1) if they mentioned that a symbol or part of the artificial word was associated with a word class or semantic category (e.g., objects or actions). They did not have to specifically name the correct category for each symbol. Based on these criteria, we found 15 participants met this criterion. Some example responses of semantically aware participants were “Each word would be a different word class on the basis of this final letter” and “Some of the symbols were related to the type of word (i.e., noun, verb etc).” Participants were considered graphotactically aware if they described any association between the medial

vowel and the last symbol. No participants were coded as graphotactically aware according to this criterion.

Figure 3.3. Mean proportion of correct responses (semantic and graphotactic trials) and preference index (preference trials) by awareness status in the fill-in-the-blank task in Experiment 1 (chance = .5).



Note. Error bars show 95% confidence intervals.

Following Singh et al. (2021), we asked whether the overall above-chance learning of semantic patterns was driven by the subgroup of participants who were coded as semantically aware. We were also interested in whether and how explicit awareness of the semantics might interact with learning of the graphotactic cue. Figure 3.3 replots Figure 3.2 with participants split by semantic awareness. For semantic trials, there was strong evidence that semantic awareness was associated with accuracy (β = -1.82, *SE* = 0.29, *z* = -6.19, *p* < .001), and breaking this down, participants who were aware of semantic patterns were above chance (β

= 1.80, *SE* = 0.25, *z* = 7.26, *p* < .001) while there was no evidence of learning for semantically unaware participants, (β = -0.02, *SE* = 0.16, *z* = -0.15, *p* = 0.88). For graphotactic trials, there was no evidence that awareness of semantic patterns had a significant effect on accuracy (β = -0.19, *SE* = 0.17, *z* = -1.14, *p* = 0.25). For the preference trials, unsurprisingly, we found that awareness of semantic patterns had a significant effect on preference (β = -1.69, *SE* = 0.22, *z* = -7.55, *p* < .001). Semantically aware participants were significantly more likely to choose the semantic option than the graphotactic option, with an overall preference index significantly above 0.5 (β = 1.44, *SE* = 0.18, *z* = 7.83, *p* < .001). Interestingly, semantically unaware participants had an overall preference index below 0.5 indicating a graphotactic preference, and this was marginally significant (β = -0.24, *SE* = 0.13, *z* = -1.95, *p* = 0.05).

Relationship between learning and word category knowledge

As shown in Table 3.4, participants were able to provide some correct responses in the word category task. However, overall mean accuracy was only 43.59%. When analysed by lexical category, participants were more accurate for nouns (67.25%) and verbs (55.38%). Accuracy was relatively lower for adjectives (28.38%) and adverbs (23.25%). Performance in the word category task correlated with fill-in-the-blank performance for both semantic ($r_s(35) = .60$, $p < .001$) and preference trials ($r_s(35) = .68$, $p < .001$). This suggests that participants who were more successful in generalising their knowledge in the semantic trials and had a strong semantic bias in the preference trials also showed better learning of word meanings. This is unsurprising given that learning the former depends on learning the latter.

Table 3.4. Descriptive statistics for Experiments 1 and 2.

	Experiment 1				Experiment 2 (Version 1)				Experiment 2 (Version 2)			
Task (max)	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Fill-in-the-blank task												
Semantics (16)	10.14	3.54	5	16	9.87	3.14	3	16	8.47	3.24	1	16
Graphotactics (16)	7.95	1.99	3	13	8.30	2.12	4	13	8.04	1.81	4	13
Preference (16)	9.41	3.55	3	16	9.14	3.05	1	16	8.41	3.00	1	15
Word category task												
Overall (32)	13.95	6.82	3	26	11.78	5.50	0	27	11.00	4.94	1	26
Noun (8)	5.38	2.34	0	8	4.44	2.30	0	8	3.51	1.96	0	8
Verb (8)	4.43	2.50	0	8	4.13	2.16	0	8	3.22	1.88	0	7
Adjective (8)	2.27	2.06	0	7	2.19	1.68	0	7	2.57	2.05	0	7
Adverb (8)	1.86	1.81	0	6	1.03	1.36	0	5	1.70	2.12	0	8
Spelling test												
Overall (88)	76.65	7.36	55	87	75.83	7.49	56	87	74.53	9.82	48	88

Discussion

We examined sensitivity to semantic and graphotactic patterns embedded in an artificial orthography. Words were assigned to one of four lexical categories (i.e., noun, verb, adjective and adverb) as indicated by their occurrence in distributional contexts. Unbeknown to the participants, nouns and verbs were assigned consistent spelling patterns in which the last symbol of the word indicated its respective lexical category (semantic pattern). Adjectives and adverbs were not differentiated by a spelling pattern, but for these words, the medial vowel predicted the word-final symbol (graphotactic pattern). Overall, participants were above chance in the fill-in-the-blank task for semantic trials but not for graphotactic trials.

Responses on the post-experiment awareness questionnaire indicated that 41% of participants were able to report some knowledge of the semantic patterns (i.e., a symbol or a part of the artificial words was associated with a word class or semantic category), while none could report the graphotactic patterns. Further analyses showed that the above-chance learning of semantics – as indicated by above-chance performance on semantic trials and semantic bias on the preference trials – was driven by the subgroup of participants who were later able to describe the semantic patterns. The absence of graphotactic learning in our experiment is inconsistent with Singh et al. (2021) where participants learned graphotactic patterns regardless of awareness status. This may suggest that the simultaneous learning of both semantic and graphotactic patterns in the same artificial orthography and the same experiment impacts graphotactic learning. Some responses in the awareness questionnaire are consistent with this. For example, when describing the semantic patterns, some participants described the pattern in terms of a relationship between the last symbol of the artificial word and lexical categories in general, not specifically to nouns and verbs. It is therefore possible that once some participants discovered a semantic pattern, they might have tried to apply this

pattern to all artificial words in the exposure phase and subsequently failed to discover the graphotactic patterns.

In addition, there was an overall semantic bias in the preference trials, but again this was driven by the subgroup of participants who were semantically aware. Correlation analyses confirmed that participants who had higher accuracy in the semantic trials and stronger semantic bias in the preference trials had stronger scores in the word category task, consistent with the fact that learning of these patterns depends on identifying the word classes from the distributional contexts. For participants who were unable to describe the semantic patterns, there was no evidence that they differed from chance on the semantic trials, while on the preference trials they showed a trend (marginally significant, $p = .05$) for a graphotactic bias. This last result suggests the intriguing possibility that graphotactic learning only occurs in those participants who do not explicitly learn the semantics. However, it is important to note that there was no evidence of learning in the graphotactic trials in either group.

In summary, the results of this experiment suggest that learning semantic patterns is easier than graphotactic patterns. This may even occur at the expense of graphotactic learning. However, the semantic patterns in Experiment 1 were relatively simple and salient: they were conditioned on the lexical categories of *nouns* and *verbs* (i.e., associated with concrete objects and actions). The results of the word category task further supported this, showing that participants were more accurate at identifying nouns and verbs than adjectives or adverbs. This suggests that semantic patterns might have been easier to learn because of the constraints in our experimental set-up: semantic learning was only tested with nouns and verbs but not with adjectives and adverbs (which followed graphotactic patterns).

Experiment 2

To determine if semantic patterns were easier to learn in Experiment 1 because they were assigned to nouns and verbs, but not adjectives and adverbs, we conducted a pre-registered experiment (https://osf.io/j3ubk) that counterbalanced the assignment of lexical categories and pattern types. Experiment 2 contained two versions. Version 1 was a replication of Experiment 1, while in Version 2, we flipped the design so that semantic patterns were conditioned on adjectives and adverbs, and graphotactic patterns were nouns and verbs. We analyse the data using similar methods, however instead of frequentist *p*-values, we use the *Bayes Factor* as our inferential statistic – with priors informed by Experiment 1 and previous studies. This has the advantage that it allows us to distinguish between null and ambiguous results. The analyses reported below were conducted according to the pre-registered hypotheses and analysis plan, except where noted (see Appendix 3E).

Method

Sample size and power calculations

The most surprising and potentially important finding from Experiment 1 came from the preference trials. Although there was no evidence of graphotactic learning elsewhere, there was some evidence that participants who were not aware of semantic patterns showed a graphotactic bias (i.e., they chose the option that followed the graphotactic patterns more often than chance) in the preference trials. Given that this effect in Experiment 1 was small, we aimed to recruit a sufficiently large sample in the current experiment to determine whether this effect would replicate.

To estimate the required sample size, we adopted a Monte Carlo simulation approach. For a range of sample sizes (N = 25, 35, 45, 55, 65, 75, 85, 95, and 100), we generated 1000 datasets using the fixed and random effect parameters from the preference trials of Experiment 1. Based on the proportion of aware/unaware participants in Experiment 1, we randomly assigned 40% of participants to be aware of semantic patterns and the remaining 60% as unaware of semantic patterns in each simulation. These simulated datasets represented the data we might expect if H_1 were true and the same size as in Experiment 1. We then repeated this using the same parameters but with the intercept for unaware participants at 0 (chance), representing the case where H_0 was true. For each random dataset, we inspected the intercept for the unaware participants and calculated the Bayes Factor using the same method planned for the current experiment. For both H_1 and H_0 simulations, and for each sample size, we calculated the proportion of cases in which the Bayes Factor provided substantial evidence for both the alternative and the null hypotheses (i.e., BF > 3 for H_1, BF < 1/3 for H_0).

The simulation results suggested that 75 participants per version were needed to achieve over 80% power to detect the alternative hypothesis (i.e., BF > 3). On the other hand, over 300 participants per version were needed to reach 75% power to detect the null hypothesis (i.e., BF < 1/3). Since recruiting 300 participants in each version was impractical for this experiment, we decided on 75 participants per version (150 participants in total). This sample size provides over an 80% chance of accepting the alternative hypothesis and about 50% of rejecting the null hypothesis, if true. All simulation details, including the R script, are available in our pre-registration on OSF.

Participants

The selection criteria, recruitment, consent and compensation procedures were the same as Experiment 1. One-hundred and fifty-six native English-speaking adults (59 female; mean age = 36.30; *SD* = 13.61) were recruited through Prolific (*N* = 131) and the research participation scheme at the University (*N* = 25). Participants were randomly assigned to Version 1 and Version 2. Following the pre-registered data exclusion criteria, one participant was removed from the final analyses as the time taken in the exposure phase was over 3 *SD* from the mean. The final sample for analysis therefore included 155 participants, meeting the target specified in our pre-registration.

Materials

The design matched Experiment 1 except that there were two versions to counterbalance the assignment of lexical categories to semantic/graphotactic patterns. As summarised in Table 3.1, Version 1 was a replication of Experiment 1 while in Version 2, the semantic patterns were conditioned on adjectives and adverbs while graphotactic patterns occurred in nouns and verbs. The artificial words were the same across the two versions (see Appendix 3C and 3D for the full list).

Artificial words were created using the same sets of rules in Experiment 1, as described in Table 3.1. Minor changes were made to increase variations in the stimuli. First, each C_1 symbol from the list (“T”, “P”, “N”, “S”, “D”, “M”, “L”, “G”) had an equal probability of occurrence in each lexical category in Experiment 2. This is to ensure that participants would not be able to associate C_1, which was not the target symbol, with lexical categories. This was not strictly controlled in Experiment 1 as only four symbols were used as C_1 in each word class (e.g., “T”, “P”, “N” and “S” were the C_1 of nouns). Second, we created another 48

artificial words using a new set of eight C_1 symbols referred to as "1", "2", "3", "4", "5", "6", "7" and "8" for preference trials in Experiment 2. This is because some words in the graphotactic trials were also used in the preference trials in Experiment 1. The addition of new words was to ensure that performance was not affected by the repetition of words in the testing phase.

Similarly, sentence frames used across the two experiments were largely similar, except for two modifications. First, since preference trials were used for testing whether participants preferred semantic or graphotactic patterns, the missing words had to be of the same lexical categories as those assigned to the semantic pattern in the exposure phase. As semantic patterns were assigned to adjectives and adverbs in Version 2, we were unable to use the same sentence frames as in Version 1 (which were established for nouns and verbs). Sixteen new sentences with 8 missing adjectives and 8 missing adverbs were therefore created. Second, in the exposure phase, we modified three sentences to reduce repetitions of target meanings within sentence frames. For example, "Leo's miniplane…" in Experiment 1 was changed to "Leo's helicopter…" in Experiment 2, as the word "plane" was the intended meaning of a missing word in one of the fill-in-the-blank trials.

Procedure

This was identical to Experiment 1, with participants also completing the awareness questionnaire and a test of English spelling knowledge.

Statistical analyses

We used *Bayes Factors* (rather than *p*-values) as our inferential statistic in Experiment 2 to perform the Bayesian equivalent of significance testing. This has the advantage of providing

information that a *p*-value cannot: a "null" result (i.e., $p > 0.05$) does not tell us whether we have evidence for the null hypothesis, or no evidence for any conclusion at all (or even evidence against the null).

To compute Bayes Factors, we used the approach advocated by (Dienes, 2008, 2014) and Diene's calculator (implemented in R by Baguley & Kaye, 2010). This requires three numbers to test the hypothesis that two means are different. The first two are a data summary that comprises (i) the mean difference in the sample for the hypothesis in question (i.e., the difference between mean performance and chance or mean difference between two conditions) and (ii) an associated *SE.* Following Silvey et al. (2024), we ran logistic mixed effect models as in Experiment 1 and extracted the beta and *SE* values for the relevant coefficient in the model. Note that these are in log-odds space, meeting the normality conditions of the calculator. The third number required is a rough estimate of the predicted difference (i.e., predicted size of the beta) for the hypothesis, and this must also be in log odds space. Appendix 3E shows how these were estimated in the current work, based on estimates obtained from Experiment 1 and Singh et al. (2021). The predicted value is used as a parameter (or the "scale factor") in a model representing the plausibility of different effect sizes, if H_1 is true. We used a half-normal distribution with a mean of 0 and *SD* set to the scale factor. Dienes (2014) recommends this in the situation where a directional effect is predicted but only a ballpark estimate of the effect is available, and when smaller values are more likely. Note that we are thus testing a series of one-tailed predictions. The direction of testing will be made clear in the reporting.

Diene's calculator tests whether the data summary is more likely under this model of H_1 than under a model representing the null (using a point null where the only plausible effect is 0).

The result is a Bayes factor: a ratio representing the relative strength of evidence for H_1 versus the null. Values above 1 indicate more evidence for H_1 whereas values below 1 indicate more evidence for H_0. Bayes Factors can therefore be interpreted continuously, however for hypothesis testing, we refer to discrete evidential categories. We used $BF > 3$ as indicative of moderate/substantial evidence for H_1 and a $BF < \frac{1}{3}$ to indicate moderate/substantial evidence for H_0. Values in-between are interpreted as inconclusive evidence (i.e., the data is insensitive to test the hypothesis). Note that $BF > 3$ represents a similar level of conservatism to the more familiar $p < .05$, and these indicators very often align, though this is not guaranteed. *BF*s > 10 or $< \frac{1}{10}$ are considered strong evidence for the hypothesis or the null, respectively (Lee & Wagenmakers, 2014; Schönbrodt & Wagenmakers, 2018).

We report Bayes Factors using the notation $BF_{(0,x)}$ where *x* is the predicted value (scale factor) for the model of H_1. Since Bayes Factors are sensitive to the choice of values for x, and as there is some subjectivity in this choice, we also calculated "robustness regions" for each *BF*, reported as Robustness Region (*RR*) = [x1, x2]. These show the range of predicted values we could have used as the parameter (scale factor) for the model of H_1 and still have drawn the same conclusion, based on the cut-offs of $BF > 3$ or $BF < \frac{1}{3}$ [i.e., x1 and x2 represent how low/high a value we could have used and still obtained a *BF* which was greater than 3 ($BF > 3$), lower than $\frac{1}{3}$ ($BF < \frac{1}{3}$), or between 1/3 and 3 ($\frac{1}{3} < BF < 3$)]. Further details of both the mixed effects models and the computation of the Bayes factors are in Appendix 3E. Note that in addition to reporting Bayes Factors, we also report *p*-values since these are more familiar, but we do not interpret them.

Results

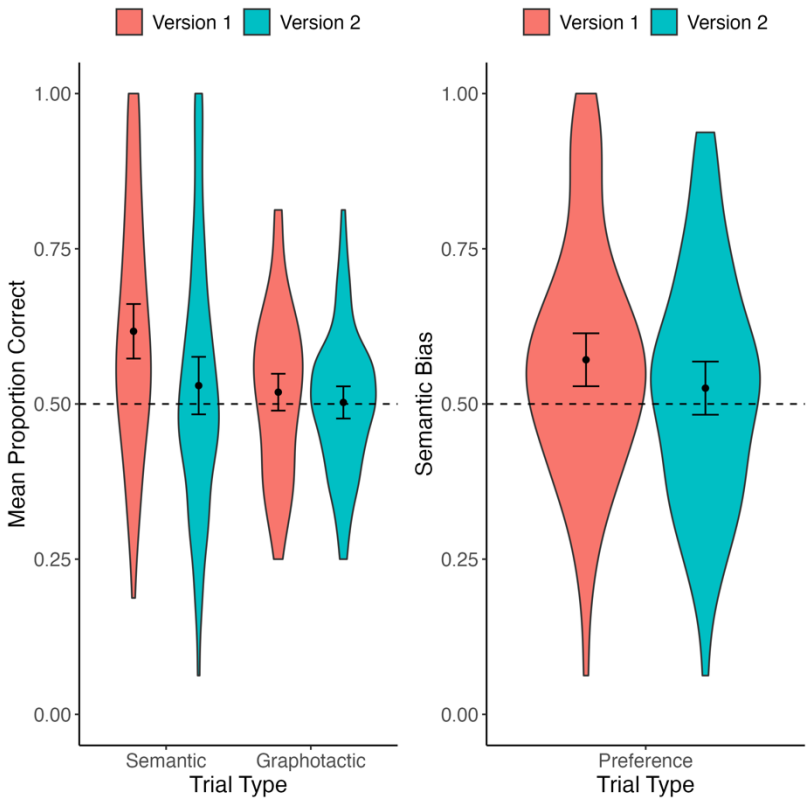
Following the pre-registered analysis plan, versions 1 and 2 were analysed separately. Version 1 is essentially a pre-registered replication of Experiment 1 (where semantic effects are conditioned on nouns and verbs) while version 2 tests whether the same effects are found when the semantic effects are instead conditioned on adjectives and adverbs. Two exploratory analyses were then conducted to gain a better understanding of semantic and graphotactic learning across both experiments. The findings from the exploratory are presented at the end of this results section.

Version 1: Fill-in-the-blank overall performance

Figure 3.4 shows the mean proportion of correct responses in the semantic and graphotactic trials (left) and their semantic bias in the preference trials (right). For semantic trials, there was strong evidence of above-chance learning, $BF_{(0,0.73)} = 106282.70$, *RR* [0.02, > 4.59] (model intercept: $\beta = 0.55$, $SE = 0.11$, $p < .001$). For graphotactic trials, the evidence as to whether performance was above chance was inconclusive, $BF_{(0,0.24)} = 0.94$, *RR* [0, 0.72] (model intercept: $\beta = 0.08$, $SE = 0.06$, $p = 0.20$). For preference trials, there was strong evidence that participants chose stimuli that went with semantic patterns above chance, $BF_{(0,0.44)} = 80.41$, *RR* [0.03, > 4.59] (model intercept: $\beta = 0.32$, $SE = 0.10$, $p < .001$).[9]

[9] We pre-registered an analysis for Experiment 2 to examine whether participants learn the semantic patterns better than the graphotactic patterns (i.e., main effects of trial types). We consistently found inconclusive/null findings for graphotactic learning in both versions of Experiment 2. These results are reported in the online supplementary materials and on OSF (https://osf.io/jkzyt/).

Figure 3.4. Mean proportion of correct responses (semantic and graphotactic trials) and semantic bias index (preference trials) in the fill-in-the-blank task in Experiment 2 (chance = .5).

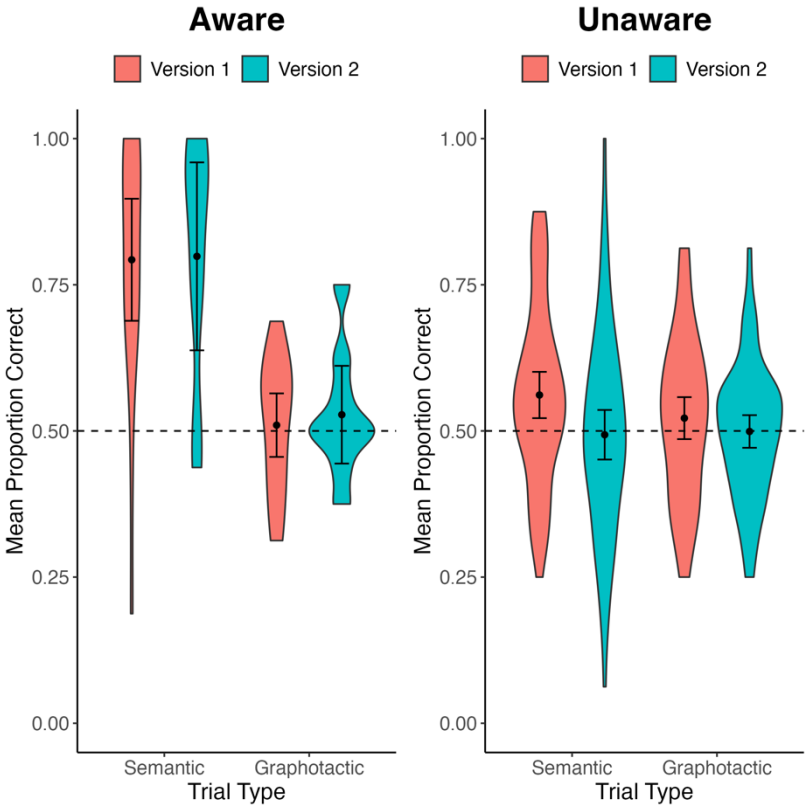


Note. Error bars show 95% confidence intervals.

Version 1: Performance by awareness status

We determined awareness status using the same criteria as in Experiment 1. Nineteen out of 79 participants were aware of semantic patterns whereas no participant was able to describe the graphotactic pattern (i.e., the association between the medial vowel and the last symbol). Therefore, the following analyses only focused on the association between semantic awareness and learning performance.

Figure 3.5. Mean proportion of correct responses (semantic and graphotactic trials) by awareness status in the fill-in-the-blank task in Experiment 2 (chance = .5).

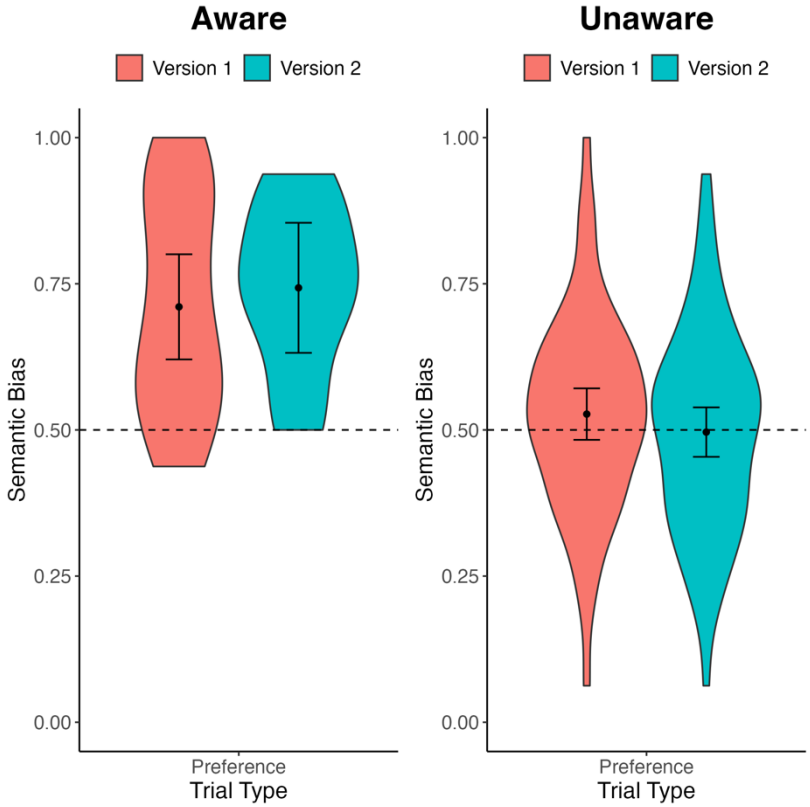


Note. Error bars show 95% confidence intervals.

Figure 3.5 replots the data from Figure 3.4 on performance in semantic and graphotactic trials, with participants split by semantic awareness. For semantic trials, there was very strong evidence that aware participants performed better than unaware participants, $BF_{(0,1.82)}$ = 240172.6, *RR* [0.05, > 4.59] (model coefficient: β = -1.23, *SE* = 0.23, $p < .001$). Note that the evidence for semantically aware participants being above chance was very strong, $BF_{(0,0.73)}$ = 9608846316, *RR* [0.03, > 4.59] (model intercept: β = 1.50, *SE* = 0.21, $p < .001$), and there was also substantial evidence for above-chance learning by the unaware participants, $BF_{(0,0.73)}$ = 7.29, *RR* [0.05, > 4.59] (model intercept: β = 0.27, *SE* = 0.10, $p < .01$). For graphotactic trials, we found moderate evidence against the hypothesis that semantically unaware participants have better performance than aware participants, $BF_{(0,1.69)}$ = 0.11, *RR* [0.54, >

4.59] (model coefficient: β = 0.05, *SE* = 0.14, *p* = 0.73). There was also inconclusive evidence as to whether either of the two groups differed from chance, respectively (semantically aware, $BF_{(0,0.24)}$ = 0.58, *RR* [0, 0.47] (model intercept: β = 0.04, *SE* = 0.12, *p* = 0.74); semantically unaware participants, $BF_{(0,0.24)}$ = 1.06, *RR* [0, 0.84] (model intercept: β = 0.09, *SE* = 0.07, *p* = 0.20).

Figure 3.6. Mean proportion of semantic bias in preference trials by awareness status in the fill-in-the-blank task in Experiment 2 (chance = .5).



Note. Error bars show 95% confidence intervals.

Figure 3.6 replots data from Figure 3.4 on the preference trials, with participants split by awareness of semantic patterns. For preference trials, we found strong evidence that semantically aware participants showed more semantic bias than the unaware participants, $BF_{(0,1.69)}$ = 835.27, *RR* [0.06, > 4.59] (model coefficient: β = -0.86, *SE* = 0.21, *p* < .001).

Breaking this down, there was strong evidence that semantically aware participants showed a semantic bias, $BF_{(0,0.44)}$ = 78803.84, *RR* [0.04, > 4.59] (model intercept: β = 0.98, *SE* = 0.19, $p < .001$), while the evidence for unaware participants was inconclusive, $BF_{(0,0.44)}$ = 0.75, *RR* [0, 1.03] (model intercept: β = 0.12, *SE* = 0.10, $p = 0.24$). Given the findings of Experiment 1, we also tested whether semantically unaware participants might show a graphotactic bias. To do this, we used Bayes Factors to test the hypothesis that their performance would be below 50% (i.e., a one-tailed test in the opposite direction). This showed evidence for the null, $BF_{(0,0.44)}$ = 0.11, *RR* [0.13, > 4.59], indicating that semantically unaware participants did not show a graphotactic bias in the preference trials.

Version 1: Relationship between learning and word category knowledge

Overall, participants achieved an average accuracy of 36.81% in the word category task (see Table 3.4 for full results). Performance follows the pattern seen in Experiment 1 with the highest accuracy for nouns (55.50%), then verbs (51.63%) and the lowest accuracy for adjectives (27.38%) and adverbs (12.88%).

To examine the relationship between performance in semantic and preference trials and word category knowledge, we report Spearman's correlations as in Experiment 1, and an associated Bayes Factor (the latter was not pre-registered due to an oversight when completing the pre-registration). Again, we tested hypotheses in a specific direction (one-tailed; i.e., testing for positive correlations) using a model of H_1 which was informed by an estimate of effect size for the effect. Details are given in Appendix 3F, but note that the computation involves computing fisher's *z* transformation of Pearson's *r* and the scale factor using an estimate of the predicted *r* with fisher *z*'s transformation applied. Predicted *r* values were taken from Experiment 1 wherever the relevant correlation effect was significant in that dataset. There

was strong evidence for a positive correlation between performance in the semantic trials and word category knowledge, $r_s(77) = 0.57$, $z_r = 0.65$, $BF_{(0,\ 0.69)} = 2062504$, *RR* [0.02, > 4.59]. The evidence for a positive correlation between the performance on the preference trials and word category knowledge was inconclusive, $r_s(77) = 0.20$, $z_r = 0.20$, $BF_{(0,\ 0.83)} = 1.24$, *RR* [0.23, 3.16].

Version 2: Fill-in-the-blank overall performance

As shown in Figure 3.4 (green violin plots), the evidence for whether participants performed above chance in semantic trials tended towards the null but was inconclusive, $BF_{(0,0.73)} = 0.63$, *RR* [0, 1.43] (model intercept: $\beta = 0.14$, $SE = 0.11$, $p = 0.18$). For graphotactic trials, we found moderate evidence against the hypothesis that participants were above chance, $BF_{(0,0.24)} = 0.27$, *RR* [0.19, > 4.59] (model intercept: $\beta = 0.01$, $SE = 0.06$, $p = 0.86$). For preference trials, the evidence that there was a semantic bias was inconclusive, though again it tended towards the null, $BF_{(0,0.44)} = 0.74$, *RR* [0, 1.02] (model intercept: $\beta = 0.11$, $SE = 0.10$, $p = 0.23$).

Version 2: Performance by awareness status

Nine out of 76 participants in Version 2 were aware of the semantic patterns and only one participant was able to describe the graphotactic patterns. We therefore conducted further analyses by semantic awareness only.

As shown in Figure 3.5 (green violin plots), for semantic trials, there was strong evidence that aware participants performed better than unaware participants, $BF_{(0,1.82)} = 48544.11$, *RR* [0.07, > 4.59] (model coefficient: $\beta = -1.56$, $SE = 0.32$, $p < .001$). Semantically aware participants were above chance in the semantic trials, $BF_{(0,0.73)} = 51621.47$, *RR* [0.06, > 4.59]

(model intercept: β = 1.53, *SE* = 0.30, $p < .001$) while there was moderate evidence against the semantically unaware performing above chance, $BF_{(0,0.73)} = 0.10$, *RR* [0.22, > 4.59] (model intercept: β = -0.03, *SE* = 0.10, $p = 0.78$). For graphotactic trials, there was evidence for the null hypothesis against the hypothesis that semantically unaware participants perform better than aware participants, $BF_{(0,1.69)} = 0.20$, *RR* [1.01, > 4.59] (model coefficient: β = -0.13, *SE* = 0.18, $p = 0.48$). The evidence that semantically aware participants generalised correctly was inconclusive, $BF_{(0,0.24)} = 0.99$, *RR* [0, 0.97] (model intercept: β = 0.12, *SE* = 0.17, $p = 0.47$) and for semantically unaware participants there was substantial evidence against this hypothesis, $BF_{(0,0.24)} = 0.24$, *RR* [0.16, > 4.59] (model intercept: β = -0.004, *SE* = 0.06, $p = 0.94$). For preference trials, there was strong evidence for H_1 in that semantically aware participants had a stronger semantic bias than the unaware participants, $BF_{(0,1.69)} = 2659.72$, *RR* [0.07, > 4.59] (model coefficient: β = -1.15, *SE* = 0.27, $p < .001$). For aware participants, we found strong evidence that they had a semantic bias, $BF_{(0,0.44)} = 1921.52$, *RR* [0.06, > 4.59] (model intercept: β = 1.14, *SE* = 0.25, $p < .001$), while there was evidence against the hypothesis that unaware participants showed a semantic bias, $BF_{(0,0.44)} = 0.17$, *RR* [0.21, > 4.59] (model intercept: β = -0.02, *SE* = 0.09, $p = 0.86$). As in version 1, we tested whether semantically unaware participants showed a graphotactic bias: there was evidence for the null for this hypothesis, $BF_{(0,0.44)} = 0.22$, *RR* [0.29, > 4.59].

Version 2: Relationship between learning and word category knowledge

As shown in Table 3.4, mean accuracy was 34.38% in the word category task. Despite reversing the word class assignments of semantic and graphotactic patterns, the pattern of performance was similar to version 1, and to the pattern reported in Experiment 1, with the highest accuracy observed for nouns (43.88%), and then verbs (40.25%) and lowest for adjectives (32.13%) and adverbs (21.25%).

In examining the relationship between performance in semantic and preference trials and word category knowledge, we found evidence supporting the null hypothesis for a positive correlation between performance in the semantic trials and word category task, $r_s(74) = 0.06$, $z_r = 0.06$, $BF_{(0,\ 0.69)} = 0.26$, *RR* [0.53, > 4.59]. However, there was strong evidence for a positive correlation between semantic bias in the preference trials and performance in the word category task, $r_s(74) = 0.40$, $z_r = 0.42$, $BF_{(0,\ 0.83)} = 162.82$, *RR* [0.04, > 4.59].

Exploratory analyses comparing Versions 1 and 2

Our pre-registration stated that we would not combine or compare versions in the initial analyses, but that we might do this in a targeted way to further understand our data.

Comparing the results across the two versions of Experiment 2, one potentially important difference concerns the strong evidence for learning semantic patterns in Version 1 (*BF* = 106282.70) whereas in Version 2, the evidence was inconclusive (*BF* = 0.63). To see whether performance across the two versions is reliably different, we combined the data to test the main effect of Version.[10] There was strong evidence for this main effect, $BF_{(0,0.35)} = 16.86$, *RR* [0.07, 3.75] (model coefficient: $\beta = -0.40$, $SE = 0.15$, $p < .01$). In other words, even though we cannot conclude that there was no semantic learning in Version 2 (because there was no evidence for the null), we can conclude that there is less semantic learning in Version 2 than in Version 1.

We also considered the evidence against graphotactic learning in the graphotactic trials in each version. In Version 2, we found evidence for the null for graphotactic learning but in

[10] We set the predicted value for the difference between conditions to be the grand mean (i.e., the intercept of the model). This value was not pre-registered, however see Silvey (2024) for an explanation of why the grand mean is a reasonable estimate of effect size for a main effect in some contexts.

Version 1, the evidence was inconclusive. We did not conduct Bayes analyses in Experiment 1, but since the two experiments are similar, and Bayes Factors remain a valid measure of evidence from a combined sample (and if there is truly evidence for H_1, the strength of evidence, and thus the Bayes Factor, should be larger in a larger sample, and similarly, the Bayes Factor will be smaller if the strength of evidence is larger for H_0), we conducted a further analysis over the pooled data from the graphotactic trials from Experiment 1 and Experiment 2 Version 1, to see if there is evidence for the null. This was still inconclusive, $BF_{(0,0.24)} = 0.53$, *RR* [0, 0.39] (model intercept: $\beta = 0.05$, $SE = 0.05$, $p = 0.32$), although the Bayes Factor tended more towards the null than the analysis on data from Experiment 2 Version 1 alone.

Discussion

The findings of Experiment 2 replicated and extended those of Experiment 1. The chief addition was to counterbalance the assignment of lexical categories and pattern types. This allowed us to examine the extent to which the pattern seen in Experiment 1 – i.e., learning of the semantic rules but not graphotactic rules – depended on the lexical categories over which the semantic rules operated. We also employed Bayes Factor analyses with the goal of being able to distinguish between null and ambiguous results.

As predicted, the results of Version 1 were largely similar to those of Experiment 1. There was strong evidence that participants learned the semantic patterns, as shown by overall above-chance performance in the semantics trials and a semantic bias in the preference trials. As in Experiment 1, the awareness questionnaire confirmed that some participants (24%) were aware of the semantic patterns, and, as before, there was strong evidence that these participants showed stronger performance in the semantic trials and a stronger semantic bias

in the preference trials than unaware participants. However, in contrast to Experiment 1, when we looked at the performance of participants categorised as “aware” and “unaware” separately, we found moderate evidence that even the unaware group showed some sensitivity to the semantic patterns (i.e., moderate evidence for above-chance performance in semantic trials). We found no evidence of learning of the graphotactic patterns, although the Bayes Factor was around 1 meaning the evidence was highly inconclusive (i.e., no evidence of learning but also no evidence of no learning). We also looked to see if, in line with the trend seen in Experiment 1, there was any evidence that semantically unaware participants showed a graphotactic bias in the preference trials. This pattern did not replicate, and instead we found moderate evidence for the null indicating that the trend in Experiment 1 was likely a chance finding.

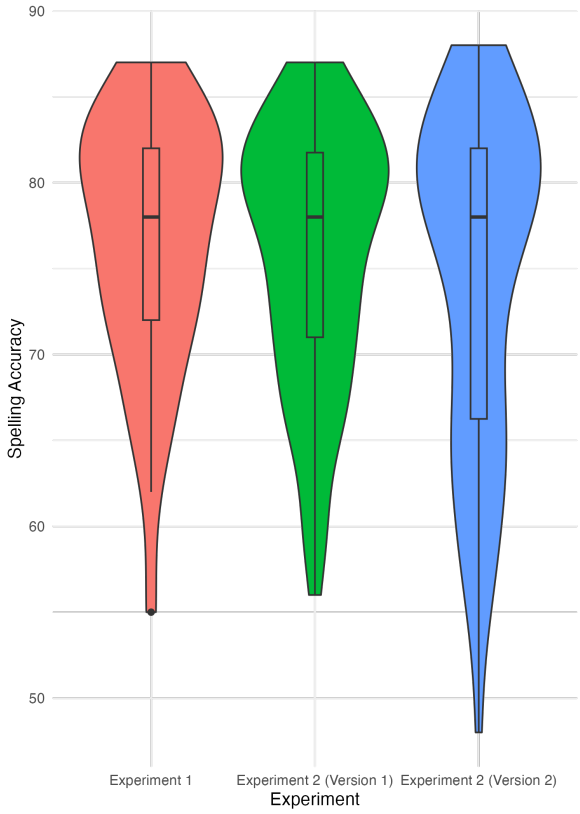
In the second version of Experiment 2, we switched the assignment of lexical categories and pattern types so that adjectives and adverbs were assigned to semantic patterns and nouns and verbs to graphotactic patterns. Overall, we did not find any evidence of semantic learning, although the data is inconclusive both with respect to whether there was above-chance learning in the semantic trials and whether there is a semantic bias in the preference trials. Replicating the previous experiments, some participants reported awareness of the semantic patterns, and there was evidence that the performance of these participants differed from those who were not aware. Specifically, when we looked at the “aware” participants in isolation, there was strong evidence that they were above chance on the semantic trials, and that they showed a semantic bias in the preference trials. This suggests that at least some participants learned the semantic patterns. Turning to the graphotactic patterns, there was no evidence of learning in the graphotactic trials, and here there was moderate evidence for the null. Replicating the null findings of Version 1, semantically unaware people did not show a

graphotactic bias in the preference trials, with again moderate evidence to the null, adding further weight to this observation in Experiment 1 being spurious.

Experiments 1 and 2: Relationship with English Spelling Ability

In both experiments, participants completed a spelling recognition task adopted from Andrews and Hersch (2010). The rationale for this was to examine the relationship between their ability to learn statistical patterns in an artificial orthography and their spelling ability in English. Participants were presented with 88 words and told that half were spelled incorrectly, and that their task was to indicate the ones with incorrect spellings. Descriptive data are summarised in Table 3.4. Three participants in Experiment 2 obtained very low scores that were more than 3 *SD* away from the mean. Inspecting these cases, it looks as if these participants misread the instructions and marked the correct items, rather than the incorrect ones. Although not pre-registered, we decided to remove these three participants from the analyses. Figure 3.7 shows the distribution of spelling accuracy scores across Experiments 1 and 2. Across the two experiments, the mean accuracy is similar at approximately 85% (Experiment 1: $M = 76.65$, SD = 7.36; Experiment 2 Version 1: $M = 75.83$, $SD = 7.49$; Experiment 2 Version 2: $M = 74.53$, $SD = 9.82$).

Figure 3.7 Distribution of spelling accuracy scores across Experiments 1 and 2.



To examine the relationship between individual differences in spelling and learning in the experiments, we conducted Spearman's correlations and computed Bayes factor for each analysis. The Bayes model for H_1 used values from Singh et al. (2021, Experiment 3; z_r = .46 for the correlation between spelling ability and learning as assessed by legality judgement in their experiment), as detailed in Appendix 3F. Note that this value was not pre-registered.

Table 3.5. Correlations between participants' performance in the fill-in-the-blank task, word category knowledge and spelling ability in Experiments 1 and 2.

Experiment and trial type	Statistics	Spelling
Experiment 1 (*N* = 37)		
Semantic trials	$BF_{(0,0.46)}$	**0.27[a] [0.36, >5.59]**
	p	.71
	z_r (r_s)	-0.06 (-0.06)
Preference trials	$BF_{(0,0.46)}$	**0.29[a] [0.39, >4.59]**
	P	.80
	z_r (r_s)	-0.04 (-0.04)
Word category	$BF_{(0,0.46)}$	0.47 [0, 0.68]
	P	.71
	z_r (r_s)	0.06 (0.06)
Experiment 2 – Ver 1 (*N* = 78)		
Semantic trials	$BF_{(0,0.46)}$	1.50 [0, 2.29]
	P	.11
	z_r (r_s)	0.18 (0.18)
Preference trials	$BF_{(0,0.46)}$	0.60 [0, 0.87]
	P	.35
	z_r (r_s)	0.11 (0.11)
Word category	$BF_{(0,0.46)}$	**4.22[b] [0.08, 0.71]**
	P	.03
	z_r (r_s)	0.25 (0.24)
Experiment 2 – Ver 2 (*N* = 74)		
Semantic trials	$BF_{(0,0.46)}$	0.81 [0, 1.20]
	P	.25
	z_r (r_s)	0.14 (0.14)
Preference trials	$BF_{(0,0.46)}$	**3.71[b] [0.09, 0.61]**
	P	< .05
	z_r (r_s)	0.25 (0.24)
Word category	$BF_{(0,0.46)}$	**192[b] [0.04, >4.59]**
	P	< .001
	z_r (r_s)	0.42 (0.40)

[a] Substantial evidence for H_0

[b] Substantial evidence for H_1

Note. We did not perform the correlation analyses with graphotactic trials as there was no conclusive evidence for above-chance graphotactic learning in either Experiment 1 or 2.

As summarised in Table 3.5, there was no evidence for an association between spelling ability and performance in semantic trials in either of the two experiments. Where evidence was conclusive, it was moderate evidence for the null. We found moderate evidence for a

positive correlation between spelling ability and a tendency to show a semantic bias in the preference trials in Experiment 2 Version 2, but moderate evidence for the null for the same correlation in Experiment 1 alongside inconclusive evidence in Experiment 2 Version 1. Finally, there was evidence for a positive correlation between spelling ability and performance in the word category knowledge in both versions of Experiment 2, yet evidence towards the null in Experiment 1. Taken together, these results generally suggest that there was no meaningful association between spelling ability and learning of orthographic patterns in the experiments. There was however some evidence to suggest that better spellers were better at inferring and recalling word categories in both versions of Experiment 2, alongside inconclusive evidence from Experiment 1. We will return to discuss the implications of these findings in General Discussion.

General Discussion

Using both frequentist statistics (Experiments 1 and 2) and Bayes factors (Experiment 2), four key results emerged from our experiments. First, learning of spelling patterns conditioned on semantic patterns (as captured by regularities between form and meaning) happened, as seen in the above-chance performance in the semantic trials, and evidence of semantic bias in the preference trials. This learning was stronger when patterns were conditioned on nouns and verbs (Experiment 1 and Experiment 2 Version 1) rather than adjectives and adverbs (Experiment 2 Version 2). Second, and contrary to previous studies (e.g., Singh et al., 2021), we found no evidence of graphotactic learning in any of the two experiments. The results were non-significant or inconclusive, other than in Experiment 2 Version 2 where there was evidence for the null. Third, where learning did occur, it was stronger in participants who reported some awareness of the form-to-meaning mappings in the post-test questionnaire. Fourth, there was no evidence that learning form-to-meaning

mappings in our experiments is related to spelling ability, though we note that some results were inconclusive, according to Bayes analyses. We discuss the implications of these findings below.

Statistical learning of form-to-meaning mappings

The English orthography contains regularities that go beyond phoneme-grapheme mappings. One additional consistency that helps inform spelling relates to morphological knowledge and associated mappings between meaning and orthographic form (Berg et al., 2014). Computational analyses of English corpora show that English derivational markers are salient markers of lexical category (Berg & Aronoff, 2017; Ulicheva et al., 2020). For example, the written suffix -ous is strongly associated with adjectives. This means that even though an individual might not be familiar with words such as "seditious" and "loquacious", they may be able to infer their adjectival nature from the -ous ending. Behavioural measures have provided consistent evidence that adults are sensitive to these form-meaning regularities in English and that they use this knowledge to help choose between alternative spellings (Berg & Aronoff, 2017; Treiman et al., 2021; Ulicheva et al., 2020, 2021).

Nevertheless, our understanding of how orthographic form-meaning regularities are learned is limited. Previous studies have largely focused on whether children and adults can infer the lexical categories of pseudowords, based on their suffix endings, or if they use specific suffix spellings when asked to spell pseudowords that contain phonological sequences such as /əs/ and /ɨk/. While this evidence indicates that experienced spellers are aware of form-meaning regularities in their writing system, it does not explain how individuals develop sensitivity to these regularities. Therefore, one of our main goals was to examine whether participants could learn form-to-meaning associations via a short amount of exposure to an artificial

orthography containing multiple patterns. Successful learning of these regularities would support a statistical learning account of orthographic learning.

Taken as whole, our results showed that participants were able to learn spelling patterns that were conditioned on lexical category. These findings align with previous studies showing that adults are sensitive to form-meaning regularities in written language. They further suggest that for some people at least they can be learned relatively rapidly, following a brief period of exposure. In our experiments, orthographic patterns were independent of phonology, indicating that people can directly connect spellings to meanings, just as other work has shown that meaning-to-form mappings can help individuals choose among alternative spellings for a specific phonological sequence (e.g., Treiman et al., 2021).

Although form-to-meaning mappings were learned, this learning was not straightforward. Across all experiments, we consistently observed more semantic learning with nouns and verbs, and less with adjectives and adverbs. This suggests that learning was stronger (and possibly limited to) the more salient contexts where semantic patterns were associated with nouns and verbs. Why might this be? One plausible explanation could be that nouns and verbs are critical to the grammaticality of a sentence whereas adjectives and adverbs are optional components (Huddleston & Pullum, 2002). Consider the sentence context "Ava does not allow anyone to touch anything on her  ." Omitting the artificial word would make the sentence grammatically incorrect and therefore prompt participants to consider potential words that can follow the pronoun *her*. In contrast, when the artificial word is an adjective such as *beautiful* in the sentence "Ava loves her garden where she has planted many  flowers and plants," the sentence remains grammatically correct even without the adjective. Thus, participants may allocate less attention to the corresponding meaning of the artificial

word, and its lexical category, which in turn makes it harder to make mappings between form and meaning. This interpretation is consistent with conclusions drawn by Heyer (2021) and Treiman et al. (2021). When analysing how suffix spellings were used when participants were asked to "fill in the blank" in a sentence context by writing a pseudoword, they argued that correct identification of the lexical category of the pseudoword is needed as a *prerequisite* for the successful application of semantic knowledge when spelling a novel word.

Despite overall above-chance learning, it is clear from the distribution in each experiment that many participants did not learn the semantic patterns, regardless of which lexical categories were used in the experimental manipulation. Why did some people learn the form-to-meaning mappings whilst others did not? One possibility is that some participants simply struggled with using nonphonological information from the sentence context. As the artificial orthography was deliberately designed to exclude phonological information, participants needed to infer the meanings of the artificial words through sentence context in order to learn the semantic patterns. This contrasts with learning English spelling, where phonological information is primary, rather than form-meaning regularities. This type of challenge has been observed in other pseudoword spelling studies. As discussed in the Introduction, analyses of the English language demonstrate a strong association between the phonological sequence /əs/ and suffix spelling -ous in adjectives, yet the actual usage of -ous in spelling /əs/ in pseudowords within adjectival contexts is much lower than predicted, even in simple sentence contexts (Heyer, 2021; Treiman et al., 2021; Ulicheva et al., 2020). Treiman et al. (2021) suggested that this might be because people rely on phonological information to spell pseudowords, rather than using knowledge that can be drawn from sentential context, such as inferring lexical category and from this, choosing to use the -ous spelling pattern. Although

there was no phonological information about the artificial words in our experiments, participants will bring their language experience with them to the task in hand. This may include a bias to rely on phonological information when learning new word forms and this may reduce the opportunity for regularities concerning lexical categories to be learned.

Absence of graphotactic learning

Based on previous studies (Samara et al., 2019; Samara & Caravolas, 2014; Singh et al., 2021), we expected participants to learn some graphotactic patterns (i.e., the associations between medial vowels and lexicon-final symbols). Surprisingly, we found no evidence of this type of learning, with evidence tending towards the null in both experiments and crossing the threshold of evidence for the null in Version 2 of Experiment 2. Perhaps not surprisingly given this absence of learning, only one person from all 192 participants was able to describe the graphotactic patterns in the post-test questionnaire. Why did participants fail to learn the graphotactic patterns in our experiments? One major difference between our design and previous studies on graphotactic learning is that we presented graphotactic patterns in artificial words within meaningful sentences, whereas prior studies presented the artificial words in isolation. It is therefore possible that people attended more to word meaning in our experiment that this served to reduce attention to the graphotactic patterns.

Another plausible explanation is that learning graphotactic patterns, especially when presented with another type of pattern, requires more repetition. In our experiment, graphotactic patterns were embedded in 16 unique artificial words, each repeated 5 times during the exposure phase. Thus, participants only saw the graphotactic patterns for a total of 80 times. This number of repetitions is much lower than in previous work. For example, Singh et al. (2021) also had 16 unique artificial words as their exposure items, but each word

was repeated for 18 times, resulting in a total of 288 exposures to the graphotactic patterns. The contrast between the robust graphotactic learning in Singh et al. (2021) and the absence of such learning in our experiments suggests the importance of repetitions in graphotactic learning. That is, the number of repetitions or the frequency with which people encounter these strings may be crucial for developing sensitivity to graphotactic patterns, especially under learning conditions where two types of pattern exist in the artificial lexicon. These results might potentially reflect how orthographic learning in the real world is protracted, especially when multiple imperfect cues are present. Future studies should use a longer exposure phase with more exemplars of the graphotactic patterns to examine whether simultaneous learning of nonphonological patterns is possible via implicit exposure.

Limits of statistical learning

Our data, along with other studies using both natural language and artificial orthography, demonstrate that people become sensitive to regularities and quasi-regularities in the writing system as they gain more experience with it. This suggests that statistical learning may underpin the process of learning to spell. While our findings generally support this view, they also highlight some limits to learning multiple types of pattern via exposure to the statistics only, namely that: (1) while it was possible to learn form-to-meaning mappings, this was dependent on the salience of word meanings and (2) learning graphotactic patterns embedded in meaningful context was challenging.

Similar limits in learning two types of pattern via exposure were observed by Rastle et al. (2021), although their experiment focused on symbol-phoneme and symbol-meaning mappings. In their experiment, despite 10 days of extensive training and a much more varied training paradigm, only 80% of participants in the discovery-learning group (i.e., learning via

exposure) achieved 75% accuracy in generalising symbol-phoneme mappings. Only 33% reached this level of accuracy in generalising symbol-meaning mappings and nearly half of the group were at chance. This is much lower than in the explicit instruction group. These findings clearly show that learning via exposure is slow and does not guarantee successful learning of regularities in written languages, especially when there is more than one pattern in the input.

In addition to the difficulty of learning two types of pattern, our results further revealed how explicit awareness influences whether participants could generalise their knowledge of these newly learned patterns. By separating participants into aware and unaware groups based on their ability to verbalise the patterns in the post-experiment questionnaire, we observed that the aware group outperformed the unaware group. In fact, only in Experiment 2 Version 1 did both groups demonstrate above-chance performance in semantic trials, though with a marked difference in the mean accuracy between the two groups (semantically aware: 81.82%; semantically unaware: 56.73%). This finding is consistent with other learning studies where participants learned graphotactic constraints. In Singh et al. (2021), for example, unaware participants achieved about 60% accuracy in a fill-in-the-blank task while aware participants were close to ceiling. Furthermore, while some participants were able to verbalise the patterns learned through exposure, the proportion of those able to do so remains relatively small. In both our study and in Singh et al.'s, only approximately 30% of adult participants were able to describe the patterns. These findings highlight that developing explicit awareness of the orthographic patterns is not guaranteed when learning via a short amount of exposure, but that in our experiments at least it was crucial for successful generalisation. We determined awareness based on responses in the post-experiment questionnaire and it is important to acknowledge that being unable to verbalise patterns does not unnecessarily

equate to being unaware of them, especially when novel symbols are used as stimuli. Nevertheless, aware participants outperformed their unaware peers across our experiments. This shows that while some "unaware" participants may have noticed some regularities in the input, being able to verbalise the patterns learned through exposure was associated with better performance on the generalisation task. Like Singh et al. (2021), we recognise that a post-experiment questionnaire does not provide precise information as to how and when explicit awareness emerges. As suggested by Singh and colleagues, it is possible that aware participants were actively searching for patterns and testing their guesses against later exemplars during testing, and/or that they only became aware of the regularities when prompted to reflect on the stimuli at the end of the experiment. Future studies should examine how explicit knowledge develops as orthographic learning builds through exposure.

Associations between individual differences in statistical learning and spelling ability

Evidence from pseudoword spelling (Treiman et al., 2021; Treiman & Boland, 2017; Treiman & Kessler, 2006) supports the idea that good spellers are more likely to mirror the patterns that exist in their natural language when spelling novel words, consistent with sensitivity to orthographic regularities being associated with spelling experience and spelling ability. This leads to the prediction that there should be a correlation between spelling (or reading) ability and orthographic learning of an artificial orthography. However, the extant evidence is not consistent. For example, Schmalz et al. (2021) noted the absence of a significant correlation between learning outcomes in their artificial orthography learning paradigm and reading ability in German-speaking adults. Singh et al. (2021) only found a correlation when children received explicit instructions about the orthographic regularities in the novel language, and not in those who learned through exposure only. It is worth noting

that in these two studies, the information to be learned was relatively simple (i.e., graphotactic patterns or grapheme-phoneme mappings). It might be that individual differences in spelling ability are associated with learning if the patterns to be learned are more complex, or if they are embedded in a more complex environment. This prompted us to examine patterns of association between spelling ability and learning in our experiments.

Across the two experiments, there was either inconclusive evidence or evidence for the null for such associations. This is consistent with Singh et al. (2021) and Schmalz et al. (2021). However, we did observe significant positive correlations between spelling ability and performance on the word category task. This task measured whether participants had inferred the meaning of the artificial words during the exposure phase, as indexed by whether they correctly identified the lexical category of each word. The positive correlations with spelling suggest that good spellers are better at learning the meanings of the artificial words presented in the exposure phase, at least in the sense that they were able to identify the correct lexical category of the target word. This ability to infer and recall the intended word class of the artificial words appears to be independent of the ability to extract the embedded orthographic patterns and generalise them to novel words. However, it is important to acknowledge the limitations of the word category task. Although participants were asked to provide an English word with the same/closest meaning to each artificial word, we coded accuracy only based on whether their responses matched the intended word class. We did not assess whether participants accurately recalled the exact meaning of the artificial words. Therefore, the accuracy in this task may in fact reflect either learning of the word's meaning or its lexical category. Moreover, this task may have also underestimated category learning, as participants may be less inclined to give a response if they could not remember the exact word meaning. However, our preliminary review of the data suggests that these speculations may not be

supported. Our initial findings indicate that accuracy would have been very low had we scored the responses based on exact item recall. In addition, most participants provided responses for the majority of artificial words, with only a few answering "I don't know."

One unexpected finding was the positive correlation between semantic bias in the preference trials and spelling accuracy in Experiment 2 Version 2. This is particularly intriguing because no correlation was observed between performance in the semantic trials and spelling accuracy, nor between performance in semantic trials and semantic bias within the same experiment. This suggests that the correlation between semantic bias and spelling accuracy was not driven by successful learning of form-to-meaning mappings. One potential explanation is that the options in preference trials are more distinctive than those in the semantic trials. Specifically, in semantic trials, both options (e.g., "BOX" vs. "BOY") followed the semantic patterns seen during the exposure phase. To choose the correct answer, participants needed to know the association between "X" and "Y" and their specific word class. In contrast, options in the preference trials (e.g., "1AZ" vs. "1AX") were more distinctive from one another because "AZ" was never associated with the intended word class of the missing word in the trial, and "A" never appeared as a medial vowel when "X" was the final symbol during the exposure phase. Participants with better spelling ability may have taken this difference as an additional cue when completing the preference trials. However, given that this correlation was not observed in the other experiments, we acknowledge that this correlation between semantic bias and spelling ability in Version 2 (which was only supported by moderate evidence) may be spurious.

Taken together, our experiments provide little evidence for there being an association between learning artificial orthographic patterns and spelling ability. However, there was

some indication that good spellers were better at inferring and recalling word meanings of artificial words (at least their lexical categories) from the exposure phase. However, we must interpret these findings with caution as while we observed some learning of the form-to-meaning mappings, it was mostly in a small group of participants who could verbalise the patterns. The lack of correlation might reflect the lack of overall learning in our artificial orthography learning paradigm. Future studies should design experiments with complex patterns and prolonged exposure durations to assess the correlation between orthographic learning and spelling ability.

Conclusion

This study provides evidence in support of a statistical learning account of orthographic learning, though with some important caveats. After a short exposure to graphotactic patterns and form-to-meaning mappings in an artificial orthography, participants were able to generalise knowledge of the form-meaning regularities to novel items. However, this learning depended on the salience of word meanings, with more learning being shown for nouns and verbs than adjectives and adverbs. Participants who could explicitly verbalise the form-meaning regularities at the end of the experiment showed better learning than those who could not. There was no evidence of graphotactic learning. We also failed to find evidence for the association between learning the artificial orthography and spelling ability. These findings highlight the limits of learning via short exposure and without explicit instruction.

Chapter 4. The Impact of Semantic and Phonological Cues on Graphotactic Learning (Study 3)

Abstract

Orthographies such as English and French reflect multiple types of regularity conditioned to varying extents on phonological, graphotactic and semantic cues. Artificial orthography learning experiments have shown that people can develop sensitivity to graphotactic regularities (i.e., spelling patterns concerning legal order and combinations of graphemes) after brief exposure. However, most previous studies focused on learning in meaningless contexts, which does not reflect natural language learning where meaning is present. In our study, we designed an artificial orthography that modelled aspects of French article-noun pluralisation patterns to examine whether adult participants could learn graphotactic constraints through exposure, and whether semantic cues would facilitate this learning. We also investigated whether phonological cues impacted graphotactic learning by presenting the artificial orthography in Latin alphabets (Experiment 3) and symbols (BACS-2 font; Experiment 4). Across both experiments, participants developed sensitivity to the graphotactic patterns. Learning was stronger when the artificial orthography was presented in Latin alphabets, although some learning still occurred with symbols. Participants who could verbalise the patterns at post-test showed better generalisation. There was limited evidence that semantic cues facilitated graphotactic learning. These findings support the view that statistical learning processes underlie orthographic learning.

Introduction

Alphabetic writing systems such as English reflect regularities beyond phoneme-grapheme correspondences in their spelling. Graphotactic patterns, which govern the legal combinations and orders of letters, are also important to spelling in alphabetic languages. For example, in English, a consonant usually becomes a doublet after a short vowel (e.g., “puff” /pʌf/ vs. *puf), but remains as a singlet when following a long vowel (e.g., “paper” /peɪpə/ vs. *papper). Some graphotactic patterns can also be purely visual and independent of phonology. For instance, English consonant doublets are legal at word-final positions (e.g., “hill”, “bluff”) but are illegal at word-initial positions (e.g., *hhil, *bbluf).

Pseudoword experiments (e.g., Hayes et al., 2006) have shown that both children and adults are sensitive to graphotactic patterns, and this sensitivity impacts their spelling. As young children show sensitivity to these regularities before receiving formal instruction, some research has explained graphotactic learning from the statistical learning perspective. This perspective postulates that people extract regularities from the input materials and that the learning process is unintentional and implicit (Turk-Browne et al., 2005). Artificial orthography learning experiments such as Singh et al. (2021), where people are exposed to and tested with novel graphotactic patterns, provide support for this account, though learning levels are generally low.

However, current evidence from this paradigm has primarily focused on learning purely formal graphotactic patterns embedded in meaningless pseudowords, yet in naturalistic contexts, language conveys semantic relationships. In the current study, we explore whether the learning of graphotactic patterns differs when the patterns to be learned are/are not associated with semantics. Building on our previous findings (Law et al., 2025), we also

explored whether participants develop explicit awareness of these regularities from exposure and asked whether this awareness affects test performance. In the following sections, we begin by reviewing evidence from pseudoword and artificial orthography learning experiments and the evidence this provides for the statistical learning perspective in orthographic learning. We then consider how form-meaning regularities in natural language might impact graphotactic learning.

Sensitivity to graphotactic regularities in spelling

Much of the existing evidence on spellers' sensitivity to graphotactic patterns, such as consonant and vowel doubling, comes from pseudoword choice or spelling tasks. Cassar & Treiman (1997) tested children on their knowledge about consonant and vowel doublets using a pseudoword choice task, where they were asked to select the more word-like option from pseudoword pairs. Results across three experiments demonstrated that even first graders were sensitive to *where* consonant doublets are allowed (e.g., choosing "heniss" over "hhenis"). By the second half of first grade, children could also identify *which* consonants and vowels are allowed to be doublets (e.g., choosing "baff" over "bahh" or choosing "sook" over "saak"). These findings highlight young children's early sensitivity to graphotactic regularities in written language, which impacts their spelling choices.

In addition to *where* and *which* consonants and vowels can become doublets, spellers also determine whether the consonant should be extended based on the quality and/or spelling of the preceding vowel. Phonologically, if the vowel is what we traditionally considered as a short vowel (e.g., /ɪ/, /ɛ/, /æ/, /ɑ/, /ʌ/, /ʊ/), the following consonant tends to be extended. Graphotactically, if the preceding vowel is spelled with only one letter, the following consonant tends to be extended, regardless of the vowel pronunciations. Hayes et al. (2006)

studied whether children and adult spellers were sensitive to this using pseudoword choice and spelling tasks. Their results show that participants were impacted by both the vowel quality (i.e., consonant extension after a short vowel) and vowel spelling (i.e., consonant extension after a vowel spelled with one letter) when spelling consonants at the coda position in monosyllabic pseudowords (see also Treiman & Boland, 2017; Treiman & Kessler, 2016; Treiman & Wolter, 2018). Some evidence has even further suggested that the graphotactic account explains consonant doubling better than the phonological account (Hayes et al., 2006; Treiman & Boland, 2017; Treiman & Kessler, 2016).

Statistical learning of graphotactic patterns

Evidence from pseudoword experiments clearly shows the importance of graphotactic knowledge in producing accurate spellings. However, these graphotactic patterns are not usually explicitly taught in school (Treiman & Wolter, 2018). There are also simply too many of such patterns to all be included in classroom teaching. How do people, especially young children, learn these patterns? One account that has been proposed is statistical learning, which posits that people extract patterns from the materials they are exposed to without any explicit instruction or feedback (Turk-Browne et al., 2005). Some support for this account comes from evidence that knowledge of graphotactic patterns is implicit and not accessible to conscious awareness. In particular, Treiman and Wolter (2018) found that only 25% of their participants were able to verbalise some explicit knowledge of the relationship between consonant doubling and the quality of the preceding vowel, while none could describe the relationship between consonant doubling and the spelling of the preceding vowel. Among those who were not explicitly aware of the phonological patterns, there was still significant evidence that both vowel quality and vowel spelling impacted their choices in consonant doubling. This indicates that knowledge of the patterns may be implicit, and is consistent

with an account in which this knowledge results from an ability to extract regularities from the input through statistical learning.

However, natural language is complex and rich with different types of statistical patterns. To directly examine the learning of a particular orthographic pattern, some research has adopted the artificial orthography learning paradigm where people are taught and tested on new orthographic patterns – either using a familiar alphabet or novel orthography – under experimental conditions. As the orthographic patterns are new to participants, this paradigm allows experimenters to have complete control over the input statistics, and thus enables them to test hypotheses about the occurrence and co-occurrences of letters and letter combinations. Learning and generalisation can therefore be tracked as a function of experimental manipulations.

Singh et al. (2021) adopted this paradigm to investigate how people learn graphotactic patterns with no phonotactic counterparts in monosyllabic artificial words. Their experiments were conducted with English speakers and used the familiar Latin alphabet. They took advantage of the phenomenon that English consonant singlets and doublets share the same pronunciations (e.g., “rus” and “russ” for the pronunciation /rʌs/), and created stimuli embedding an association between the medial vowel (“u” or “e”) and consonant doubling. Under an incidental learning condition, both English-speaking children and English-speaking adults learned the association between the medial vowel and consonant doubling. Though learning effects were small, they were able to generalise this graphotactic knowledge to novel words. They also explored the extent to which learning was explicit using a post-experiment questionnaire. This showed that 30% of adult participants could describe the graphotactic patterns, however, while these ‘aware’ participants had better performance, those who could

not verbalise these patterns still showed above-chance generalisation performance. This evidence supports the statistical learning perspective in orthographic learning in which learners can develop sensitivity to graphotactic patterns through exposure, and this knowledge is not necessarily accessible to conscious awareness (see also Samara et al., 2019; Samara & Caravolas, 2014).

Sensitivity to morphological regularities in spelling

One limitation of the experimental studies discussed above is that they examined the learning of spelling patterns within meaningless pseudowords. This contrasts with natural language where spelling patterns may also convey meaning, since letters and letter groups often function as morphemes. Consider the written English past tense suffix -ed, which corresponds to different pronunciations including /d/ (e.g., "sav<u>ed</u>" /seɪv<u>d</u>/), /t/ (e.g., "jump<u>ed</u>" /jump<u>t</u>/) and /ɪd/ (e.g., "wait<u>ed</u>" /weɪt<u>ɪd</u>/). Given the variations in the pronunciations, learners may find it easier to directly associate the concept of past tense with the written form -ed than relying on its pronunciations when learning its spelling. On the other hand, different morphemes can share the same phonological forms but have distinct spellings. For instance, while the suffixes -us in "bonus" and -ous in "nervous" are both pronounced as /əs/, the written suffix -ous is predominantly used in English adjectives (Ulicheva et al., 2021). In these cases, recognising the lexical category, where the former refers to a non-adjective and the latter to an adjective, can then help determine the correct spelling. These examples highlight the importance of meaning to the learning-to-spell process.

As with graphotactic patterns, evidence from pseudoword spelling experiments (Heyer, 2021; Treiman et al., 2021; Ulicheva et al., 2020) shows that individuals are sensitive to the associations between forms and meaning in written language. For example, Ulicheva et al.,

(2020) used a pseudoword classification task and asked participants to categorise pseudowords that contained English suffixes (e.g., adjectives: “cevable”, “dolous”, “tumish”; nouns: “tobness”, “jumer”, “nadence”) as nouns or adjectives. Importantly, these English suffixes have high diagnosticity according to their corpus, meaning that the suffix spelling highly predicts a particular lexical category in English. Their results indicated that participants accurately classified pseudowords based on their suffixes, demonstrating sensitivity to the relationship between spelling patterns and meaning.

To further examine the learning of spelling patterns which are associated with semantic cues, Law et al. (2025) designed an artificial orthography learning experiment in which native English-speaking adult participants were exposed to artificial words embedded in English sentences during a reading task. Critically, the sentence contexts ensured that the artificial words were associated with lexical categories, and the words embedded spelling patterns associated with lexical semantics (the last symbol of the artificial word indicates the lexical category) or graphotactic regularities (the last symbol of the artificial word is determined by the medial symbol). They were constructed using a mix of new symbols to present the consonants and English vowels in the Latin alphabet to create stimuli that could be learned in a short session whilst eliminating phonological cues. Despite only 20 minutes of exposure, participants performed above chance in generalising the form-meaning associations to novel items (e.g., they knew that nouns should end with one symbol and verbs another), suggesting that they recognised part of the artificial word indicates its lexical category. However, in contrast to previous work, here the post-questionnaire responses indicated that learning was driven by participants who were able to verbalise the form-meaning associations. Specifically, when participants who could verbalise the patterns were removed from the dataset, there was very limited evidence that learning was above chance. This indicates that

their knowledge of these patterns was relatively explicit. In addition, contrary to other graphotactic learning experiments such as Singh et al. (2021), there was no evidence of graphotactic learning (i.e., they did not know that the choice of final consonant depends on the vowel).

Co-occurrence of form-meaning and graphotactic regularities

The findings from Law et al. (2025) suggest that when spelling patterns are embedded in meaningful contexts, learners can quickly pick up on form-to-meaning mappings, albeit quite explicitly. However, the absence of graphotactic learning in their study suggests that purely formal spelling rules may be harder to learn when embedded in meaningful contexts compared to when they are embedded in meaningless pseudowords.

In natural language, there are also cases where spelling rules apply to particular morphemes, yet these rules could potentially be learned as formal graphotactic patterns. One example is the noun pluralisation patterns in French. Similar to English, French nouns form plurals by adding suffixes, but the choice of suffixes is also influenced by graphotactic constraints. In most cases, French nouns take the plural suffix -s (e.g., “enfants”, “fleurs”, “maisons”). However, a small number of nouns with specific spelling patterns require -x instead. For example, for nouns ending in “eau”, they must take -x as a plural suffix (e.g., “bateaux”). Those ending with “eu” and “ou” can take either -x or -s, though “eu” more commonly takes -x (e.g., “feux”, “jeux”) and “ou” more often takes -s (e.g., “choux”, “trous”). Thus, these graphotactic regularities in French noun pluralisation are also probabilistic: they vary in frequency (with -s being a more frequent suffix than -x) and in consistency (i.e., “eau” ending only takes -x but “ou” ending can take either -s or -x). Importantly, these French plural

suffixes are most often "silent[11]." For example, in the singular noun phrase "un stylo" ("a pen") /œ̃ stilo/ and its pluralised form "des stylos" ("some pens") /de stilo/, the plural suffix -s is present in the written form but it has no phonological counterpart. As a result, the singular and plural nouns share the same pronunciation. Spellers must rely on the article to recognise whether it is a singular or plural noun phrase, and the vowel in the stem determines the correct spelling of the plural suffix.

How important is the morphological status of the suffix in this learning process? One possibility is that learners could acquire these patterns as purely graphotactic constraints, becoming sensitive to the pattern without any consideration of its meaning. That is, they recognise that in certain contexts (e.g., following "des"), the noun requires an additional consonant, and the identity of this consonant depends on the ending of the word. Alternatively, learners may recognise the additional letter in the noun as a morpheme indicating plurality, and that this morphological regularity is further constrained by graphotactic patterns based on spelling of the noun ending.

Some experimental evidence has suggested that the learning of French plural spellings is facilitated by semantics. Klasen et al. (2023) examined this with Luxembourgish fourth graders who were learning French as their second language. Participants completed two gap dictation tests eliciting real and pseudo nouns, verbs and adjectives. In each trial, they listened to a sentence and filled in the missing word. For example, in the sentence "Mes __/kɔpɛ̃/__ mangent une pizza" ("My friends eat a pizza"), participants filled out the plural noun "copains*"* ("friends") based on the auditory input. A key finding was that participants were more accurate at pluralising nouns than adjectives. While this may partly reflect that

[11] Due to liaison, some plural endings might be pronounced when the following word begins with a vowel.

nouns are more frequently taught in the curriculum, it could also be because noun plurals refer to tangible objects and are more semantically grounded. In contrast, adjective agreement is purely formal. That is, in the noun phrase “les petites filles” (“the little girls”), the -s in “filles” represents plurality (i.e., there are more than one girl), whereas the -s in “petites” does not carry any additional meaning. This semantic grounding likely helps learners understand the function and importance of the suffix, which results in better spelling performance in noun plurals (see also Fayol et al., 2006). However, no study to date has directly compared the learning of these noun pluralisation spelling patterns with and without semantics. Given that Law et al. (2025) found that meaningful contexts are not necessarily helpful to the learning of graphotactic patterns, the present study set out to examine the impact of semantic cues on learning this type of graphotactic patterns.

In this study, we created an artificial orthography modelling aspects of the article-noun agreement in French to examine whether English-speaking adult participants could learn statistical graphotactic patterns via exposure, and the impacts of semantic cues on this learning. Furthermore, following Singh et al. (2021) and Law et al. (2025), we used a post-test questionnaire to examine whether participants developed explicit knowledge of the target patterns, and how this relates to their performance at post-test.

Experiment 3

This pre-registered experiment (https://osf.io/7qad4/) examined whether adult participants can learn graphotactic patterns in an artificial orthography created from Latin alphabets, and the impact of semantic cues on this learning. We also explored whether participants’ ability in recalling the word meaning was related to their graphotactic learning. Finally, we assessed

whether participants could verbalise these patterns at post-test, and how related this is to their learning of these graphotactic patterns.

Method

Sample size and power calculations

One main goal of this experiment was to examine the impact of semantic cues on learning the graphotactic patterns. Therefore, to determine the sample size for this experiment, we conducted a priori power analysis in G*Power 3.1, targeting the effect of condition (SE condition vs. GR condition) using an independent *t*-test (one-tailed). The analysis indicated that a total of 102 participants would be needed to achieve a statistical power of .80 for detecting a medium effect size ($d = 0.5$). As described in our pre-registration, we applied a stopping rule using Bayesian statistics. Specifically, we analysed the data after testing 56 participants (i.e., 28 participants in each condition with 7 participants in each of the four randomisation lists within each condition). If the Bayes Factor for this effect was larger than 3 or smaller than 1/3 (i.e., there was conclusive evidence for the alternative or null hypothesis), we would stop testing. Otherwise, we would continue testing in increments of 8 participants (4 per condition) and analysing the data until the Bayes Factor crossed the pre-determined thresholds or the total sample size reached 102 participants. Following this procedure, we tested the full planned sample of 102 participants.

Participants

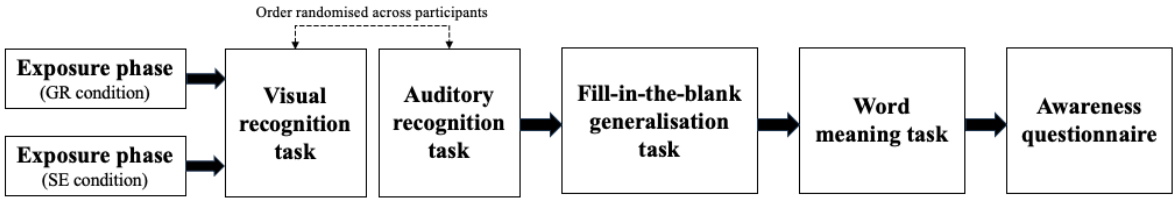
A total of 102 native English-speaking adults (57 female; mean age = 33.78; *SD* = 13.88) were recruited, all with self-reported normal or corrected-to-normal vision and no known neurological and learning impairments (see Appendix 4A for additional information on

participants' language backgrounds). Among these participants, 77 were recruited through the online participant recruitment platform Prolific and 25 were recruited through the University's research participation scheme. The former were paid £6 and the latter received course credit. Most of the participants (N = 75) had obtained or were currently pursuing an undergraduate or postgraduate degree at the time of the study. As per our pre-registered criteria, we excluded six participants who did not meet the 75% accuracy benchmark in the visual and/or auditory recognition tasks. Additionally, three participants were excluded as they indicated French as their second language. The final sample contains 93 participants with 46 participants in the graphotactic constraints only (GR) condition and 47 participants in the semantics cues (SE) condition.

Design

As shown in Figure 4.1, the experiment comprises an exposure phase, with one between-participant manipulation (GE/SE condition) followed by four post-test tasks and an awareness questionnaire.

Figure 4.1. Schematic depiction of Experiment 3.



The *Visual and Auditory Recognition Tasks* were included to provide a baseline attention check (participants who did not meet the criteria on this test are assumed not to have sufficiently attended the experimental stimuli and are excluded). The *Fill-in-the-blank Generalisation Task* provides the key data assessing how well participants in each of the conditions learned the different types of graphotactic patterns embedded in the stimuli. The *Word Meaning Task* assessed whether they had learned the meanings of the words. Note that data from this last task are only meaningful for the SE group; however, for consistency, all participants took all four tests. The post-test awareness questionnaire tested their ability to verbalise the graphotactic patterns.

Materials

Exposure phase

Our artificial orthography consists of 48 singular pseudo-noun phrases (e.g., "yim bligma") and their pluralised form (e.g., "vop bligmax") (see Appendix 4B for the full list). Each pseudo-noun phrase contains the pseudo-article "yim" /jɪm/ (for singular nouns) or "vop" /vɒp/ (for plural nouns), followed by a pseudo-noun (hereafter referred to as noun phrase, noun and article). Singular nouns can be categorised into four types, each with a unique vowel ending: "a", "i", "o", and "u". The critical experimental manipulation is the pluralisation patterns, which are conditioned on the ending of the singular noun. Six plural suffixes namely -x, -n, -t, -k, -d and -v are mapped to singular noun endings with varying levels of consistency (i.e., how predictable the plural suffix is based on the singular noun ending) and frequency (i.e., the number of items participants encounter the noun ending-suffix mappings during the exposure phase). Table 4.1 shows the pluralisation patterns in our stimuli.

Table 4.1. Randomisation of pluralisation pattern

Frequency	Singular noun ending	Consistency	Absolute frequency	Plural suffix (Word list 1)	Plural suffix (Word list 2)	Plural suffix (Word list 3)	Plural suffix (Word list 4)
High	a (e.g., bligma)	100%	48	-x	-n	-t	-k
Low	i (e.g., lomi)	100%	24	-v	-x	-n	-t
High	o (e.g., begro)	75%	36	-k	-d	-v	-x
		25%	12	-d	-v	-x	-n
Low	u (e.g., timmu)	75%	18	-n	-t	-k	-d
		25%	6	-t	-k	-d	-v

Note. Absolute frequency refers to the number of times participants encounter each specific suffix ending in the entire exposure phase.

As shown in Table 4.1, four versions were devised for counterbalancing purposes. In word list 1, singular nouns ending in “a” and “i” consistently take the plural markers -x and -v, respectively. In contrast, singular nouns ending in “o” take the plural marker -k in 75% of the items and -d for the remaining items. Similarly, singular nouns ending in “u” take the plural marker -n in 75% of the items and -t in the remaining items. This results in inconsistent mappings between singular noun endings and plural markers. In addition to the consistency, the frequency of the noun ending-suffix mappings was also manipulated. Singular nouns ending in “a” and “o” have high frequency (i.e., each noun ending occurs in 16 unique training items) whereas those ending in “i” and “u” have low frequency (i.e., each noun ending occurs in 8 unique training items). Each pair of noun phrase was repeated 3 times in the exposure phase, resulting in a total of 144 exposure items.

Importantly, singular and plural nouns had identical pronunciations in the auditory input (e.g., “yim bligm**a**” /jɪm blɪgmə/ vs. “vop bligm**ax**” /vɒp blɪgmə/), making the plural suffixes only visual in the written forms. To correctly generalise the plural markers at post-test, participants must learn the association between the plural article and plural suffix in the noun, as well as the context-conditioned graphotactic patterns of the plural suffix in its spelling. Note that these stimuli were only presented as two-word phrases to participants. Participants were not told that the stimuli correspond to article-noun phrases, although we expected some participants to discover this structure during the exposure phase. This was assessed through the awareness questionnaire.

In the exposure phase, participants were randomly assigned to one of the two conditions – (1) graphotactic constraints only (GR; i.e., no semantic cues) and (2) semantic cues (SE). The critical difference was whether participants were also exposed to the word meanings

alongside the artificial words during exposure. In the SE condition, each artificial word was paired with an object from one of eight categories (animal, clothing, food, tool, containers, people, toy and vehicle). Singular noun phrases were paired with a single object (e.g., one elephant) whereas plural noun phrases were paired with three objects to indicate plurality (e.g., three elephants). Object images were obtained from Szekely et al. (2004). Each semantic category appeared with equal probabilities across artificial words with the same noun ending.

All the singular nouns were pseudowords selected from the English Lexicon Project (Balota et al., 2007) based on the following criteria: (1) maximum of 3 orthographic neighbours, (2) word length between 4 to 7 letters, (3) two-syllable structure, and (4) high nonword identification accuracy according to the database (i.e., above 0.75 in the proportion of accurate responses for recognising it as a nonword). The two articles were also chosen from the same corpus. As three-letter words, they had nine orthographic neighbours but a nonword identification accuracy of 1. The corresponding audios for each phrase were created using Microsoft Speech Studio. To ensure that these pseudowords do not resemble real words either in its visual or spoken form, we asked native English speakers who did not participate in this study to read through and listen to both the exposure and testing items. They were asked to judge whether (1) the pseudowords were similar to any real English words in either the visual or auditory form, (2) the audios matched the visual form of the words and (3) the audios sounded natural. All final exposure and testing items were judged to be natural and not resemble any real English words.

Visual and auditory recognition tasks

The purpose of these tasks was to assess if participants had sufficiently attended to the training stimuli. Both the visual and auditory recognition tasks contained 12 trained items and 12 foil items, respectively. The 12 trained items were the singular noun phrases from the exposure, with 3 items selected for each noun ending. Three types of foil items were created: (1) a novel article combined with a novel noun ending with a learned suffix (e.g., "bes homa"), (2) a learned article combined with a novel noun ending in a novel suffix (e.g., "yim aining") and (3) a novel article combined with a novel noun ending in a novel suffix (e.g., "cen admusts"). There are four items for each type of foil items (see Appendix 4C for the full list).

Fill-in-the-blank task

An additional 32 singular noun phrases (i.e., 8 per singular noun ending) were created for the fill-in-the-blank task (the critical test of generalisation) following the same criteria as the items in the exposure phase (see Appendix 4C for the full list). In each trial, participants were presented with six options. Each option featured one of the possible plural suffixes from the exposure phase.

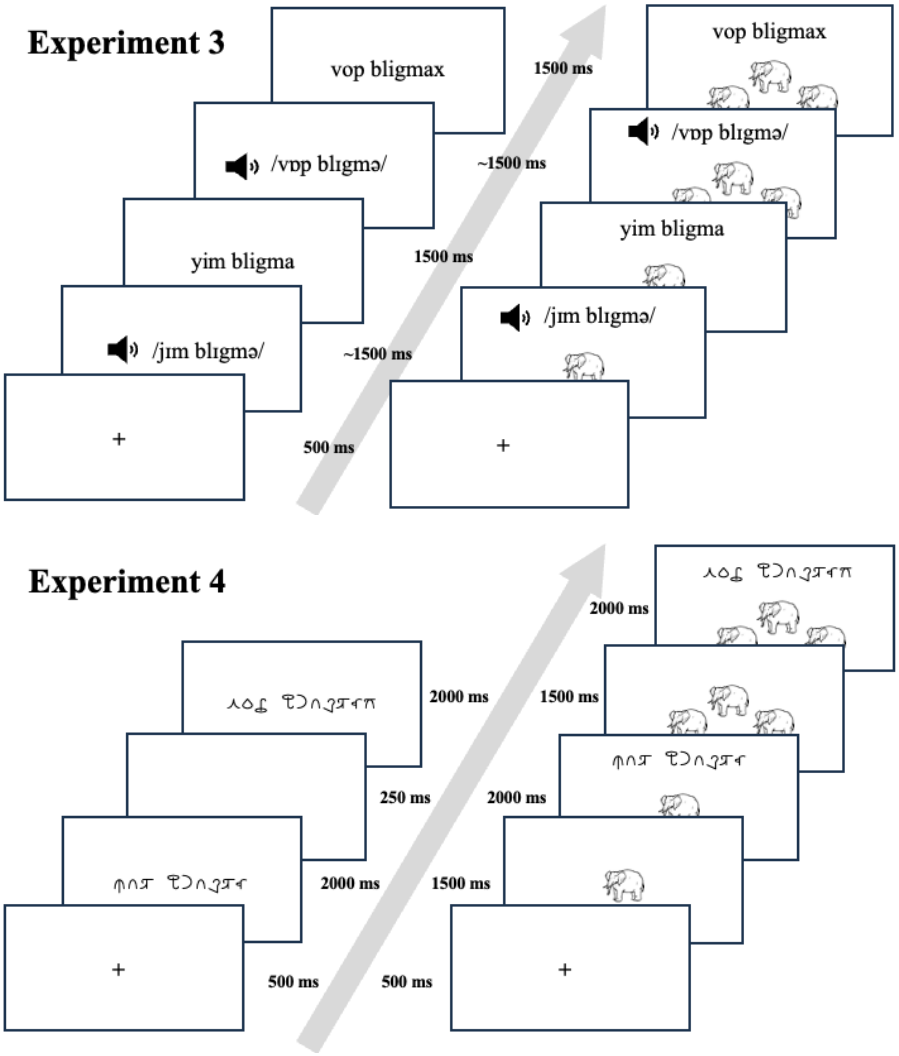
Procedure

The experiment was conducted online using Gorilla.sc (Anwyl-Irvine et al., 2020) and the link was distributed to participants through Prolific (www.prolific.co) or the Research Participation Scheme at the University. Participants completed the experiment on their own computer or laptop. The entire experiment took approximately 40 minutes to complete.

Exposure phase

This contained a passive-viewing task where participants saw and listened to the 48 singular-plural noun phrases. Figure 4.2 (top) illustrates the presentation order for each trial within the GR (left) and SE (right) conditions in Experiment 3.

Figure 4.2. Presentation order of each noun phrase in the exposure phase in Experiment 3 (top) and Experiment 4 (bottom) for graphotactic constraints only (GR) (left) and semantic cues (SE) conditions (right).



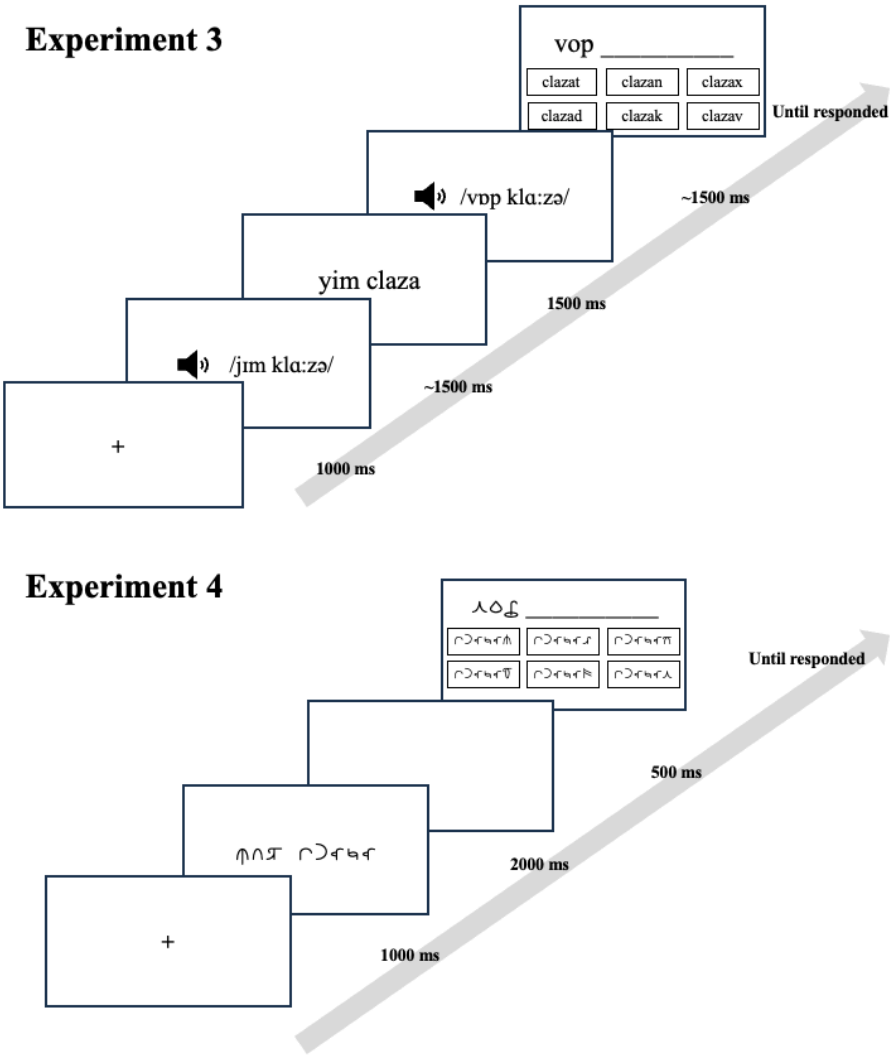
As shown on the left of Figure 4.2, for each trial, after the fixation point (500ms), participants in the GR condition first heard the singular noun phrase (e.g., /jɪm blɪgmə/) and then saw its written form (e.g., “yim bligma”). This was followed by the pluralised noun phrase first in auditory (e.g., /vɒp blɪgmə/) and then its written form (e.g., “vop bligmax”). Participants in the SE condition were exposed to the stimuli following the same format, but with the addition of pictures. Specifically, pictures indicating one object (for singular noun phrases) and multiple objects (for plural noun phrases) appeared on the screen alongside the written form of the phrase. The exposure phase comprised of a total of 3 blocks with each block showing the 48 phrases once.

Testing phase

Visual and auditory recognition task. In the recognition task, participants completed both visual and auditory trials. For the visual task, they were presented with 12 trained singular noun phrases and 12 foil phrases, one at a time, and asked if they had seen the words earlier during the exposure phase. The auditory task followed the same procedure but presented the phrases in auditory forms only, and participants were asked to judge whether they had heard the words earlier. The task order was counterbalanced across participants. This task served as an attention check. Participants had to achieve at least 75% overall accuracy in both tasks in order to be included in the final analyses.

Fill-in-the-blank task. This tested whether participants had learned the graphotactic patterns. Before the task, participants were informed that they would see some phrases with a missing word and they were instructed to choose one out of the six choices they thought best fit the phrase. Figure 4.3 illustrates the presentation order of each trial.

Figure 4.3. Presentation order of each trial in the fill-in-the-blank task in Experiment 3 (top) and Experiment 4 (bottom).



As illustrated in Figure 4.3 (top), the presentation order of each trial in the fill-in-the-blank task followed that of the exposure phase. After a fixation point (1000ms), participants encountered the auditory form of a singular noun phrase, followed by its written form. They then heard the auditory form of the plural noun phrase, and the written form appeared on the screen, with the critical pluralised noun left blank for participants to fill in. Six options appeared below the blank space for the critical item. The positions of the items were randomised across trials and participants.

For items with consistent noun ending-suffix mappings (i.e., those ending in “a” and “i”), there is only one correct answer among the six options. For those with inconsistent mappings (e.g., nouns ending with “o” take -k in 75% of the items and -d in the remaining items), there are two plausible answers, with one being the dominant suffix and the other one the non-dominant suffix. There was a total of 32 trials in this task (i.e., 8 trials per singular noun ending × 4 singular noun endings). Participants, no matter which condition or word list they were assigned to during the exposure phase, received the same fill-in-the-blank task.

Word meaning task. This task tested whether participants in the SE condition learned the word meanings from exposure. Participants were shown the nouns and articles they had seen in the exposure phase, and were asked to write down an English word that had the closest meaning for each. They were given 1 score if their answers matched the intended meaning including synonyms. All participants completed this task. However, given that the GR condition did not learn any meanings during the exposure phase, this task was only given to them to keep the experimental procedures constant for both groups. Only accuracy from the SE condition was used in our analyses.

Awareness questionnaire. Participants completed an awareness questionnaire after completing all the other tasks. This asked whether they (i) were aware of the purpose of the experiment, (ii) could explicitly verbalise any patterns and rules in the phrases, and (iii) used any strategies when choosing the words to fit into the phrases in the fill-in-the-blank task.

Statistical analyses

We used Bayes Factors to perform the Bayesian equivalent of significance testing, as they provide information that *p*-value from frequentist statistics cannot. Specifically, while a *p*-

value below 0.05 indicates significant evidence for the alternative hypothesis, a *p*-value above 0.05 cannot distinguish between evidence for the null, against the null, or there is no evidence for any conclusion at all. In contrast, Bayes Factors can make this distinction, allowing us to better understand the data.

To compute Bayes Factors, we followed the method recommended by Dienes (2008, 2014) (see also Silvey et al., 2024) and Diene's calculator (implemented in R by Baguley and Kaye, 2010). This method requires three numbers to test the hypothesis that two means differ. The first two numbers are the data summary that consist of the mean difference in the sample for the hypothesis being tested (e.g., the difference between mean performance and chance, or the mean difference between two conditions) and its associated standard error (*SE*). To obtain these values, we used logistic mixed-effects models (using the glmer function in the lme4 package; Bates et al., 2015) and extracted the beta and *SE* values for the relevant coefficient in the model. The third number required is an approximate estimate of the predicted difference (i.e., predicted size of the beta) for the hypothesis. Note that all three values are in log-odds space, in order to meet the normality conditions of the calculator.

Appendix 4D shows the structure of the glmer models used to extract the coefficients[12]. In general, our approach was to put all experimented variables into the models as fixed effects but only extract the coefficients relevant to the predictions. We generally used a centred coding for our fixed effects, so that the intercept represents the grand mean for the data set in question, and all predictors can be interpreted as main effects (though in some cases we used simple coding to get at simple effects). We had random intercepts for participants and full

[12] The final analyses differ slightly from the pre-registered analysis plan, as it became clear that including all experimented variables in a single model and the frequency variable in the random slope structure better represents the design of our experiment.

random slopes structure. For each test that we did, we extracted the relevant estimate (beta) and *SE* from the coefficient, and we also needed a predicted estimate. We determined these values based on based on methodologically similar experiments such as Law et al. (2025) and Singh et al. (2021) (full details of how we determined these values, which were pre-registered, are given in the Appendix 4D). These predicted values are used as a parameter (or the "scale factor") in our model of H_1, which is a model which represents the plausibility of different effect sizes if H_1 is true. In almost all cases, we had a directional prediction and thus used a half-normal distribution (to test a one-tailed distribution) with a mean of 0 and SD set to the scale factor. This approach, as recommended by Dienes (2014), is suited for situations where a directional effect is predicted but only a rough estimate of the effect is available, and when smaller values are more likely. Diene's calculator tests whether the data summary is more likely under this model of H_1 than under a model representing the null (using a point null where the only plausible effect is 0).

The result is a Bayes factor (*BF*), which quantifies the relative strength of evidence for H_1 versus the null as a ratio. Values greater than 1 suggest more evidence for H_1, while values below 1 indicate evidence for the null. Bayes Factor can be interpreted on a continuous scale, however we refer to discrete evidential categories for hypothesis testing. Specifically, we used $BF > 3$ as indicative of moderate/substantial evidence for H_1 and $BF < \frac{1}{3}$ to indicate moderate/substantial evidence for H_0. Values between these thresholds are interpreted as inconclusive evidence, meaning that the data is insensitive to test the hypothesis. *BF*s > 10 or $< \frac{1}{10}$ are considered strong evidence for H_1 or the null, respectively (Lee & Wagenmakers, 2014; Schönbrodt & Wagenmakers, 2018). Note that $BF > 3$ represents a similar level of conservatism to the more familiar $p < .05$, and these indicators very often align but are not guaranteed.

Bayes Factors are sensitive to the choice of values for the predicted effects and there is some subjectivity in this choice. Therefore, in addition to Bayes Factors, we also calculated "robustness regions" for each *BF*. Robustness regions are reported as *RR* [x:y] in our main text. These show the range of predicted values that we could have used as the scale factor for the model of H_1 and still have gotten the same conclusion, based on the cut-offs of $BF > 3$ or $BF < \frac{1}{3}$ [i.e., x and y represent how low/high a value we could have used and still obtained a *BF* which was greater than 3 ($BF > 3$), lower than $\frac{1}{3}$ ($BF < \frac{1}{3}$), or between 1/3 and 3 ($\frac{1}{3} < BF < 3$)]. Further details of both the mixed effects models and the computation of the Bayes factors are in Appendix 4D. Note that in addition to reporting Bayes Factors, we also report the *p*-values since these are more familiar, but we do not interpret them.

Results

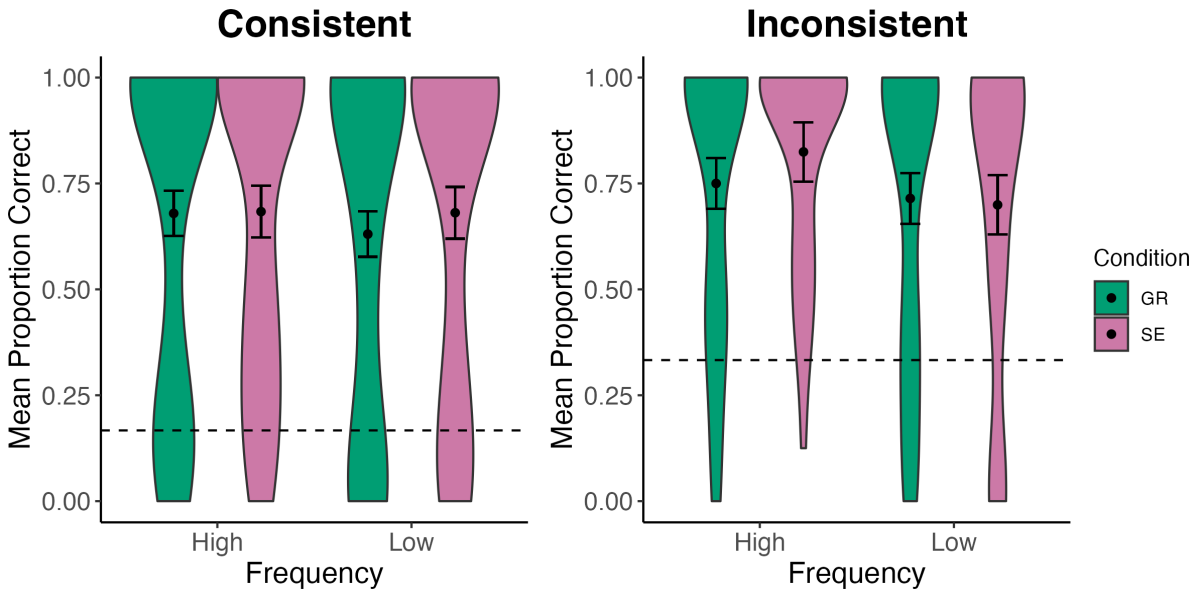
The key analyses focus on the results of our fill-in-the-blank task. Figure 4.4 shows the mean proportion of correct responses in this task for consistent and inconsistent items, split by frequency and condition. Note that here, picking either of the possible options for the inconsistent items was coded as correct. Since there are 6 options, chance is thus 1/6 (-1.609 in log odds) for consistent items and 1/3 (-0.693 in log odds) for inconsistent items. Given this difference, we ran separate models over the data from the consistent and inconsistent items. All data and analysis scripts are available on https://osf.io/3kx82/.

Fill-in-the-blank overall performance

We can see in Figure 4.4 that participants are well above chance. This was confirmed in the models by comparing the overall intercept for consistent items, $BF_{(0,1.81)} = 39420277876$, *RR* [0.09, > 4.59] (model intercept: $\beta = 4.44$, $SE = 0.60$, $p < .001$) and inconsistent items,

$BF_{(0,0.89)}$ = 897678302601, *RR* [0.06, > 4.59] (model intercept: β = 3.44, *SE* = 0.42, *p* < .001), to the relevant chance value. This suggests that participants have successfully learned the permissible noun ending-suffix mappings for both types of items.

Figure 4.4. Mean proportion of correct responses for consistent items (left) and inconsistent items (right), split by frequency and condition, in the fill-in-the-blank task in Experiment 3.



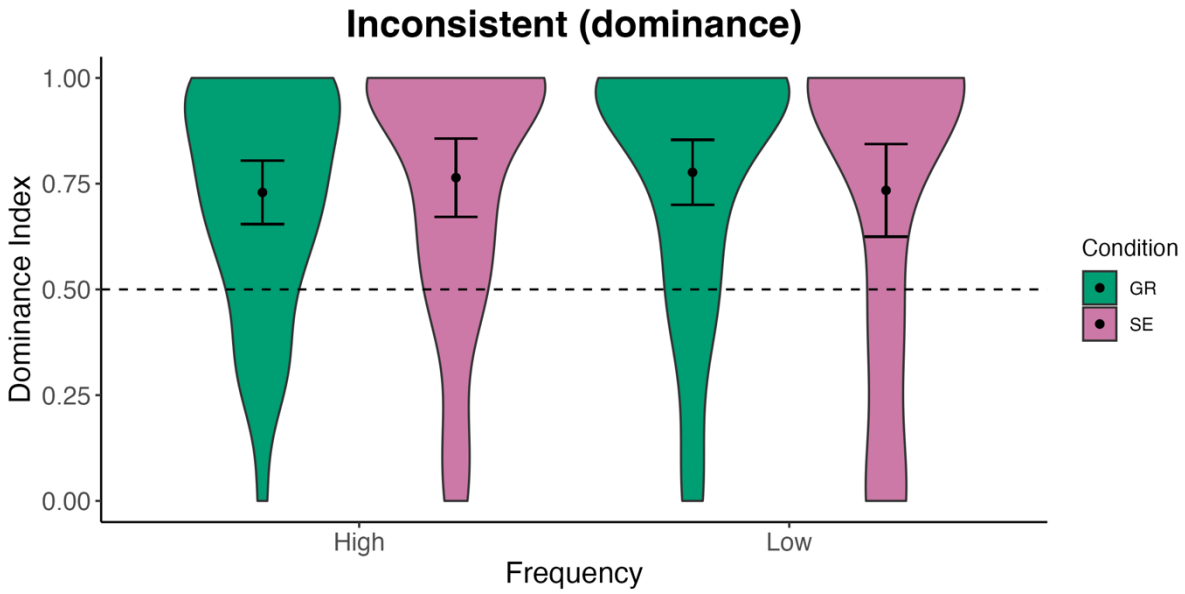
Note. Error bars show 95% confidence intervals. Dotted lines represent the chance level.

For the frequency effect, the evidence for a high-frequency benefit was ambiguous for both consistent, $BF_{(0,0.91)}$ = 0.60, *RR* [0, 1.90] (model coefficient: β = 0.001, *SE* = 0.67, *p* = .998) and inconsistent item, $BF_{(0,0.45)}$ = 1.27, *RR* [0, 3.71] (model coefficient: β = 0.47, *SE* = 0.55, *p* = .39).

For condition, the evidence for better performance in the SE than GR condition was ambiguous for both consistent items, $BF_{(0,0.41)}$ = 1.09, *RR* [0, 4.58] (model coefficient: β = 0.51, *SE* = 1.01, *p* = .62) and inconsistent items, $BF_{(0,0.41)}$ = 0.83, *RR* [0, 1.83] (model coefficient: β = -0.09, *SE* = 0.72, *p* = .90). Thus, we cannot determine whether semantic cues did or did not facilitate graphotactic learning in this dataset. However, there was strong

evidence for above-chance performance in consistent items in both the GR condition, $BF_{(0,1.81)}$ = 13919.99, *RR* [0.20, > 4.59] (model intercept: β = 4.32, *SE* = 0.88, $p < .001$) and SE condition, $BF_{(0,1.81)}$ = 362172.90, *RR* [0.16, > 4.59] (model intercept: β = 4.54, *SE* = 0.81, $p < .001$). Similarly, there was strong evidence for above-chance performance in inconsistent items for both the GR condition, $BF_{(0,0.89)}$ = 5288.46, *RR* [0.15, > 4.59] (model intercept: β = 3.73, *SE* = 0.71, $p < .001$), and the SE condition, $BF_{(0,0.89)}$ = 4869510, *RR* [0.09, > 4.59] (model intercept: β = 3.20, *SE* = 0.50, $p < .001$). The interaction between condition and frequency (which were registered as exploratory and tested with two-tailed Bayes Factors) also yielded no conclusive evidence for H_1 or null in either model (see Appendix 4E).

Figure 4.5. Dominance index in inconsistent items, split by frequency and condition, in the fill-in-the-blank task in Experiment 3.



Note. Error bars show 95% confidence intervals. Dotted lines represent the chance level.

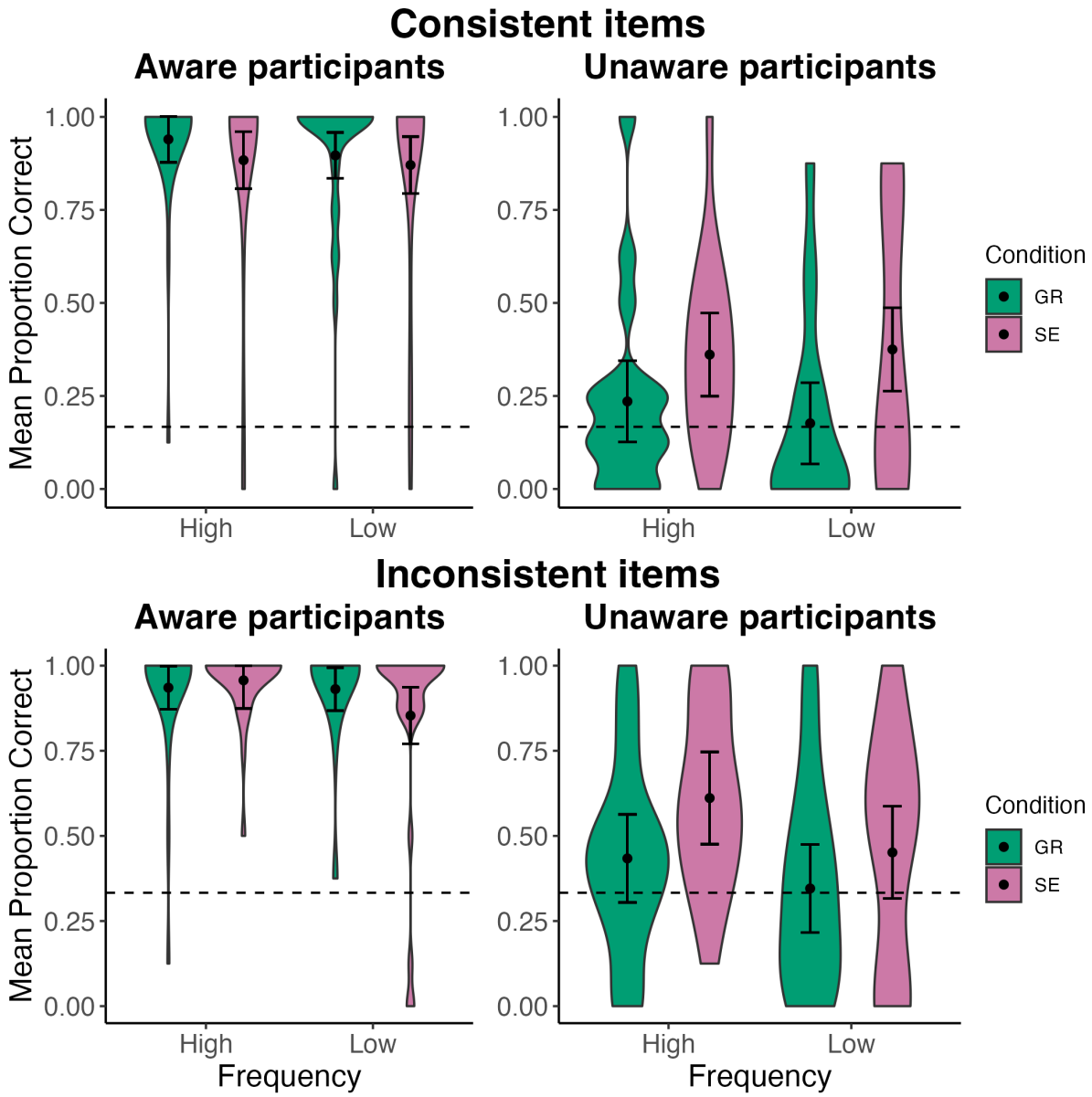
For inconsistent items, we were also interested in whether participants learned that there were two plausible suffixes, but that one was more dominant than the other. To examine this, we only analysed trials where participants selected one of the two plausible suffixes and coded the responses using a dominance index where 1 indicates that they selected the dominant suffix and 0 the non-dominant suffix. Thus, performance above 0.5 (0 log odds) indicates a bias towards the dominant suffix and below 0.5 a bias towards the non-dominant suffix. Figure 4.5 shows data for the inconsistent trials replotted with this coding, again split by frequency and condition. We ran a glmer model with the same structure as the previous model. There was strong evidence that participants' overall dominance index (the intercept) was greater than 0 (chance), $BF_{(0,1.10)}$ = 1.02726e+12, *RR* [0.04, > 4.59] (model intercept: β = 2.02, *SE* = 0.26, $p < .001$), suggesting a tendency to choose the dominant suffix over the non-dominant suffix. There was evidence against the prediction of a frequency effect in the direction of a higher dominant index in higher-frequency items, $BF_{(0,0.55)}$ = 0.31, *RR* [0.51, >

4.59] (model coefficient: β = -0.81, *SE* = 0.49, *p* = .10). There was no conclusive evidence for an effect of condition or an interaction effect between condition and frequency on the dominance index (see Appendix 4E).

Performance by awareness status

Following Singh et al. (2021) and Law et al. (2025), we were interested in whether explicit awareness impacted graphotactic learning within each of the SE and GR conditions. To examine this, we coded the data from the awareness questionnaire as to whether each participant could verbalise the graphotactic patterns. Participants were considered as aware (coded as 1) if they showed an understanding of the relationship between the singular noun ending and the additional suffix. Some example responses of aware participants were "I noticed the silent endings of words was linked to which vowel the singular finished with" and "… last letter of second word followed n after i, t after a, k after u, after o was either x or v (majority of time it was v)." Based on these criteria, 29 participants in each of the GR and SE conditions were coded as aware, leaving 17 (GR) and 18 (SE) participants coded as unaware. In Figure 4.6, we have replotted the accuracy data in Figure 4.4 for the consistent (top) and inconsistent (bottom) items, split by condition and "awareness status" (i.e., where the participants are coded as aware/unaware).

Figure 4.6. Mean proportion of correct responses in consistent (top) and inconsistent (bottom) items for aware (left) and unaware (right) participants by condition in the fill-in-the-blank task in Experiment 3.



Note. Error bars show 95% confidence intervals. Dotted lines represent the chance level.

We also ran glmer models for the consistent and inconsistent datasets for each of the SE and GR groups. For consistent items, we found strong evidence for an awareness effect in the SE condition, $BF_{(0,\ 1.23)} = 2663.77$, *RR* [0.24, > 4.59] (model coefficient: $\beta = 5.94$, $SE = 1.12$, $p < .001$). There was strong evidence for above-chance performance for aware participants,

$BF_{(0,1.81)}$ = 1052614, *RR* [0.25, > 4.59] (model intercept: β = 11.52, *SE* = 1.64, $p < .001$), and moderate evidence for above-chance performance for unaware participants, $BF_{(0, 1.81)}$ = 6.19, *RR* [0.19, 4.12] (model intercept: β = 0.85, *SE* = 0.35, $p < .05$). Similarly, for participants in the GR condition, there was strong evidence for an awareness effect on accuracy in consistent items, $BF_{(0,1.23)}$ = 116849.20, *RR* [0.19, > 4.59] (model coefficient: β = 6.92, *SE* = 1.09, $p < .001$). Breaking this down, there was strong evidence for above-chance performance in aware participants, $BF_{(0,1.81)}$ 2673.99, *RR* [0.33, > 4.59] (model intercept: β = 8.02, *SE* = 1.56, $p < .001$). For unaware participants, there was moderate evidence for the null, suggesting that they performed at chance, $BF_{(0,1.81)}$ = 0.14, *RR* [0.68, > 4.59] (model intercept: β = -0.46, *SE* = 0.48, $p = .34$).

For inconsistent items, we again found strong evidence for an awareness effect on accuracy in the SE condition, $BF_{(0,1.23)}$ = 55268.39, *RR* [0.15, > 4.59] (model coefficient: β = 3.85, *SE* = 0.71, $p < .001$). There was strong evidence that aware participants performed above chance, $BF_{(0,0.89)}$ = 49.80, *RR* [0.31, > 4.59] (model intercept: β = 5.80, *SE* = 1.27, $p < .001$). There was also moderate evidence for above-chance performance for unaware participants, $BF_{(0,0.89)}$ = 6.05, *RR* [0.21, 2.50] (model intercept: β = 0.76, *SE* = 0.34, $p < .05$). For the GR condition, we also found strong evidence for an awareness effect on inconsistent items, $BF_{(0,1.23)}$ = 14224493, *RR* [0.12, > 4.59] (model coefficient: β = 5.01, *SE* = 0.75, $p < .001$). Aware participants again showed above-chance performance, $BF_{(0,0.89)}$ = 34.84, *RR* [0.35, > 4.59] (model intercept: β = 7.64, *SE* = 1.55, $p < .001$). In contrast, there was inconclusive evidence whether unaware participants performed above chance, $BF_{(0, 0.89)}$ = 0.42, *RR* [0, 1.14] (model intercept: β = 0.10, *SE* = 0.30, $p = .74$).

We also looked at the interaction between awareness and frequency/condition in both consistent and inconsistent items in each of the SE and GR conditions. None of these analyses reached conclusive evidence for either H_1 or the null hypothesis (see Appendix 4E).

Relationship between learning and word meaning knowledge

In the word meaning task, participants' mean accuracy was 21.40%, suggesting that there was only limited lexical knowledge from the short exposure session. We predicted that there could be a relationship between the accuracy in the fill-in-the-blank task and the word meaning task, which could be either a positive or negative correlation. Thus, we tested this with two-tailed Bayes factors. Our analyses showed moderate evidence for the null for a correlation in both consistent items, $r_s(45) = 0.085$, $z_r = 0.085$, $BF_{(0,\ 0.57)} = 0.30$, RR [0.50, > 4.59], and inconsistent items, $r_s(45) = 0.047$, $z_r = 0.047$, $BF_{(0,\ 0.57)} = 0.27$, RR [0.45, > 4.59]. Based on these findings, we can conclude that learning the semantics of the artificial words is not associated with graphotactic learning.

Discussion

This experiment investigated whether native English-speaking participants could learn graphotactic patterns via exposure, and the impact of semantics on this learning. Modelling aspects of the noun pluralisation patterns in French, we created an artificial orthography using Latin alphabets where nouns could be pluralised by adding a single consonant as a suffix, and the noun endings determine the choice of consonant. Critically, these plural noun suffixes were only present in the written form, but not in the auditory input. Participants were either exposed to just the artificial words (GR condition), or the artificial words alongside reference pictures which made it clear that the consonants were plural affixes (SE condition).

Participants in both conditions were successful in learning the graphotactic pattern, as evidenced by strong evidence for above-chance performance in both conditions for both consistent and inconsistent items. They were also sensitive to the fact that there were two plausible suffixes in the inconsistent items, but one suffix was more dominant than the other, though there was no conclusive evidence of frequency effects.

The finding that individuals can become sensitive to graphotactic patterns after a short exposure is consistent with Singh et al. (2021). However, learning is notably stronger in the current experiment: Adult participants in Singh et al. (2021) achieved a mean accuracy of approximately 0.55 (with chance level at 0.50), whereas participants in our experiment reached a mean accuracy of 0.67 in consistent items (with chance level of 0.167) and 0.75 for inconsistent items (with chance level of 0.333). This is surprising given that both experiments examined a similar type of graphotactic pattern (where the final consonants are conditioned by the preceding vowels) and our graphotactic patterns are more complex, with a greater number of patterns and the additional manipulation of consistency and frequency. One possible explanation for this stronger learning effect is that the final consonant in our stimuli has the morphological status as a plural suffix, potentially making the pattern easier to extract. However, this cannot fully explain the results as strong learning was also observed in the GR condition where no semantic cues were provided. A more plausible explanation is that seeing the noun consecutively – first in its singular form (without the additional consonant) and then in its plural form (with the additional consonant) – may have given the additional consonant a suffix status even without the semantic cues. This makes the additional consonant more salient and thus facilitates learning of the graphotactic constraints.

The fact that we did not see evidence for the benefit of semantic grounding in our experiment to some extent contrasts with the findings of Klasen et al. (2023) and Fayol et al. (2006), who showed that children were better at spelling plural suffixes in French nouns, as compared to adjectives. They attributed this effect to the greater semantic grounding of nouns, as they refer to tangible objects, in addition to the fact that nouns have higher frequency and are introduced earlier in the curriculum. Given that we do not find evidence for *no* difference between the learning conditions in our data, we cannot draw strong conclusions. However, one question is whether participants were actually aware of the semantic cues. We know that participants achieved very low accuracy when asked to recall the word meanings at post-test (and there was also no conclusive evidence for a positive correlation between graphotactic learning and their ability to recall word meanings). While this is unsurprising given the brief exposure, it suggests that participants developed limited lexical knowledge of the artificial noun phrases. This raises the possibility these pluralisation patterns may not have been more semantically grounded than in the condition where participants were only exposed to the written graphotactic patterns. However, although we did not directly test whether participants recognised the additional suffixes as plural markers, over 70% of our participants were able to indicate that the determiners “yim” and “vop” referred to singularity and plurality in the word meaning task. This finding suggests that the majority of participants in the SE condition did indeed learn that some elements of the phrases were connected to the singular/plural meaning, therefore distinguishing the two learning conditions.

Another more general explanation for not seeing differences between conditions is that performance is generally too high in both conditions (i.e., ceiling effects). It is worth noting that this is largely driven by a high proportion of participants (60% across the conditions) for whom the awareness questionnaire at post-test indicated that they had become explicitly

aware of the graphotactic patterns. As in previous studies, our analyses revealed that aware participants outperformed unaware participants across both item types and learning conditions. Nevertheless, there was also moderate evidence that even unaware participants performed above chance, at least in the group who saw semantic cues during exposure. This provides tentative evidence that while being able to explicitly verbalise these patterns leads to better generalisation at post-test, some incidental learning is also present.

To briefly summarise, our experiment demonstrates that participants quickly became sensitive to graphotactic patterns conditioning consonant endings after a short exposure, regardless of whether these graphotactic patterns were presented with semantic cues which grounded them as plural affixes. Since the consonant endings were "silent", we view these patterns as graphotactic rather than phonotactic. However, given the fact that we used the familiar Latin alphabet, it remains possible that participants themselves articulated the words (out loud or covertly) while learning. If so, the learning we observed may be of phonological regularities, but not purely graphotactic. This potential use of unintended phonological cues in learning may also contribute to the high accuracy we observed for participants in both learning conditions. As noted above, this may lead to ceiling effects which block us from seeing facilitative effects from semantic cues. Therefore, in the next experiment, we explored whether the same learning patterns could be observed when all phonological cues are eliminated.

Experiment 4

To investigate whether graphotactic learning can still occur when unintended phonological cues are eliminated, we conducted a pre-registered experiment (https://osf.io/yf39j) replicating Experiment 3 using the Brussels Artificial Character Sets (Vidal et al., 2017). The

BACS-2 font contains novel symbols that closely match the visual characteristics of Latin letters, allowing us to remove any phonological cues that participants may have used from Latin alphabets to facilitate learning in Experiment 3.

Method

Sample size and power calculations

The procedure for estimating sample size and applying the stopping rule was the same as that described in Experiment 3.

Participants

The selection criteria, recruitment, consent and compensation procedures were the same as Experiment 3. Fifty-seven[13] native English-speaking adults (28 female; mean age = 37.41; *SD* = 13.03) were recruited through Prolific (*N* = 50) and the research participation scheme at the University (*N* = 7) (see Appendix 4A for additional information on participants' language backgrounds). Most of the participants (*N* = 36) have obtained or were currently pursuing an undergraduate or postgraduate degree at the time of the study. As per our pre-registered criteria, we excluded 4 participants who indicated French as their second language and 2 participants who did not meet the 75% accuracy benchmark in the visual recognition task. One participant was removed due to learning disabilities. The final sample contains 50 participants with 24 participants in the GR condition and 26 participants in the SE condition.

[13] As stated in our pre-registration, we planned to analyse the data after testing 56 participants (half of the total sample size) and would stop testing if we found conclusive evidence for the semantic effect. Since we did find such evidence at this stage, we discontinued testing.

Materials

The artificial words used in the exposure phase and the fill-in-the-blank task were identical to those in Experiment 3, except presented in the BACS-2 font (Vidal et al., 2017). This ensures consistent word lengths and the number of letter/symbol repetitions across experiments for better comparability. In addition, given that the goal of using symbols is to eliminate all phonological cues, no additional auditory input was provided in this experiment.

For the visual recognition task, which is used as a baseline attention check, modifications were made due to the greater difficulty of developing strong mental representations of lexical items comprised of novel symbols within such a short exposure phase. Specifically, we adjusted the foil items to create a stronger contrast with the exposure items. Three types of foil items were created: (1) a novel five-letter article combined with a novel noun ending with a learned suffix (e.g., “coreb homa”), (2) a learned article at the wrong position combined with a novel noun ending with a novel suffix (e.g., “aining yim”) and (3) phrases written in a different font (BACS-1 font). There are four items for each type of foil items (see Appendix 4C for the full list). Moreover, instead of applying a 75% overall accuracy benchmark as in Experiment 3, we included participants who met this threshold in either the exposure or foil items in the recognition task.

Procedure

Participants completed all tasks in the same order as Experiment 3, except for the auditory recognition task, which was removed as there was no auditory input during the exposure phase in this experiment.

As all the stimuli were presented in symbols, we expected participants to take more time to read each phrase compared to when they were written in Latin alphabets. There was also no auditory input in each trial. Therefore, modifications were made to the presentation order and timing of each trial in the exposure phase. In the GR condition (Fig. 4.2, bottom left), each trial began with a fixation point (500ms), followed by the written singular noun phrase (2000 ms), a short blank (250 ms) and and then the written plural noun phrase (2000 ms). In the SE condition (Fig. 4.2, bottom right), each trial started with a fixation point (500 ms), followed by the picture of a singular object (1500 ms), and then the written singular noun phrase was displayed alongside the object (2000 ms). This was followed by the picture indicating multiple objects (1500 ms), then the written plural noun phrase was displayed alongside the objects (2000 ms). For the fill-in-the-blank task, similar modifications were also made. Each trial began with a fixation point (1000 ms), followed by the written singular noun phrase (2000 ms), followed by a short blank (500 ms). The written plural noun phrase with the missing word was presented until participants responded.

Statistical analyses

Analyses followed the same plan as Experiment 3. As pre-registered, we did not routinely conduct analyses comparing the two experiments, but only did so when the learning patterns differed across the experiments.

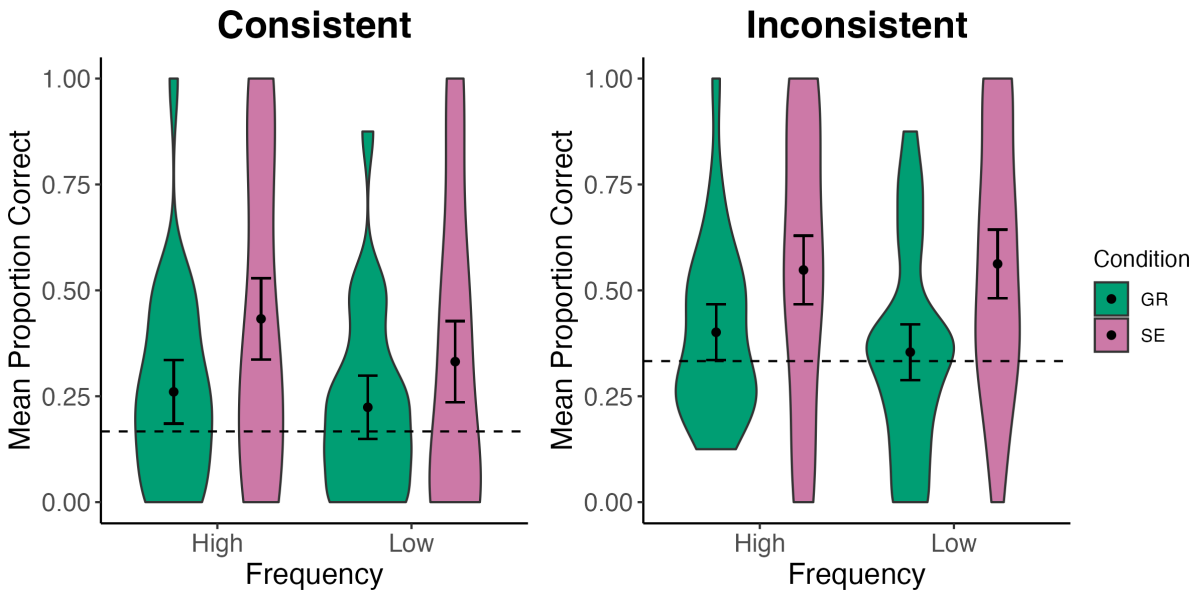
Results

Fill-in-the-blank overall performance

Figure 4.7 shows the mean proportion of correct responses in the fill-in-the-blank task for consistent and inconsistent items in Experiment 4, split by frequency and condition. Overall,

we found strong evidence that participants performed above chance in inconsistent items, $BF_{(0,0.89)}$ = 69.16, *RR* [0.06, > 4.59] (model intercept: β = 0.59, *SE* = 0.18, $p < .01$). The evidence for overall above-chance learning in consistent items approached 3 but was inconclusive, $BF_{(0,1.81)}$ = 2.70, *RR* [0, > 4.59] (model intercept: β = 0.52, *SE* = 0.24, $p < .05$).

Figure 4.7. Mean proportion of correct responses for consistent items (left) and inconsistent items (right), split by frequency and condition, in the fill-in-the-blank task in Experiment 4.



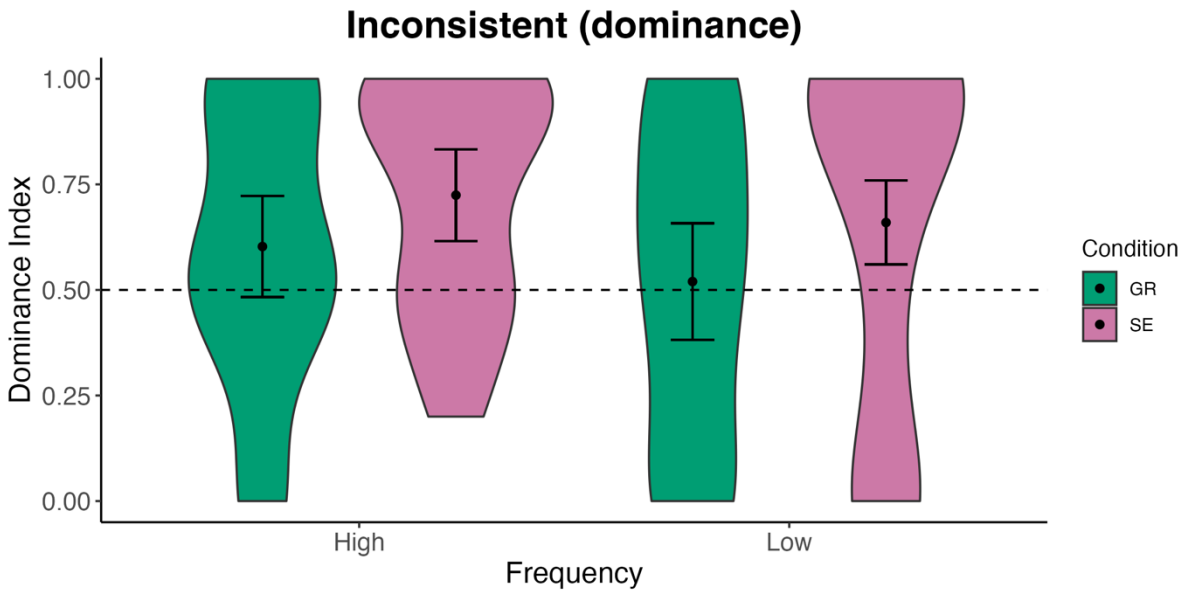
Note. Error bars show 95% confidence intervals. Dotted lines represent the chance level.

For the frequency effect, the evidence for a high-frequency benefit was ambiguous for both consistent, $BF_{(0,0.91)}$ = 2.20, *RR* [0, > 4.59] (model coefficient: β = 0.52, *SE* = 0.31, $p = .09$) and inconsistent item, $BF_{(0,0.45)}$ = 0.53, *RR* [0, 0.77] (model coefficient: β = 0.08, *SE* = 0.18, $p = .67$).

As for the effect of condition (SE/GR), we found inconclusive evidence in consistent items, $BF_{(0,\ 0.41)}$ = 2.46, *RR* [0, > 4.59] (model coefficient: β = 0.80, *SE* = 0.47, $p = .09$). Breaking this down, there was inconclusive evidence for above-chance performance for both the SE

condition, $BF_{(0,\ 1.81)}$ = 1.89, *RR* [0, > 4.59] (model intercept: β = 0.79, *SE* = 0.46, *p* = .09), and the GR condition, $BF_{(0,\ 1.81)}$ = 0.46, *RR* [0, 2.50] (model intercept: β = 0.27, *SE* = 0.22, *p* = .22). In contrast, there was moderate evidence for an effect of condition in inconsistent items, $BF_{(0,\ 0.41)}$ = 8.05, *RR* [0.18, > 4.59] (model coefficient: β = 0.91, *SE* = 0.36, *p* < .05). Breaking this down, there was strong evidence for above-chance performance in the SE condition, $BF_{(0,\ 0.89)}$ = 84.49, *RR* [0.12, > 4.59] (model intercept: β = 1.07, *SE* = 0.32, *p* < .001), whereas the evidence for above-chance performance in the GR condition was inconclusive, $BF_{(0,\ 0.89)}$ = 0.39, *RR* [0, 1.06] (model intercept: β = 0.13, *SE* = 0.19, *p* = .49). This suggests that semantic cues facilitated the learning of graphotactic patterns, at least in the inconsistent items. Exploratory analyses of the interactions between condition and frequency for both item types did not result in conclusive evidence supporting either H_1 or null (see Appendix 4E).

Figure 4.8. Dominance index in inconsistent items, split by frequency and condition, in the fill-in-the-blank task in Experiment 4.



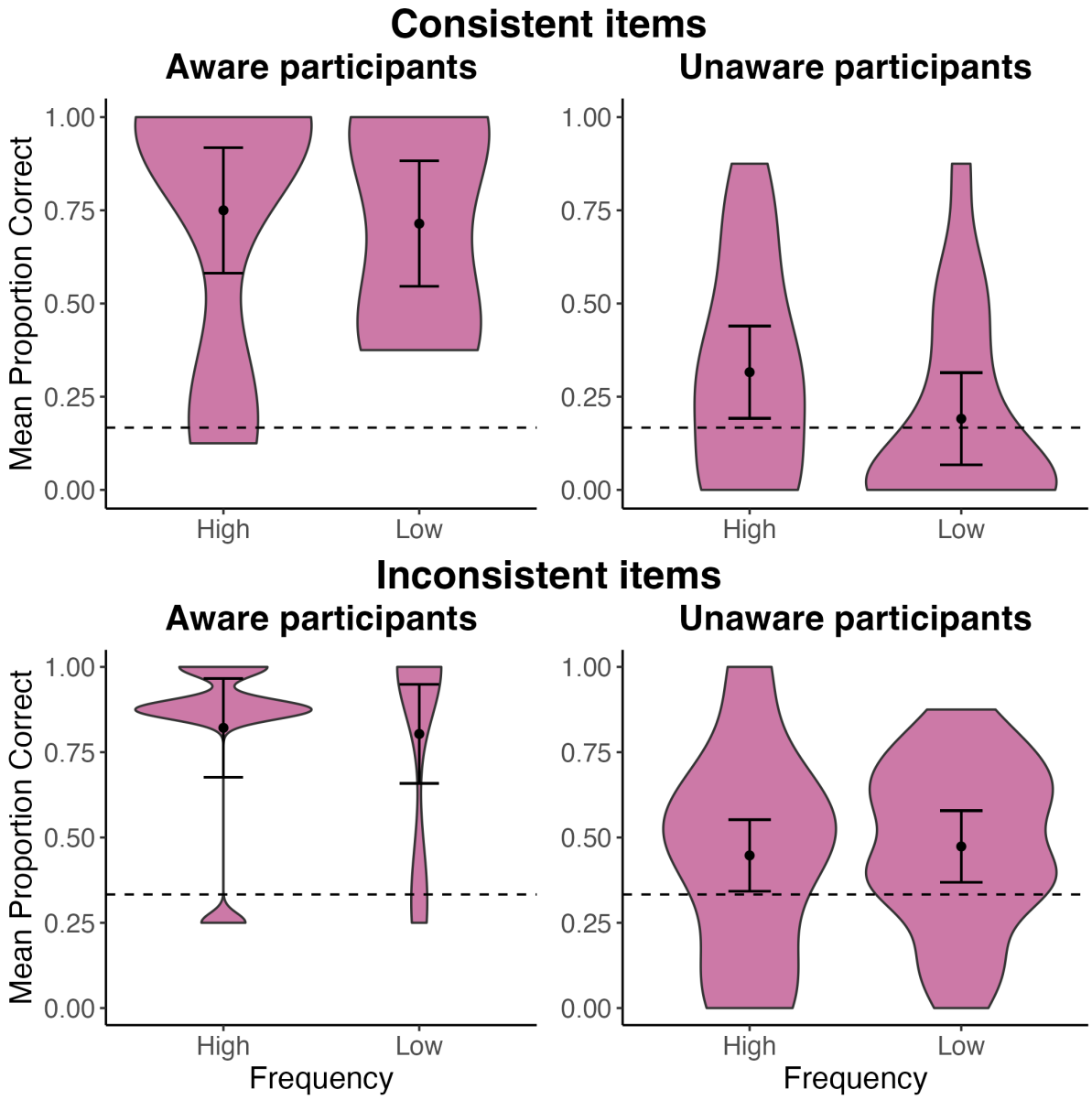
Note. Error bars show 95% confidence intervals. Dotted lines represent the chance level.

Figure 4.8 shows the dominance index in Experiment 4, split by frequency and condition. There was strong evidence showing that participants chose the dominant suffix over the non-dominant suffix in inconsistent items, $BF_{(0,\ 1.10)} = 26.71$, *RR* [0.13, > 4.59] (model intercept: $\beta = 0.87$, $SE = 0.30$, $p < .01$). There was no conclusive evidence for a frequency effect on the dominance index, $BF_{(0,\ 0.55)} = 0.64$, *RR* [0, 1.27] (model coefficient: $\beta = 0.04$, $SE = 0.41$, $p = .92$). There was again no conclusive evidence for an effect of condition or an interaction between condition and frequency on the dominance index, as suggested in our exploratory analyses (Appendix 4E).

Performance by awareness status

We determined awareness status using the same criteria as in Experiment 3. Seven out of 26 participants in the SE condition were aware of the graphotactic patterns, whereas no participants in the GR condition were able to describe the patterns. Therefore, the following analyses were conducted with participants in the SE condition only.

Figure 4.9. Mean proportion of correct responses in consistent (top) and inconsistent (bottom) items for aware (left) and unaware (right) participants in the SE condition in the fill-in-the-blank task in Experiment 4.



Note. Error bars show 95% confidence intervals. Dotted lines represent the chance level.

As shown in Figure 4.9, for consistent items, there was strong evidence that aware participants performed better than unaware participants, $BF_{(0,\ 1.23)} = 393.37$, *RR* [0.24, > 4.59] (model coefficient: $\beta = 3.76$, $SE = 0.89$, $p < .001$). Participants who were aware of the patterns performed above chance, $BF_{(0,\ 1.81)} = 67.49$, *RR* [0.38, > 4.59] (model intercept: β =

3.72, *SE* = 1.10, $p < .001$). There was moderate evidence for the null suggesting that unaware participants performed at chance, $BF_{(0,\ 1.81)} = 0.16$, *RR* [0.81, > 4.59] (model intercept: β = -0.29, *SE* = 0.46, $p = .53$). Similarly, for inconsistent items, there was strong evidence that aware participants outperformed unaware participants, $BF_{(0,\ 1.23)} = 138.24$, *RR* [0.21, > 4.59] (model coefficient: β = 2.31, *SE* = 0.65, $p < .001$). However, the evidence for above-chance performance was inconclusive for aware participants, $BF_{(0,\ 0.89)} = 2.28$, *RR* [0, 1.24] (model intercept: β = 4.22, *SE* = 1.93, $p < .05$), and unaware participants, $BF_{(0,\ 0.89)} = 2.73$, *RR* [0, > 4.59] (model intercept: β = 0.49, *SE* = 0.26, $p = .06$). This finding is somewhat surprising given that Figure 4.9 suggests above-chance learning. However, these results should be interpreted with caution, as they are based on only seven participants who were coded as aware in the post-test questionnaire. None of the exploratory analyses examining the interaction between awareness and frequency reached conclusive evidence for either H_1 or the null hypothesis (see Appendix 4E).

Relationship between learning and word meaning knowledge

Given that artificial words were presented in symbols, we predicted that it would be challenging for participants to develop strong lexical representations of these artificial words after a short exposure. Our results confirmed this, as only one out of 26 participants achieved over 30% accuracy in this task. Most participants scored below 5% (*M* = 4.08%, *SD* = 6.23%), with the majority of correct responses coming from recalling the meaning of the two artificial articles, “yim” and “vop”, which represented singular and plural forms. Due to such low accuracy and lack of variability in participants’ performance, we were unable to conduct the planned correlation analyses to examine the relationship between graphotactic learning and word meaning knowledge.

Exploratory analyses comparing Experiments 3 and 4

Our pre-registration stated that we would not combine or compare experiments in the initial analyses, but that we might do this in a targeted way to further understand our data. To explore potential differences in learning outcomes between Experiment 3 and Experiment 4, we tested the effect of Experiment to examine whether the absence of phonological information led to differences in learning effects. In two separate models (one for consistent items and another for inconsistent items), the centred variable for Experiment was included as a fixed effect, with random intercepts for participants. There was strong evidence that performance was better in Experiment 3 than Experiment 4 for consistent items, $BF_{(0,0.40)}$ = 593.87, *RR* [0.10, > 4.59] (model coefficient: β = 2.83, *SE* = 0.51, $p < .001$), and inconsistent items, $BF_{(0,0.40)}$ = 5282.93, *RR* [0.07, > 4.59] (model coefficient: β = 2.06, *SE* = 0.37, $p < .001$).

Another potentially important difference concerns the moderate evidence supporting a condition effect (i.e., SE participants outperformed GR participants) in the inconsistent items in Experiment 4, whereas the evidence for this effect was inconclusive in Experiment 3. To test whether the effect of condition in inconsistent items was reliably different across the two experiments, in a separate model, we tested for an interaction between Condition and Experiment. For this, the centred variables for Condition, Frequency and Experiment were entered as fixed effects along with their interactions, and random intercepts and random slopes for Frequency by participants. There was no conclusive evidence that SE participants outperformed GR participants to a greater extent in Experiment 4 than Experiment 3, $BF_{(0,0.20)}$ = 0.77, *RR* [0, 0.98] (model coefficient: β = -1.08, *SE* = 0.82, $p = .19$). Thus, though we have evidence for a difference between conditions in Experiment 4 but not Experiment 3, we cannot draw strong conclusions that this is due to the change in stimuli.

Discussion

Experiment 4 extended Experiment 3 by replicating the experiment with artificial orthography (BACS-2 font) in place of the Latin alphabet. These novel symbols eliminated any unintended phonological cues, allowing us to investigate whether participants could still learn the noun ending-suffix mappings when the graphotactic patterns are purely visual.

Many of the results of Experiment 4 were similar to those of Experiment 3. Overall, participants became sensitive to the statistical regularities embedded in the artificial orthography after a short exposure, despite the absence of linguistic information. This is reflected in the strong evidence for above-chance performance in inconsistent items, although the evidence for consistent items was inconclusive. A closer look at the inconsistent items revealed that participants were again more likely to choose the dominant over the non-dominant suffix in the inconsistent items. There was again no conclusive evidence for a high-frequency benefit on learning the graphotactic patterns.

As in Experiment 3, some participants were able to verbalise the graphotactic patterns in the post-test awareness questionnaire. However, unlike Experiment 3, where a similar number of participants developed explicit awareness of the graphotactic patterns in both learning conditions, we found aware participants only in the SE condition. More specifically, only about 27% of the participants in the SE condition demonstrated this awareness, which is significantly lower than the 62% in Experiment 3. Consistent with previous findings, participants who could verbalise the patterns outperformed those who could not at post-test. However, this finding should be taken with caution as there were only 7 participants in Experiment 4 who were coded as aware.

The key difference in the results between the two experiments is the impact of semantic cues on graphotactic learning. In Experiment 3, there was no conclusive evidence that exposure to semantic cues alongside the artificial words facilitated participants' learning of the graphotactic patterns. However, in Experiment 4, semantic cues had a clear facilitative effect in inconsistent items: there was evidence for stronger learning of the graphotactic conditioning of the final consonant, when that consonant could be interpreted as a plural morpheme. This is in line with our prediction that morphological status can affect graphotactic learning. It is possible that the fact that this effect is seen here but not in Experiment 3 is somehow due to the absence of linguistic information in the artificial words used in this version of the artificial orthography. However, our exploratory analysis did not provide conclusive evidence that the effect of condition was stronger in Experiment 4 than in Experiment 3 for inconsistent items. In other words, we cannot confidently conclude that the difference in performance between the SE and GR group was greater than the corresponding difference in Experiment 3.

General Discussion

In two experiments, we examined whether native English-speaking adults could learn graphotactic constraints through exposure. There were three key manipulations. The first was a within-participant manipulation based on the frequency and consistency of mappings between noun endings and suffixes in the graphotactic constraints. Some mappings appeared more frequently than others in the exposure phase (e.g., the pairing of the noun ending "a" with the suffix -x occurred 48 times, while the pairing of "i" and -v appeared only 24 times). These mappings can also be consistent (e.g., the noun ending "a" can only take the suffix -x) and inconsistent (e.g., the noun ending "o" can take either -k or -d as its suffix). The second was a between-participant manipulation examining the impact of semantic cues on

graphotactic learning. In each experiment, one group of participants was exposed to the graphotactic constraints embedded in artificial words (GR group) with no reference, while another group saw referents alongside the written phrases, which made it indicate that the endings being learned were plural morphemes (SE group). The third was another between-participant manipulation exploring the impact of phonological cues on this learning. The artificial orthography was presented in Latin alphabets (Experiment 3) and symbols (Experiment 4), with the latter designed to eliminate any unintended phonological cues. Further to these experimental manipulations, we also administered a post-experiment questionnaire to assess whether participants could verbalise the graphotactic patterns and how this explicit awareness impacted their learning.

Three key results emerged from our experiments. First, participants successfully learned the graphotactic constraints and became sensitive to the statistical regularities within them after a short exposure. In the generalisation post-test, we found strong evidence for above-chance learning in both consistent and inconsistent items in Experiment 3, and for inconsistent items in Experiment 4. Within the inconsistent items, participants tended to choose the dominant suffix more often than the non-dominant suffix, which further suggests their sensitivity to the statistical regularities embedded in the input. Second, there is evidence that semantic cues impacted graphotactic learning, but the evidence for this effect is limited to inconsistent items in Experiment 4. Third, explicit awareness – measured by participants' ability to verbalise the graphotactic patterns in a post-experiment questionnaire – led to better generalisation of the patterns at post-test. In both experiments, participants who could verbalise the patterns consistently performed better than those who could not. There was also moderate evidence that participants who could not verbalise the patterns still achieved above-chance

performance, but this was only found in the SE group in Experiment 3. We discuss the implications of these findings below.

Statistical learning of graphotactic patterns

Across the two experiments, our findings showed that people developed sensitivity to the target graphotactic patterns after a short exposure (15 – 20 minutes), as evidenced by their above-chance performance in the generalisation task. Within inconsistent items, participants tended to choose the dominant suffix over the non-dominant suffix, despite both being plausible suffixes. Consistent with previous learning experiments (Samara et al., 2019; Samara & Caravolas, 2014; Singh et al., 2021), these findings show that statistical learning of graphotactic patterns occurs rapidly. Furthermore, the evidence of above-chance performance in Experiment 4 confirms that this learning can be purely graphotactic, given that artificial symbols could not be covertly articulated. This aligns with other learning experiments, such as Chetail (2017), which also demonstrated that people could become sensitive to bigram frequencies in the form of visual statistical regularities.

However, evidence of learning was much stronger in Experiment 3 than Experiment 4 for both consistent and inconsistent items, which suggests that unintended phonological cues had an impact on their learning. In other words, although the suffixes were absent in the auditory input in Experiment 3, it is likely that participants still articulated them (whether out loud or covertly) and used these phonological cues as an additional regularity to facilitate this learning. Nevertheless, the sample size for Experiment 3 was considerably larger than that of Experiment 4, which may have also contributed to the stronger effects.

Impact of semantic cues on graphotactic learning

Previous research has shown that the English writing system is rich in meaningful regularities (Ulicheva et al., 2020), and that individuals are sensitive to these form-meaning associations, which influence their spelling choices (e.g., Treiman et al., 2021). This indicates that in natural language, spelling patterns are not only governed by graphotactic constraints, but also by meaning. To examine how morphological and graphotactic regularities may interact and impact the learning of spelling patterns, our study contrasted two learning conditions. In the GR condition, participants only saw the written phrases so that learning the mappings of the noun ending and suffix is purely graphotactic. In the SE condition, singular noun phrases were paired with images of single objects, while plural phrases were paired with images of multiple objects. In this case, the additional suffix represents a morphological regularity that is further governed by graphotactic patterns.

Overall, we found limited evidence for a difference between the two conditions at post-test. The only conclusive evidence was observed in inconsistent items in Experiment 4, where there was strong evidence for above-chance learning in the SE condition, but the evidence remained inconclusive in the GR condition. In all other comparisons, the evidence for a condition effect remained inconclusive (Experiment 3 consistent items: $BF = 1.09$, Experiment 3 inconsistent items: $BF = 0.83$, Experiment 4 consistent items: $BF = 2.46$). The observed difference between the SE and GR conditions may suggest that participants in the SE condition recognised the additional suffix as a morphological regularity indicating plurality. The reference to tangible objects likely made it easier for learners to understand the function and importance of the suffix (Klasen et al., 2023), which in turn may have helped them focus on the additional graphotactic patterns.

However, if this explanation holds, why was the same difference not observed in the consistent items in Experiment 4? It is important to note that the evidence here was ambiguous, and in fact tended towards more evidence for H1 than the null ($BF = 2.46$). Tentatively, this may be due to the difference in scoring criteria between item types. In the generalisation post-test, inconsistent items were scored more leniently as both plausible suffixes were accepted as correct responses whereas consistent items required an exact match to be considered correct. This may have made the inconsistent items easier for participants to perform correctly, although consistent mappings would typically be easier to learn. As a result, there was a stronger learning effect for inconsistent items, which may have also amplified the observed difference between the SE and GR groups.

Similarly, we also found inconclusive evidence as to whether there was a difference between the SE and GR conditions in Experiment 3, meaning we cannot determine whether participants in the SE condition outperformed those in the GR condition, or whether they had comparable performance. We also note that the comparison with Experiment 4 was inconclusive. This means we cannot rule out that participants in this experiment do not also benefit from the suffix being semantically grounded. Tentatively, the availability of unintended phonological cues in this experiment might have made the graphotactic patterns much easier to detect and learn, as is evident from the fact that participants achieved over 60% accuracy across all item types in both conditions after only 15 to 20 minutes of exposure. This overall high accuracy may not allow space to see the added benefit of the semantically grounded suffix.

Explicit knowledge from statistical learning

Statistical learning is often assumed to produce minimal explicit knowledge of the underlying statistical structure of the stimuli (e.g., Conway & Christiansen, 2005; Turk-Browne et al., 2005), such that people are unaware of what they have learned (Treiman, 2018a). However, evidence from artificial orthography learning experiments challenges this view. For example, Singh et al. (2021, Experiment 1) used a post-experiment questionnaire and found that approximately 30% of adult participants were able to describe the graphotactic patterns after incidental exposure. While unaware participants (i.e., those who could not verbalise the patterns) achieved about 60% accuracy in a fill-in-the-blank task, aware participants were close to ceiling. Law et al. (2025) observed similar results when examining the learning of form-to-meaning mappings and graphotactic patterns through exposure. They found that approximately 10-20% of participants could verbalise the nonphonological form-meaning regularities, and these participants were successful in generalising the patterns to novel items. Contrary to Singh et al. (2021, Experiment 1), there was limited evidence that unaware participants successfully learned the form-to-meaning mappings.

As in previous studies, our participants were not informed that patterns were embedded in the alien language, nor were they instructed to learn them or made aware that they would later be tested on their knowledge of these patterns. We applied a relatively lenient criterion for assessing participants' ability to verbalise the graphotactic patterns where we classified them as "aware" as long as they mentioned any relationship between the noun ending and suffix. Similar to Singh et al. (2021) and Law et al. (2025), some participants developed explicit awareness of the target patterns. In Experiment 3, over 60% of participants in both the GR and SE conditions were able to verbalise the graphotactic patterns. In contrast, in Experiment 4, only 20% of the participants in the SE condition was able to do so, and none in the GR

condition. When analysing the effect of this awareness status on participants' performance in the generalisation task, those who could verbalise the patterns consistently performed better than those who could not. However, as seen in Law et al. (2025), the evidence for unaware participants performing above chance is limited. We only found moderate evidence for above-chance learning in unaware participants in the SE condition in Experiment 3, with inconclusive evidence or evidence for the null in other analyses.

Collectively, the evidence reviewed above suggests that the statistical learning process does not only result in implicit knowledge of the target patterns, but explicit knowledge can also emerge. One possible explanation for this observation is that explicit learning mechanisms were at play during exposure. That is, despite the absence of explicit learning instructions, some participants had conscious intention to detect regularities in the input information. Our post-experiment questionnaire provided some evidence supporting this: when asked how they determined which alien word better fit the phrase in the testing phase, some participants reported that they had already noticed some patterns during the exposure phase, and they used this knowledge in completing the testing phase.

However, the inferences we made regarding explicit learning based on post-experiment questionnaires should be interpreted with caution. First, as noted in Law et al. (2025), the inability to verbalise the target patterns does not necessarily mean that explicit learning did not occur. Participants may have struggled to verbalise the patterns for various reasons, including differences in their oral language abilities. This challenge is even greater in Experiment 4 when the stimuli were entirely presented in symbols, making it particularly difficult to articulate the patterns. Second, while we suggest that explicit learning may have occurred for some participants in our experiments, this fundamentally differs from explicit

learning conditions where participants are informed of all the patterns beforehand (e.g., Singh et al., 2021, Experiment 3). As in Law et al. (2025), explicit knowledge acquired through exposure is incomplete in our case. For example, in the post-test awareness questionnaire, some participants only reported noticing the dominant suffix in the inconsistent items, rather than recognising that there were two plausible suffixes. This suggests that while some participants were conscious with searching for patterns in the input, they did not necessarily become aware of all possible patterns in the input. Future research should compare learning that occurs through emerging explicit awareness versus learning where participants are explicitly informed of the patterns before exposure. Such comparisons would not only clarify how explicit learning contributes to people's ability to generalise these orthographic patterns, but it also shed light on the interaction between implicit and explicit learning.

Conclusion

In conclusion, our study provides evidence supporting the view that the statistical learning processes underlie graphotactic learning. After a short exposure, native English-speaking adult participants became sensitive to the statistics in the graphotactic patterns modelling aspects of French noun pluralisation patterns. When semantic information was available, participants recognised that the graphotactic patterns were linked to a morphological regularity. There was some limited evidence that this further facilitated their learning of the graphotactic patterns. Additionally, explicit awareness – measured by participants' ability to verbalise the patterns at post-test – led to better generalisation of these patterns. Future research should further explore the interaction between nonphonological graphotactic and morphological regularities and its impacts on the learning-to-spell process.

Chapter 5. General Discussion

This thesis systematically reviewed evidence from existing artificial orthography learning experiments on reading and spelling acquisition (Chapter 2) and adopted this paradigm to examine whether native English-speaking adults could learn semantic and graphotactic regularities (Chapters 3 & 4) through exposure to words embedding those patterns. Across all three studies, my research has explored how evidence from this paradigm informs our understanding of statistical learning in reading and spelling acquisition. This section first outlines the key findings from each of the three studies. These findings are then synthesised to highlight their contribution to our understanding of spelling acquisition and statistical learning within this context, along with suggestions for future research directions.

Summary of Thesis Findings

Chapter 2 presented a systematic review of artificial orthography learning experiments investigating reading and spelling acquisition. The review began by outlining the key characteristics of these experiments, and the factors which have been explored for their impacts on orthographic learning within this paradigm. Results showed that most experiments focused on orthography-phonology mappings in native English-speaking adults, reflecting the broader emphasis on alphabetic languages in literacy research (Share, 2008). Nonetheless, the paradigm offers valuable insights into orthographic learning, highlighting the impacts of both linguistic (e.g., consistency, frequency) and non-linguistic external factors (e.g., explicit instruction). A second aim of this review was to examine how this paradigm informs our understanding of statistical learning in reading and spelling acquisition. Findings from orthotactic learning experiments reveal that individuals can quickly develop sensitivity to statistical regularities in novel orthographies, supporting the view that statistical learning

processes underlie orthographic learning. However, there was considerable variability in experimental designs, particularly in the exposure phase, which could lead to learning processes that ranged from implicit to explicit. Methods for assessing and measuring learning also varied across experiments, further complicating comparisons and synthesis of findings. Additionally, among the 19 statistical learning experiments that investigated individual differences (according to our pre-set criteria), there was no consistent relationship between learning outcomes and individual difference measures in reading, spelling and other related skills, raising concerns about the external validity of this paradigm. This review highlights the need for future research to examine orthographic patterns beyond orthography-phonology mappings and to adopt more consistent approaches to testing and measuring learning.

Motivated by the limited research on orthographic patterns beyond orthography-phonology mappings, Chapter 3 investigated whether native English-speaking adults could simultaneously learn semantic (where possible spellings depend on grammatical word class) and graphotactic patterns (where possible spellings depend on earlier graphemes) from brief exposure to an artificial lexicon. The results from both Experiments 1 and 2 showed that participants successfully learned the semantic patterns, as evidenced by above-chance performance in generalisation at post-test. However, contrary to previous research, there was no evidence of graphotactic learning. Importantly, participants who could explicitly verbalise these patterns at test consistently outperformed those who could not at post-test. Semantic learning was also more salient when the semantic patterns were associated with nouns and verbs, as compared to adjectives and nouns. Overall, these findings support that statistical learning processes underlie orthographic learning but also highlight its limits. Specifically, learning two types of nonphonological regularities simultaneously may not be possible given the short exposure to the artificial lexicon.

The absence of graphotactic learning in Experiments 1 and 2 was somewhat surprising because other learning experiments (e.g., Samara et al., 2019; Samara & Caravolas, 2014; Singh et al., 2021) have found consistent evidence for this learning. To further investigate this, Chapter 4 comprised two artificial orthography learning experiments designed to examine whether native English-speaking participants could learn graphotactic constraints from exposure to an artificial orthography modelled on French pluralisation patterns. The experiments also investigated whether learning graphotactic patterns in meaningful contexts (where part of the graphotactic pattern functions as a plural suffix) facilitated learning of these patterns, compared to when the patterns were presented in meaningless contexts. Additionally, the influence of phonological cues was investigated by presenting the artificial orthography using Latin alphabets (Experiment 3) and symbols (BACS-2 fonts; Experiment 4). Across both experiments, participants developed sensitivity to the graphotactic patterns after brief exposure. Learning was stronger when the artificial orthography was presented in Latin alphabets, although some learning still occurred with symbols. Consistent with findings from Experiments 1 and 2, participants who could verbalise the patterns at post-test showed better generalisation. Contrary to expectations, there was limited evidence that graphotactic learning was better in meaningful contexts where part of the graphotactic pattern clearly indicated plurality. However, overall learning effects were stronger than in Experiments 1 and 2, possibly because the presentation format of the artificial noun phrases led participants to interpret part of the graphotactic patterns as a suffix, even without the semantic cues. This may have increased the salience of the graphotactic patterns, making them easier for participants to learn. Overall, findings from the two experiments reported in Chapter 4 support the view that statistical learning processes underlie orthographic learning.

Synthesis of Thesis Findings

Regularities beyond phoneme-grapheme correspondences in written language

Much of the existing literacy research has focused on alphabetic languages (Share, 2008) so it is unsurprising that a substantial body of work on reading and spelling acquisition has examined orthography-phonology mappings, as these regularities are central to reading and spelling in alphabetic languages. However, as Deacon et al. (2008) pointed out, writing systems such as English and French contain several types of regularity. Apart from phoneme-grapheme correspondences, English spelling also reflects the relationship between spelling and meaning. For example, the regular past tense in English is consistently spelled with the suffix -ed, despite its variable pronunciations (i.e., /t/, /d/ or /id/). In addition, graphotactic constraints, which concern permissible letter sequences, patterns and positional constraints in written words (Apel et al., 2006; Pacton et al., 2005), also govern English spelling patterns. For instance, while the letter "l" can appear as a doublet, the letter "h" cannot ("hill" vs *hhil).

The experimental findings presented in this thesis demonstrate that individuals are indeed sensitive to regularities in writing systems other than phoneme-grapheme correspondences. Across the experiments, native English-speaking adults successfully learned form-to-meaning mappings (Chapter 3) and graphotactic patterns related to noun pluralisation (Chapter 4), and they were able to generalise this knowledge to novel words after brief exposure. Importantly, the regularities in the artificial orthographies were designed to be independent of phonological cues. These findings align with the results of the systematic review reported in Chapter 2, which showed that both adults and children can quickly develop sensitivity to nonphonological regularities in artificial orthographies. Such regularities include graphotactic

patterns (e.g., Singh et al., 2021), bigram/trigram frequencies (e.g., Chetail, 2017), morphological structure (e.g., Lelonkiewicz et al., 2023), positional regularities in Chinese (e.g., He & Tong, 2017), and form-meaning regularities (e.g., Rastle et al., 2021). Taken together, these findings suggest that spelling acquisition extends beyond learning grapheme-phoneme correspondences. It also includes developing sensitivity to a broader range of orthographic patterns embedded within the writing system, some of which operate independently of phonology.

Statistical learning as a learning mechanism in spelling acquisition

Experiments using pseudoword choice/spelling tasks have shown that individuals are sensitive to graphotactic patterns (e.g., Hayes et al., 2006) and form-meaning regularities (e.g., Ulicheva et al., 2020), even when they are not able to verbalise them (Treiman & Wolter, 2018) and without the patterns having been explicitly taught (Cassar & Treiman, 1997). These findings are consistent with the triangle model (Harm & Seidenberg, 2004; Plaut et al., 1996; Seidenberg & McClelland, 1989), which posits that orthographic learning involves extracting statistical regularities in orthography-phonology, orthography-semantics, and orthotactic patterns via reading experience. Based on this view, statistical learning has been proposed as a key mechanism underlying spelling acquisition (Treiman, 2018a).

This thesis adopts a broad definition of statistical learning as the discovery of patterns in the input (Romberg & Saffran, 2010). This definition captures an individual's ability to extract regularities from exposure without any explicit instructions directly on the target patterns themselves, which could range from supervised to unsupervised learning. Based on this definition, the systematic review in Chapter 2 identified 31 artificial orthography learning experiments that examined orthotactic regularities. Across these experiments, participants

consistently demonstrated successful learning of various orthotactic regularities (e.g., Chetail, 2017; Singh et al., 2021), providing evidence that statistical learning processes contribute to orthographic learning. The experimental results presented in Chapters 3 and 4 of this thesis further support this claim, showing that adult participants can learn both form-meaning regularities and graphotactic patterns after brief exposure. Moreover, these findings offer several key insights into our understanding of statistical learning in orthographic learning:

First, statistical learning can occur with and without intent. The systematic review (Chapter 2) identified several orthotactic learning experiments with incidental learning conditions, where participants performed unrelated tasks while being exposed to the stimuli (e.g., Singh et al., 2021). These experiments assumed that participants extracted regularities from input without conscious awareness (Turk-Browne et al., 2005) and thus learning occurred as a "as a byproduct of mere exposure'' (Saffran et al., 1999). On the other hand, some experiments (e.g., Vidal et al., 2021) included in the review, as well as Experiments 1 and 2 in this thesis, asked participants to learn the stimuli but without informing them of the embedded patterns. Successful learning of the statistical regularities was also observed, suggesting that statistical learning can occur regardless of whether participants had the intent to learn. However, given the variability of experimental designs, it is difficult to reliably examine the impact of the intent to learn on learning outcomes. Furthermore, it is debatable whether incidental learning conditions truly reflect the absence of the intent to learn.

Second, statistical learning can result in both implicit and explicit knowledge of target patterns. In all four experiments of this thesis, participants completed a post-test awareness questionnaire designed to assess whether they could verbalise the patterns, providing a measure of explicit awareness. Results showed that some participants successfully

generalised form-to-meaning mappings (in Experiments 1 and 2) and graphotactic patterns (in Experiments 3 and 4) despite not being able to describe these patterns. These findings align with previous research (e.g., Singh et al., 2021) that statistical learning processes result in implicit knowledge that is not accessible, or at least not accessible enough to tap via the awareness questionnaire. However, the same questionnaire data also revealed that some participants were able to explicitly verbalise these patterns and those who demonstrated explicit awareness consistently outperformed those who did not. These findings challenge the view that statistical learning results in minimal explicit knowledge of the underlying statistical structure of the stimuli (Turk-Browne et al., 2005). At least within our learning experiments, there appears to be a more nuanced picture where both implicit and explicit knowledge can emerge under the same learning condition. That said, as noted in Singh et al. (2021) and throughout this thesis, using an awareness questionnaire to probe explicit awareness offers only a limited understanding of explicit learning during the statistical learning process itself. It is unclear when participants become explicitly aware of the patterns during exposure or only recognised them when prompted at the end of the experiment. Future research should explore how the implicit and explicit learning mechanisms interact over the course of learning (Karmiloff-Smith, 1991; Steffler, 2001).

Third, the ease of learning depends on the materials and the statistics to be learned (Treiman, 2018a). As highlighted in the systematic review (Chapter 2), orthotactic regularities such as bigram frequencies can be learned within minutes of exposure (e.g., Chetail, 2017). This is likely because stimuli in these experiments tend to involve a small set of highly repetitive regularities that are easier to detect and internalise. However, experimental findings from this thesis suggest that statistical learning can also be slow and incomplete (Treiman & Kessler, 2021), especially when multiple types of regularity are present in the input. For example, in

the experiments in Chapter 3, participants successfully learned form-to-meaning mappings but showed no conclusive evidence of graphotactic learning. In Chapter 4's experiments, some participants only learned the dominant suffix, but not the non-dominant suffix, in the inconsistent items. These results show that learning is not always complete and may lead to mental representations of the patterns that only partially reflect the structure of the input (Treiman & Kessler, 2021). Given the complexity of natural language, orthographic learning in natural language settings is also likely to be gradual and incomplete, taking place over many years as individuals continually build and refine their knowledge of the orthography.

The importance of explicit instruction in spelling acquisition

The previous section highlighted the limits of statistical learning processes, showing that the products from this learning process can be incomplete and slow to emerge. Learning spelling patterns through exposure, whether in lab-based settings or in the real world, is simply not adequate. Given this, explicit instruction where learners are given information concerning the rules underlying the input either during or before training may support spelling acquisition. My systematic review highlights the facilitative effects of explicit instruction for learning various orthographic patterns including form-meaning regularities (Rastle et al., 2021), graphotactic patterns (Singh et al., 2021; Singh, 2021) and phoneme-grapheme correspondences (e.g., Rastle et al., 2021).

Nevertheless, many aspects of explicit instruction remain unexplored. For example, current research in the artificial orthography learning paradigm does not determine the most effective form of explicit instruction. It can take different forms, including (1) goal-directed instructions where participants are encouraged to actively discover the underlying regularities (e.g., Verwimp et al., 2023) and (2) a brief explanation of the regularities, given before

exposure to multiple exemplars containing the patterns (e.g., Rastle et al., 2021). While both have been shown to benefit learning, they likely lead to varying levels of explicit learning and awareness of the underlying regularities during training. For example, while goal-directed instructions may encourage learners to search for regularities in the input, they do not guarantee learning of the patterns. In contrast, directly explaining the specific statistical regularities may lead to more consolidated knowledge, although this approach often requires longer learning procedures due to the additional explanations. Moreover, most experimental work examining the effect of explicit instruction focuses on phoneme-grapheme correspondences that are highly consistent. Little is known about the effect of explicit instruction or training on form-meaning regularities and graphotactic patterns, or how this effect may vary when the input contains multiple types of patterns, inconsistencies or noise, as is common in exposure to natural language.

In addition to determining the most effective form of explicit instruction in laboratory experiments, it is important to consider how this applies to classroom spelling instruction. In the UK, early spelling instruction for young children places a strong emphasis on phonics, focusing on the relationship between sounds and letters (Department for Education, 2013). However, while learning sound-letter mappings is essential, experimental findings from this thesis and prior research (e.g., Treiman, 2018b) have shown that, depending on the dosage and the number of phoneme-grapheme correspondences included, phonics may oversimplify the English writing system. This is particularly relevant as form-meaning regularities and graphotactic patterns in English orthography can be independent of phonological information. Given these limitations, explicit spelling instructions in the classroom should broaden its scope to include lessons on English morphology and graphotactics, especially for older children once they have acquired sound-spelling mappings after 1-2 years of phonics

instruction. This will allow learners to make better use of the multiple levels of consistency in the English writing system to facilitate spelling acquisition. However, while the current English curriculum in England acknowledges the importance of morphological regularities in spelling instruction (Department for Education, 2013), this does not seem to be consistently implemented in classrooms. Specifically, Esposito et al. (2023) found that only half of the sample of primary school teaching staff they surveyed reported teaching morphological regularities as a strategy to support spelling. Future research is therefore needed to determine which types of form-meaning regularities and graphotactic patterns should be included in spelling instruction to facilitate the learning of spelling patterns, and how these can be better and more systematically incorporated into classroom teaching.

Understanding L1 and L2 spelling development through learning experiments

It is worth noting that all participants in the experiments presented in this thesis were adults. As such, the learning conditions presented in this thesis resemble second language (L2) learning in many respects because participants already have extensive experience with their first language (L1) and this likely influenced how they approached the artificial orthography learning tasks. Evidence from Experiments 1 and 2 is consistent with this interpretation. Although participants on average showed above-chance learning of the form-meaning mappings, this was driven by a small number of participants who were explicitly aware of these patterns. Many participants did not learn the form-meaning mappings. One possible explanation is that participants are biased towards phonological information (as is the case in their native language English) when learning the artificial words, which could reduce the opportunity for other regularities in the writing system to be learned. Additionally, comparing graphotactic learning between adults and children, Singh et al. (2021) found that adults were more likely to become explicitly aware of the patterns (as assessed by their ability to

verbalise the patterns) and had better generalisation performance than children. This suggests that adults, who have had years of language exposure and instruction, may be better at detecting patterns in language input or more capable of verbalising patterns than children learning their L1.

Although the experimental learning conditions in this thesis resemble L2 learning, the findings remain relevant to understanding orthographic learning in L1. When considered alongside findings from previous artificial orthography learning experiments with children (e.g., Singh et al., 2021) and experiments using pseudoword choice/spelling tasks (e.g., Cassar & Treiman, 1997), the experimental findings in this thesis support a broader conclusion that statistical learning processes underlie orthographic learning, whether or not participants are familiar with the orthography. Experiments 3 and 4 provide some evidence for this idea. Regardless of whether participants were exposed to graphotactic patterns presented in a familiar orthography (Latin alphabet) or unfamiliar orthography (symbols), they developed sensitivity to the statistical regularities, although learning was stronger with Latin alphabets. While this does not directly compare L1 and L2 spelling acquisition, it nonetheless demonstrates that people can develop sensitivity to statistical regularities in an orthography even though they have no prior experience with that writing system.

Artificial orthography learning paradigm: The ways forward

Throughout this thesis, the advantages of using an artificial orthography learning paradigm to examine spelling acquisition have been highlighted. The approach allows researchers to isolate specific orthographic patterns or factors for investigation, and by introducing novel orthographic patterns, it ensures that any differences in learning outcomes between conditions are not due to differences in prior knowledge. However, the findings from the systematic

review (Chapter 2) also revealed some issues with the paradigm, particularly regarding the comparability of findings across experiments, as well as its reliability and ecological validity. While addressing these issues is beyond the scope of this thesis, efforts have been made to address these concerns in two ways.

One problem in existing artificial orthography learning experiments is that they often focus on highly consistent orthographic patterns in meaningless contexts. These have limited utility for understanding how orthographic learning occurs in natural language learning contexts in which multiple types of probabilistic patterns are present, and in meaningful contexts. To address this, Experiments 1 and 2 examined the simultaneous learning of semantic and graphotactic patterns in a meaningful context, with the artificial words embedded in English sentences. In Experiments 3 and 4, graphotactic learning was contrasted in meaningless and meaningful contexts, with the graphotactic patterns varying in frequency and consistency to mirror the probabilistic nature of regularities in written language. Across experiments, findings generally align with previous research that has used highly consistent patterns, demonstrating that participants learn regularities with short exposure (though there was no evidence for graphotactic learning in either Experiments 1 or 2). However, consistent with the findings which examined learning of orthography-phonology and orthography-semantics mappings, it was evident that statistical learning processes were slow. Furthermore, when multiple patterns were present, learning of the input's statistics was also incomplete. To more accurately understand spelling acquisition in natural language contexts, future artificial orthographic learning experiments should incorporate multiple patterns and meanings within the novel writing system.

Another issue this thesis attempted to address is how learning is measured. The systematic review (Chapter 2) highlighted the variability in the methods used to assess learning across experiments that have used the artificial orthography paradigm. Furthermore, there is no consensus on the statistical methods that should be used to determine whether learning has occurred, leading to the possibility that learning might be detected simply because certain statistical approaches are more sensitive than others. While this thesis did not attempt to standardise task designs, a key contribution of this work is demonstrating the value of Bayes factors in interpreting data when expected effect sizes are small. Silvey et al. (2024) advocate for using Bayes factors in psychological research because they help distinguish whether the data supports the alternative hypothesis, the null hypothesis, or remains inconclusive for either. This approach addresses a common issue in psycholinguistic experiments, where marginally significant (i.e., $.10 > p > .05$) and non-significant results (i.e., $p > .05$) in frequentist statistics are often over-interpreted. In our experiments, several analyses resulted in non-significant *p* values, but Bayes factors provided additional interpretation of the data. For example, in Experiment 2, evidence for graphotactic learning was inconclusive in Version 1, while there was moderate evidence for the null in Version 2. These findings reinforce the value of Bayes factors in providing a clear and more reliable interpretation of the data, particularly when results are non-significant according to frequentist methods.

In addition, Bayes factor may help inform sample sizes. The systematic review (Chapter 2) showed that sample sizes in artificial orthographic learning experiments are often quite small, with little justification apart from following previous studies. In this thesis, in addition to estimating the target sample sizes using power simulations and analyses, a stopping rule was implemented using Bayes factors. Specifically, an initial analysis was performed after collecting data from half of the sample required from the power analyses. If conclusive

evidence for the alternative hypothesis (BF > 3) or the null hypothesis (BF < 1/3) for the key effect was found, data collection stopped. Otherwise, testing continued in small increments until conclusive evidence was obtained or reached the maximum number of participants estimated in the power analyses. This approach offers a data-driven method for determining an appropriate sample size. This is especially useful when resources and funding are limited, as is often the case in postgraduate research.

To enhance the reliability and validity of the artificial orthography learning paradigm as a tool for examining reading and spelling acquisition, future research should prioritise the harmonisation of testing methods. A key consideration, as highlighted by Siegelman, Bogaerts, Christiansen, et al. (2017) in the context of statistical learning research, is the extent to which the testing tasks within this paradigm are interchangeable. Findings from the systematic review (Chapter 2) suggest that testing tasks vary in their cognitive demands and may tap into different aspects of orthographic knowledge. Making a yes/no decision in the legality judgement task is different from selecting one of six response options that best completes a phrase, as used in Experiments 3 and 4 in this thesis. The latter task, given its use of multiple foils, likely requires more specific and detailed orthographic representations to make an informed choice. Future research would benefit from comparing learning across different tasks to determine whether they yield correlated outcomes when assessing the same orthographic patterns.

Further to thinking about the tasks used to measure learning, more careful consideration should be given to their psychometric quality (Siegelman, Bogaerts, & Frost, 2017), especially when researchers want to tap into individual differences. As with many psycholinguistic experiments, artificial orthography learning experiments suffer from small

sample sizes and a limited number of test trials. Additionally, the use of multiple task versions for counterbalancing can further reduce task reliability. Zorowitz and Niv (2023) offered several recommendations that can be applied to the artificial orthography learning paradigm. First, to increase between-participant variance, post-tests should be assessed for ceiling and floor effects to ensure that they are neither too easy or too difficult. This is especially relevant to artificial orthography learning, where exposure duration is typically quite short. Chance-level performance may reflect either that participants failed to learn the orthographic patterns in the time provided or that the task is too difficult to detect learning. These observations could lead to very different conclusions about the learning processes. Second, recruiting participants from a diverse sample from platforms such as Prolific can improve the generalisability of the findings, in contrast to recruiting participants from universities, which tend to be more homogenous. However, appropriate attention checks should be implemented in online experiments to ensure participants are properly engaged in the task (see also Rodd, 2024, for further recommendations for online testing). Finally, to reduce measurement noise, researchers may consider increasing the number of trials or reducing the number of counterbalancing versions to increase the number of observations for each trial. However, using these methods to reduce measurement noise comes with some trade-offs in learning experiments. For example, more testing trials can potentially prompt learning and lead participants to adapt their responses during testing, which can confound the effects observed in a learning study. In addition, having fewer versions of the experiments may not fully account for the item-specific effects on learning and compromise the experimental validity. Ultimately, when designing an artificial orthography learning study, researchers must weigh the trade-offs between methodological rigor in a learning study and the psychometric quality of the testing phase in addressing their research questions.

Concluding Remarks

Written language is complex and rich with multiple types of regularity conditioned to varying extents on phonological, graphotactic and semantic cues. This thesis systematically reviewed evidence from the artificial orthography learning paradigm in the context of reading and spelling acquisition, and subsequently adopted this paradigm to investigate whether native English-speaking adults can learn form-meaning regularities and graphotactic patterns via exposure. Findings from the systematic review and four experiments provide new insights into the learning mechanisms underlying orthographic learning. Theoretically, the results demonstrate that individuals can develop sensitivity to statistical regularities in form-meaning mappings and graphotactic patterns through exposure. These findings support the view that statistical learning processes contribute to spelling acquisition. However, they also reveal the limits of statistical learning, showing that such learning can be slow and incomplete. From a methodological perspective, the artificial orthography learning paradigm offers several advantages for studying reading and spelling acquisition, especially in examining the underlying learning mechanism. However, challenges remain regarding the comparability of findings across experiments, as well as their validity and reliability. Finally, in terms of educational implications, this research highlights the importance of explicit awareness in generalising form-meaning mappings and graphotactic patterns, especially when learners only have brief exposure to these orthographic regularities. As such, explicit instruction may facilitate the learning of these patterns and warrants further research.

References

*Acha, J., Rodriguez, N., & Perea, M. (2023). The role of letter knowledge acquisition ability on children's decoding and word identification: Evidence from an artificial orthography. *Journal of Research in Reading*, *46*(4), 358–375. https://doi.org/10.1111/1467-9817.12432

*Adwan-Mansour, J., & Bitan, T. (2017). The Effect of Stimulus Variability on Learning and Generalization of Reading in a Novel Script. *Journal of Speech, Language and Hearing Research (Online)*, *60*(10), 2840–2851. https://doi.org/10.1044/2017_JSLHR-L-16-0293

Andrews, S., & Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *Journal of Experimental Psychology: General*, *139*(2), 299–318. https://doi.org/10.1037/a0018366

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

Apel, K., Wolter, J. A., & Masterson, J. J. (2006). Effects of phonotactic and orthotactic probabilities during fast mapping on 5-year-olds' learning to spell. *Developmental Neuropsychology*, *29*(1), 21–42. Scopus. https://doi.org/10.1207/s15326942dn2901_3

*Aravena, S., Snellings, P., Tijms, J., & van der Molen, M. W. (2013). A lab-controlled simulation of a letter-speech sound binding deficit in dyslexia. *Journal of Experimental Child Psychology*, *115*(4), 691–707. https://doi.org/10.1016/j.jecp.2013.03.009

*Aravena, S., Tijms, J., Snellings, P., & van der Molen, M. W. (2016). Predicting responsiveness to intervention in dyslexia using dynamic assessment. *Learning and Individual Differences*, *49*, 209–215. https://doi.org/10.1016/j.lindif.2016.06.024

*Aravena, S., Tijms, J., Snellings, P., & van der Molen, M. W. (2018). Predicting Individual Differences in Reading and Spelling Skill with Artificial Script-Based Letter-Speech Sound Training. *Journal of Learning Disabilities*, *51*(6), 552–564. https://doi.org/10.1177/0022219417715407

Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics*, *24*(4), 621–635. https://doi.org/10.1017/S0142716403000316

Baguley, T., & Kaye, D. (2010). Understanding psychology as a science: An introduction to scientific and statistical inference. *British Journal of Mathematical & Statistical Psychology*, *63*(3), 695–698. https://doi.org/10.1348/000711009X481027

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

*Bartolotti, J., Daniel, N. L., Marian, V., Knauff M., Sebanz N., Pauen M., & Wachsmuth I. (2013). *Spoken Words Activate Cross-Linguistic Orthographic Competitors in the Absence of Phonological Overlap* (rayyan-1110142385). 1827–1832. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85017470790&partnerID=40&md5=0ab4dfa79e3ea67dcaf48985ea3d9512

*Bartolotti, J., & Marian, V. (2019). Learning and processing of orthography-to-phonology mappings in a third language. *International Journal of Multilingualism*, *16*(4), 377–397. https://doi.org/10.1080/14790718.2017.1423073

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

Berg, K., & Aronoff, M. (2017). Self-organization in the spelling of English suffixes: The emergence of culture out of anarchy. *Language*, *93*(1), 37–64. https://doi.org/10.1353/lan.2017.0000

Berg, K., Buchmann, F., Dybiec, K., & Fuhrhop, N. (2014). Morphological spellings in English. *Written Language & Literacy*, *17*(2), 282–307. https://doi.org/10.1075/wll.17.2.05ber

*Bhide, A. (2018). Copying Helps Novice Learners Build Orthographic Knowledge: Methods for Teaching Devanagari Akshara. *Reading and Writing: An Interdisciplinary Journal*, *31*(1), 1–33. https://doi.org/10.1007/s11145-017-9767-8

*Bitan, T., & Booth, J. R. (2012). Offline Improvement in Learning to Read a Novel Orthography Depends on Direct Letter Instruction. *Cognitive Science*, *36*(5), 896–918. https://doi.org/10.1111/j.1551-6709.2012.01234.x

*Bitan, T., & Karni, A. (2003). Alphabetical knowledge from whole words training: Effects of explicit instruction and implicit experience on learning script segmentation. *Cognitive Brain Research*, *16*(3), 323–337. https://doi.org/10.1016/S0926-6410(02)00301-4

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

*Bitan, T., Manor, D., Morocz, I. A., & Karni, A. (2005). Effects of alphabeticality, practice and type of instruction on reading an artificial script: An fMRI study. *Cognitive Brain Research*, *25*(1), 90–106. https://doi.org/10.1016/j.cogbrainres.2005.04.014

*Bolger, D. J., & Perfetti, C. A. (2007). *The development of orthographic knowledge: A cognitive neuroscience investigation of reading skill* (rayyan-1110139774) [University of Pittsburgh]. https://www.proquest.com/dissertations-theses/development-orthographic-knowledge-cognitive/docview/304835850/se-2?accountid=13042

*Brennan, C., & Booth, J. R. (2015). Large Grain Instruction and Phonological Awareness Skill Influence Rime Sensitivity, Processing Speed, and Early Decoding Skill in Adult L2 Learners. *Reading and Writing: An Interdisciplinary Journal*, *28*(7), 917–938. https://doi.org/10.1007/s11145-015-9555-2

*Brennan, C., & Kiskin, J. (2020). Distinct Benefits Given Large Versus Small Grain Orthographic Instruction for English-Speaking Adults Learning to Read Russian Cyrillic. *Journal of Psycholinguistic Research*, *49*(6), 915–933. https://doi.org/10.1007/s10936-019-09684-5

*Byrne, B. (1992). Studies in the acquisition procedure for reading: Rationale, hypotheses, and data. *Reading Acquisition.*, 1–34.

*Byrne, B., & Carroll, M. (1989). Learning artificial orthographies: Further evidence of a nonanalytic acquisition procedure. *Memory & Cognition*, *17*(3), 311–317.

Cassar, M., & Treiman, R. (1997). *The Beginnings of Orthographic Knowledge: Children's Knowledge of Double Letters in Words*. 14.

*Catronas, D. M. D., Reis, A., & Araújo, S. (2020). *The Influence of Sensorimotor Training in Learning a Novel Script: A Comparison Between Handwriting and Visual Learning* (rayyan-1110139558) [Universidade do Algarve (Portugal)]. https://www.proquest.com/dissertations-theses/influence-sensorimotor-training-learning-novel/docview/2675224268/se-2?accountid=13042

Chan, J., Woore, R., Molway, L., & Mutton, T. (2022). Learning and teaching Chinese as a foreign language: A scoping review. *Review of Education*, *10*(3), e3370. https://doi.org/10.1002/rev3.3370

Chen, M. J., & Weekes, B. S. (2004). Effects of Semantic Radicals on Chinese Character Categorization and Character Decision. *Chinese Journal of Psychology*, *46*(2–3), 181–196.

***Chetail, F. (2017). What do we do with what we learn? Statistical learning of orthographic regularities impacts written word processing. *Cognition*, *163*, 103–120. Scopus. https://doi.org/10.1016/j.cognition.2017.02.015**

*Chetail, F., & Sauval, K. (2022). Diversity matters: The sensitivity to sublexical orthographic regularities increases with contextual diversity. *Psychonomic Bulletin & Review*, *29*(3), 1003–1016. https://doi.org/10.3758/s13423-021-02029-1

Chong, S. W., & Reinders, H. (2025). Autonomy of English language learners: A scoping review of research and practice. *Language Teaching Research*, *29*(2). https://doi.org/10.1177/13621688221075812

Cochrane Effective Practice and Organisation of Care (EPOC). (2017). *EPOC Resources for review authors*. epoc.cochrane.org/epoc-resources-review-authors

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

Deacon, S. H., Conrad, N., & Pacton, S. (2008). A statistical learning perspective on children's learning about graphotactic and morphological regularities in spelling. *Canadian Psychology/Psychologie Canadienne*, *49*(2), 118–124. https://doi.org/10.1037/0708-5591.49.2.118

Department for Education. (2013). *English programmes of study: Key stages 1 and 2*.

Dewey, G. (1971). *English Spelling: Roadblock to Reading*. Teachers College Press. https://books.google.co.uk/books?id=rewfAAAAMAAJ

Dienes, Z. (n.d.). *Making the most of your data with Bayes*. Making the Most of Your Data with Bayes.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan.

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

*Dong, J., Yue, Q., Li, A., Gu, L., Su, X., Chen, Q., & Mei, L. (2022). Individuals' preference on reading pathways influences the involvement of neural pathways in phonological learning. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1067561

Ellis, N. C., Natsume, M., Stavropoulou, K., Hoxhallari, L., Van Daal, V. H. p., Polyzoe, N., Tsipa, M.-L., & Petalas, M. (2004). The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading Research Quarterly*, *39*(4), 438–468. https://doi.org/10.1598/RRQ.39.4.5

Esposito, R., Herbert, E., & Sumner, E. (2023). Capturing variations in how spelling is taught in primary school classrooms in England. *British Educational Research Journal*, *49*(1), 70–92. https://doi.org/10.1002/berj.3829

Fayol, M., Totereau, C., & Barrouillet, P. (2006). Disentangling the impact of semantic and formal factorsin the acquisition of number inflections: Noun, adjective and verb agreement in written French. *Reading and Writing*, *19*(7), 717–736. https://doi.org/10.1007/s11145-005-1371-7

*Fernández-López, M., Marcet, A., & Perea, M. (2021). Does orthographic processing emerge rapidly after learning a new script? *British Journal of Psychology*, *112*(1), 52–91. https://doi.org/10.1111/bjop.12469

*Fernández-López, M., & Perea, M. (2023). A Letter is a Letter and its Co-Occurrences: Cracking the Emergence of Position-Invariance Processing. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-023-02265-7

Ferreiro, E., & Teberosky, A. (1982). *Literacy before Schooling.* Heinemann Educational Books Inc., 70 Court St., Portsmouth, NH 03801. https://eric.ed.gov/?id=ed263542

Flöel, A., Rösser, N., Michka, O., Knecht, S., & Breitenstein, C. (2008). Noninvasive Brain Stimulation Improves Language Learning. *Journal of Cognitive Neuroscience*, *20*(8), 1415–1422. https://doi.org/10.1162/jocn.2008.20098

*Gelzheiser, L. M. (1991). Learning Sound/Symbol Correspondences: Transfer Effects of Pattern Detection and Phonics Instruction. *Applied Cognitive Psychology*, *5*(4), 361–371.

*Guerra, G., Tijms, J., Tierney, A., Vaessen, A., Dick, F., & Bonte, M. (2024). Auditory attention influences trajectories of symbol–speech sound learning in children with and without dyslexia. *Journal of Experimental Child Psychology*, *237*. https://doi.org/10.1016/j.jecp.2023.105761

Harm, M. W., & Seidenberg, M. S. (2004). Computing the Meanings of Words in Reading: Cooperative Division of Labor Between Visual and Phonological Processes. *Psychological Review*, *111*(3), 662–720. https://doi.org/10.1037/0033-295X.111.3.662

*Hart, L. A., & Perfetti, C. A. (2005). *A training study using an artificial orthography: Effects of reading experience, lexical quality, and text comprehension in L1 and L2* (rayyan-1110139802) [University of Pittsburgh].

https://www.proquest.com/dissertations-theses/training-study-using-artificial-orthography/docview/305414278/se-2?accountid=13042

Hayes, H., Treiman, R., & Kessler, B. (2006). Children use vowels to help them spell consonants. *Journal of Experimental Child Psychology*, *94*(1), 27–42. https://doi.org/10.1016/j.jecp.2005.11.001

***He, X., & Tong, X. (2017). Statistical learning as a key to cracking Chinese orthographic codes. *Scientific Studies of Reading*, *21*(1), 60–75. APA PsycInfo <2017>. https://doi.org/10.1080/10888438.2016.1243541**

Heyer, V. (2021). Below the surface: The application of implicit morpho-graphic regularities to novel word spelling. *Morphology*, *31*(3), 243–260. https://doi.org/10.1007/s11525-020-09370-6

Hirshorn, E. A., & Fiez, J. A. (2014). Using artificial orthographies for studying cross-linguistic differences in the cognitive and neural profiles of reading. *Journal of Neurolinguistics*, *31*, 69–85. Scopus. https://doi.org/10.1016/j.jneuroling.2014.06.006

*Hirshorn, E. A., Wrencher, A., Durisko, C., Moore, M. W., & Fiez, J. A. (2016). Fusiform gyrus laterality in writing systems with different mapping principles: An artificial orthography training study. *Journal of Cognitive Neuroscience*, *28*(6), 882–894. https://doi.org/10.1162/jocn_a_00940

Huddleston, R., & Pullum, G. K. (2002). The Cambridge Grammar of the English language. (The Grammar for the 21st Century). *The Bookseller (London)*, 43.

Hulstijn, J. H. (2005). THEORETICAL AND EMPIRICAL ISSUES IN THE STUDY OF IMPLICIT AND EXPLICIT SECOND-LANGUAGE LEARNING: Introduction. *Studies in Second Language Acquisition, 27*(2), 129–140. https://doi.org/10.1017/S0272263105050084

Hulstijn, J. H., Long, M. H., & Doughty, C. J. (2003). Incidental and Intentional Learning. In *The Handbook of Second Language Acquisition* (pp. 349–381). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470756492.ch12

*Ise, E., Arnoldi, C. J., Bartling, J., & Schulte-Körne, G. (2012). Implicit learning in children with spelling disability: Evidence from artificial grammar learning. *Journal of Neural Transmission*, *119*(9), 999–1010. https://doi.org/10.1007/s00702-012-0830-y

*Jenkins, J. R., Bausell, R. B., & Jenkins, L. M. (1972). Comparisons of letter name and letter sound training as transfer variables. *American Educational Research Journal*, *9*(1), 75–86. https://doi.org/10.2307/1162051

Karmiloff-Smith, A. (1991). Beyond Modularity: Innate Constraints and Developmental Change. In *The Epigenesis of Mind*. Psychology Press.

Katz, L., & Frost, R. (1992). Chapter 4 The Reading Process is Different for Different Orthographies: The Orthographic Depth Hypothesis. In *Advances in Psychology* (Vol. 94, pp. 67–84). Elsevier. https://doi.org/10.1016/S0166-4115(08)62789-2

Klasen, L., Ugen, S., Dording, C., Fayol, M., & Weth, C. (2023). Do learners need semantics to spell syntactic markers? Plural spellings in real vs. pseudowords in a French L2 setting. *Reading and Writing*. https://doi.org/10.1007/s11145-023-10422-6

*Laine, M., Polonyi, T., & Abari, K. (2014). More Than Words: Fast Acquisition and Generalization of Orthographic Regularities During Novel Word Learning in Adults. *Journal of Psycholinguistic Research*, *43*(4), 381–396.

*Lally, C., Taylor, J. S. H., Lee, C. H., & Rastle, K. (2020). Shaping the precision of letter position coding by varying properties of a writing system. *Language, Cognition and Neuroscience*, *35*(3), 374–382. https://doi.org/10.1080/23273798.2019.1663222

*Law, J. M., De Vos, A., Vanderauwera, J., Wouters, J., Ghesquière, P., & Vandermosten, M. (2018). Grapheme-phoneme learning in an unknown orthography: A study in

typical reading and dyslexic children. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01393

Law, N. S. H., Samara, A., Wonnacott, E., & Nation, K. (2025). *Simultaneous learning of semantic and graphotactic regularities in spelling: An artificial orthography learning experiment*. https://osf.io/7arwb/download

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lehtonen, A., & Bryant, P. (2005). Doublet challenge: Form comes before function in children's understanding of their orthography. *Developmental Science*, *8*(3), 211–217. https://doi.org/10.1111/j.1467-7687.2005.00409.x

***Lelonkiewicz, J. R., Ktori, M., & Crepaldi, D. (2020). Morphemes as letter chunks: Discovering affixes through visual regularities. *Journal of Memory and Language*, *115*. https://doi.org/10.1016/j.jml.2020.104152**

***Lelonkiewicz, J. R., Ktori, M., & Crepaldi, D. (2023). Morphemes as letter chunks: Linguistic information enhances the learning of visual regularities. *Journal of Memory and Language*, *130*. Scopus. https://doi.org/10.1016/j.jml.2023.104411**

Li, Y., & Wang, M. (2023). A systematic review of orthographic learning via self-teaching. *Educational Psychologist*, *58*(1), 35–56. https://doi.org/10.1080/00461520.2022.2137673

*Marian, V., Bartolotti, J., Daniel, N. L., & Hayakawa, S. (2021). Spoken words activate native and non-native letter-to-sound mappings: Evidence from eye tracking. *Brain and Language*, *223*, 1. https://doi.org/10.1016/j.bandl.2021.105045

*Martin, L., Durisko, C., Moore, M. W., Coutanche, M. N., Chen, D., & Fiez, J. A. (2019). The VWFA is the home of orthographic learning when houses are used as letters. *eNeuro*, *6*(1). https://doi.org/10.1523/ENEURO.0425-17.2019

*Martin, L., Hirshorn, E. A., Durisko, C., Moore, M. W., Schwartz, R., Zheng, Y., & Fiez, J. A. (2019). Do adults acquire a second orthography using their native reading network? *Journal of Neurolinguistics*, *50*, 120. https://doi.org/10.1016/j.jneuroling.2018.03.004

*Maurer, U., Blau, V. C., Yoncheva, Y. N., & McCandliss, B. D. (2010). Development of Visual Expertise for Reading: Rapid Emergence of Visual Familiarity for an Artificial Script. *Developmental Neuropsychology*, *35*(4), 404–422.

*McMillan, I. M., El-Deredy, W., & Woollams, A. (2017). *Stimulating Reading: A Behavioural and Electrophysiological Investigation of the Impact of Brain Stimulation in Developmental Dyslexia* (rayyan-1110139625) [The University of Manchester (United Kingdom)]. https://www.proquest.com/dissertations-theses/stimulating-reading-behavioural/docview/2186362111/se-2?accountid=13042

*Mei, L., Xue, G., Lu, Z.-L., He, Q., Wei, M., Zhang, M., Dong, Q., & Chen, C. (2015). Native language experience shapes neural basis of addressed and assembled phonologies. *NeuroImage*, *114*, 38–48. https://doi.org/10.1016/j.neuroimage.2015.03.075

*Mei, L., Xue, G., Lu, Z.-L., He, Q., Zhang, M., Xue, F., Chen, C., & Dong, Q. (2013). Orthographic Transparency Modulates the Functional Asymmetry in the Fusiform Cortex: An Artificial Language Training Study. *Brain and Language*, *125*(2), 165–172. https://doi.org/10.1016/j.bandl.2012.01.006

*Mei, L., Xue, G., Zhong-Lin, L., He, Q., Zhang, M., Miao, W., Xue, F., Chen, C., & Dong, Q. (2014). Artificial Language Training Reveals the Neural Substrates Underlying Addressed and Assembled Phonologies. *PLoS One*, *9*(3). https://doi.org/10.1371/journal.pone.0093548

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

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

Pacton, S., Borchardt, G., Treiman, R., Lété, B., & Fayol, M. (2014). Learning to spell from reading: General knowledge about spelling patterns influences memory for specific words. *Quarterly Journal of Experimental Psychology*, *67*(5), 1019–1036. Scopus. https://doi.org/10.1080/17470218.2013.846392

Pacton, S., Fayol, M., & Perruchet, P. (2005). Children's Implicit Learning of Graphotactic and Morphological Regularities. *Child Development*, *76*(2), 324–339. https://doi.org/10.1111/j.1467-8624.2005.00848_a.x

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit learning out of the lab: The case of orthographic regularities. *Journal of Experimental Psychology: General*, *130*(3), 401–426. https://doi.org/10.1037/0096-3445.130.3.401

Pacton, S., Sobaco, A., Fayol, M., & Treiman, R. (2013). How does graphotactic knowledge influence children's learning of new spellings? *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00701

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, *88*, 105906. https://doi.org/10.1016/j.ijsu.2021.105906

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). *Understanding Normal and Impaired Word Reading: Computational Principles in Quasi-Regular Domains*.

*Quinn, C., Taylor, J. S. H., & Davis, M. H. (2017). Learning and retrieving holistic and componential visual-verbal associations in reading and object naming. *Neuropsychologia*, *98*, 68–84. https://doi.org/10.1016/j.neuropsychologia.2016.09.025

***Rastle, K., Lally, C., Davis, M. H., & Taylor J S H. (2021). The Dramatic Impact of Explicit Instruction on Learning to Read in a New Writing System. *Psychological Science*, *32*(4), 471–484. https://doi.org/10.1177/0956797620968790**

Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language*, *134*, 104472. https://doi.org/10.1016/j.jml.2023.104472

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

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52. https://doi.org/10.1016/S0010-0277(98)00075-4

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

***Samara, A., & Caravolas, M. (2014). Statistical learning of novel graphotactic constraints in children and adults. *Journal of Experimental Child Psychology*, *121*, 137–155. https://doi.org/10.1016/j.jecp.2013.11.009**

***Samara, A., Singh, D., & Wonnacott, E. (2019). Statistical learning and spelling: Evidence from an incidental learning experiment with children. *Cognition*, *182*, 25–30. https://doi.org/10.1016/j.cognition.2018.09.005**

Schapiro, A., & Turk-Browne, N. (2015). Statistical Learning. In A. W. Toga (Ed.), *Brain Mapping* (pp. 501–506). Academic Press. https://doi.org/10.1016/B978-0-12-397025-1.00276-1

*Schmalz, X., Mulatti, C., Schulte-Körne, G., & Moll, K. (2022). Effects of complexity and unpredictability on the learning of an artificial orthography. *Cortex*, *152*, 1–20. https://doi.org/10.1016/j.cortex.2022.03.014

Schmalz, X., Robidoux, S., Castles, A., & Marinus, E. (2020). Variations in the use of simple and context-sensitive grapheme-phoneme correspondences in English and German developing readers. *Annals of Dyslexia*, *70*(2), 180–199. https://doi.org/10.1007/s11881-019-00189-3

*Schmalz, X., Schulte-Korne, G., De Simone, E., & Moll, K. (2021). What do artificial orthography learning tasks actually measure? Correlations within and across tasks. *Journal of Cognition*, *4*(1). https://doi.org/10.5334/joc.144

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schulz, J., Hamilton, C., Wonnacott, E., & Murphy, V. (2023). The impact of multi-word units in early foreign language learning and teaching contexts: A systematic review. *Review of Education*, *11*(2), e3413. https://doi.org/10.1002/rev3.3413

Seidenberg, M. S. (2005). Connectionist Models of Word Reading. *Current Directions in Psychological Science*, *14*(5), 238–242. https://doi.org/10.1111/j.0963-7214.2005.00372.x

Seidenberg, M. S., & McClelland, J. L. (1989). *A Distributed, Developmental Model of Word Recognition and Naming*.

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an 'outlier' orthography. *Psychological Bulletin*, *134*(4), 584–615. https://doi.org/10.1037/0033-2909.134.4.584

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160059. https://doi.org/10.1098/rstb.2016.0059

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, *49*(2), 418–432. https://doi.org/10.3758/s13428-016-0719-z

Silvey, C., Dienes, Z., & Wonnacott, E. (2024). Bayes factors for logistic (mixed-effect) models. *Psychological Methods*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/met0000714

***Singh, D., Wonnacott, E., & Samara, A. (2021). Statistical and explicit learning of graphotactic patterns with no phonological counterpart: Evidence from an artificial lexicon study with 6–7-year-olds and adults. *Journal of Memory and Language*, *121*. https://doi.org/10.1016/j.jml.2021.104265**

*Singh, F. D. (2021). Statistical and Explicit Learning of Graphotactic Patterns with No Phonological Counterpart: Evidence from Artificial Lexicon Studies with 6- to 7-Year-Olds and Adults [Ph.D., University of London, University College London (United Kingdom)]. In *PQDT - UK & Ireland* (2607337844). ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/statistical-explicit-learning-graphotactic/docview/2607337844/se-2?accountid=13042

Sobaco, A., Treiman, R., Peereman, R., Borchardt, G., & Pacton, S. (2015). The influence of graphotactic knowledge on adults' learning of spelling. *Memory & Cognition*, *43*(4), 593–604. APA PsycInfo <2015>. https://doi.org/10.3758/s13421-014-0494-y

Steffler, D. J. (2001). Implicit Cognition and Spelling Development. *Developmental Review*, *21*(2), 168–204. https://doi.org/10.1006/drev.2000.0517

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., Lu, C. C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., … Bates, E. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, *51*(2), 247–250. https://doi.org/10.1016/j.jml.2004.03.002

*Tamminen, J., Newbury, C. R., Crowley, R., Vinals, L., Cevoli, B., & Rastle, K. (2020). Generalisation in language learning can withstand total sleep deprivation. *Neurobiology of Learning and Memory*, *173*. https://doi.org/10.1016/j.nlm.2020.107274

*Taylor, J. S. H. (2010). *The impact of frequency, consistency, and semantics on reading aloud: An artificial orthography learning paradigm* (rayyan-1110139739) [University of Oxford (United Kingdom)]. https://www.proquest.com/dissertations-theses/impact-frequency-consistency-semantics-on-reading/docview/2340614288/se-2?accountid=13042

*Taylor, J. S. H., Davis, M. H., & Rastle, K. (2017). Comparing and validating methods of reading instruction using behavioural and neural findings in an artificial orthography. *Journal of Experimental Psychology: General*, *146*(6), 826–858. APA PsycInfo <2017>. https://doi.org/10.1037/xge0000301

Taylor, J. S. H., Davis, M. H., & Rastle, K. (2019). Mapping visual symbols onto spoken language along the ventral visual stream. *Proceedings of the National Academy of*

Sciences of the United States of America, 116(36), 17723–17728. https://doi.org/10.1073/pnas.1818575116

*Taylor, J. S. H., Plunkett, K., & Nation, K. (2011). The influence of consistency, frequency, and semantics on learning to read: An artificial orthography paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 60–76. https://doi.org/10.1037/a0020126

*Taylor, J. S., Rastle, K., & Davis, M. H. (2014). Distinct Neural Specializations for Learning to Read Words and Name Objects. *Journal of Cognitive Neuroscience*, *26*(9), 2128–2154.

*Thakkar, V. J., Engelhart, A. S., Khodaparast, N., Abadzi, H., & Centanni, T. M. (2020). Transcutaneous auricular vagus nerve stimulation enhances learning of novel letter-sound relationships in adults. *Brain Stimulation*, *13*(6), 1813–1820. https://doi.org/10.1016/j.brs.2020.10.012

*Tong, S. X., Duan, R., Shen, W., Yu, Y., & Tong, X. (2023). Multiple mechanisms regulate statistical learning of orthographic regularities in school-age children: Neurophysiological evidence. *Developmental Cognitive Neuroscience*, *59*. https://doi.org/10.1016/j.dcn.2022.101190

*Tong, S., Zhang, P., & He, X. (2020). Statistical Learning of Orthographic Regularities in Chinese Children with and without Dyslexia. *Child Development*, *91*(6), 1953–1969. https://doi.org/10.1111/cdev.13384

Treiman, R. (2018a). Statistical Learning and Spelling. *Language, Speech, and Hearing Services in Schools*, *49*(3S), 644–652. https://doi.org/10.1044/2018_LSHSS-STLT1-17-0122

Treiman, R. (2018b). Teaching and Learning Spelling. *Child Development Perspectives*, *12*(4), 235–239. https://doi.org/10.1111/cdep.12292

Treiman, R., & Boland, K. (2017). Graphotactics and spelling: Evidence from consonant doubling. *Journal of Memory and Language*, *92*, 254–264. https://doi.org/10.1016/j.jml.2016.07.001

Treiman, R., & Kessler, B. (2006). Spelling as statistical learning: Using consonantal context to spell vowels. *Journal of Educational Psychology*, *98*(3), 642–652. https://doi.org/10.1037/0022-0663.98.3.642

Treiman, R., & Kessler, B. (2016). Choosing between alternative spellings of sounds: The role of context. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1154–1159. https://doi.org/10.1037/xlm0000225

Treiman, R., & Kessler, B. (2021). Statistical Learning in Word Reading and Spelling across Languages and Writing Systems. *Scientific Studies of Reading*, 1–11. https://doi.org/10.1080/10888438.2021.1920951

Treiman, R., & Wolter, S. (2018). Phonological and graphotactic influences on spellers' decisions about consonant doubling. *Memory & Cognition*, *46*(4), 614–624. https://doi.org/10.3758/s13421-018-0793-9

Treiman, R., Wolter, S., & Kessler, B. (2021). How sensitive are adults to the role of morphology in spelling? *Morphology*, *31*(3), 261–271. https://doi.org/10.1007/s11525-020-09356-4

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The Automaticity of Visual Statistical Learning. *Journal of Experimental Psychology: General*, *134*(4), 552–564. https://doi.org/10.1037/0096-3445.134.4.552

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience*, *21*(10), 1934–1945. https://doi.org/10.1162/jocn.2009.21131

Ulicheva, A., Harvey, H., Aronoff, M., & Rastle, K. (2020). Skilled readers' sensitivity to meaningful regularities in English writing. *Cognition*, *195*, 103810. https://doi.org/10.1016/j.cognition.2018.09.013

Ulicheva, A., Marelli, M., & Rastle, K. (2021). Sensitivity to meaningful regularities acquired through experience. *Morphology*, *31*(3), 275–296. https://doi.org/10.1007/s11525-020-09363-5

Venezky, R. L. (1967). English Orthography: Its Graphical Structure and Its Relation to sound. *Reading Research Quarterly*, *2*(3), 75–105. https://doi.org/10.2307/747031

*Verwimp, C., Snellings, P., Wiers, R. W., & Tijms, J. (2023). Goal-directedness enhances letter-speech sound learning and consolidation in an unknown orthography. *Child Development*, *94*(4), 836–852. APA PsycInfo <2023 to October Week 1 2023>. https://doi.org/10.1111/cdev.13901

Vidal, C., Content, A., & Chetail, F. (2017). BACS: The Brussels Artificial Character Sets for studies in cognitive psychology and neuroscience. *Behavior Research Methods*, *49*(6), 2093–2112. https://doi.org/10.3758/s13428-016-0844-8

*Vidal, Y., Viviani, E., Zoccolan, D., & Crepaldi, D. (2021). A general-purpose mechanism of visual feature association in visual word identification and beyond. *Current Biology*, *31*(6), 1261-1267.e3. Scopus. https://doi.org/10.1016/j.cub.2020.12.017

*Wei, M., Manis, F. R., & Lu, Z.-L. (2015). *Cognitive-Linguistic Factors and Brain Morphology Predict Individual Differences in Form-Sound Association Learning: Two Samples From English-Speaking and Chinese-Speaking University Students* (rayyan-1110139654) [University of Southern California]. https://www.proquest.com/dissertations-theses/cognitive-linguistic-factors-brain-morphology/docview/2063353729/se-2?accountid=13042

*Williams, G. P., Panayotov, N., & Kempe, V. (2020). How does dialect exposure affect learning to read and spell? An artificial orthography study. *Journal of Experimental Psychology*, *149*(12), 2344. https://doi.org/10.1037/xge0000778

*Williams, G. P., Panayotov, N., & Kempe, V. (2022). Exposure to dialect variation in an artificial language prior to literacy training impairs reading of words with competing variants but does not affect decoding skills. *Journal of Experimental Psychology*, *48*(12), 1868. https://doi.org/10.1037/xlm0001094

*Wu, Q., Fang, X., Chen, Q., Li, Y., & Deng, Y. (2011). *The learning of morphological principles: A statistical learning study on a system of artificial scripts* (rayyan-1110142442). *122*, 187–196. https://doi.org/10.1007/978-3-642-25553-3_25

*Xue, H., Zhao, L., Wang, Y., Dong, Q., Chen, C., & Xue, G. (2017). Anodal transcranial direct current stimulation over the left temporoparietal cortex facilitates assembled phonology. *Trends in Neuroscience and Education*, *8*, 10–17. https://doi.org/10.1016/j.tine.2017.08.001

*Yoncheva, Y. N., Blau, V. C., Maurer, U., & McCandliss, B. D. (2010). Attentional Focus During Learning Impacts N170 ERP Responses to an Artificial Script. *Developmental Neuropsychology*, *35*(4), 423–445.

*Yoncheva, Y. N., Wise, J., & McCandliss, B. (2015). Hemispheric specialization for visual words is shaped by attention to sublexical units during initial learning. *Brain and Language*, *145*, 23–33. https://doi.org/10.1016/j.bandl.2015.04.001

*Younger, J. W., & Booth, J. R. (2018). Parietotemporal Stimulation Affects Acquisition of Novel Grapheme-Phoneme Mappings in Adult Readers. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2018.00109

*Zhao, J., Li, T., Elliott, M. A., & Rueckl, J. G. (2018). Statistical and Cooperative Learning in Reading: An Artificial Orthography Learning Study. *Scientific Studies of Reading*, *22*(3), 191–208. https://doi.org/10.1080/10888438.2017.1414219

*Zhao, J., & Rueckl, J. G. (2012). *The Influence of Statistical Systematicities on Learning to Read: Studies with Artificial Orthographies* (rayyan-1110139707) [University of Connecticut]. https://www.proquest.com/dissertations-theses/influence-statistical-systematicities-on-learning/docview/1223343731/se-2?accountid=13042

Zorowitz, S., & Niv, Y. (2023). Improving the reliability of cognitive task measures: A narrative review. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *8*(8), 789–797. https://doi.org/10.1016/j.bpsc.2023.02.004

Appendices

Appendix for Chapter 2 (Systematic Review)

Appendix 2A. Descriptive Summary of Experimental Characteristics for Each Experiment Included in the Review

Experiment ID	Authors (Year)	Publication type	Design	Age group	Additional profiles	Sample size	Native language background	Artificial orthographic patterns		
								O-P	O-S	O
E001	Acha, Rodriguez, & Perea (2023)	JA	Behavioural	C		30	Spanish-Basque bilinguals	X		
E002	Adwan-Mansour & Bitan (2017)	JA	Behavioural	A		55	Hebrew	X		
E003	Aravena, Snellings, Tijms, & van der Molen (2013)	JA	Behavioural	C	With/without dyslexia	126	Dutch	X		
E004	Aravena, Tijms, Snellings, & van der Molen (2016)	JA	Behavioural	C	With dyslexia	55	Dutch	X		
E005	Aravena, Tijms, Snellings, & van der Molen (2018)	JA	Behavioural	C	With/without dyslexia	118	Dutch	X		
E006	Bartolotti & Marian (2019)	JA	Behavioural + eye-tracking	A		20	Spanish-English bilinguals	X		
E007	Bartolotti, Daniel & Marian (2013)	CP	Behavioural + eye-tracking	A		20	English	X		

E008	Bhide (2018)	JA	Behavioural	A		41	English	X		
E009	Bitan & Booth (2012)	JA	Behavioural	A		48	English	X		
E010	Bitan & Karni (2003)	JA	Behavioural	A		9	n/a	X		
E011	Bitan & Karni (2004)	JA	Behavioural	A		24	n/a	X		
E012	Bitan, Manor, Morocz, & Karni (2005)	JA	Behavioural + fMRI	A		16	Hebrew	X		
E013	Bolger & Perfetti (2007)	T (PhD)	Behavioural + fMRI	A		30	English	X		
E014	Brennan & Booth (2015)	JA	Behavioural	A		37	English	X		
E015	Brennan & Kiskin (2020)	JA	Behavioural	A		34	English	X		
E016	Byrne (1992) (Experiment 6)	BC	Behavioural	C		24	English*	X		
E017	Byrne (1992) (Experiment 8)	BC	Behavioural	C		13	English*	X		
E018	Byrne (1992) (Experiment 9)	BC	Behavioural	C		12	English*	X		
E019	Byrne (1992) (Experiment 10)	BC	Behavioural	C		14	English*	X		
E020	Byrne (1992) (Experiment 11)	BC	Behavioural	C		13	English*	X		
E021	Byrne & Carroll (1989) (Experiment 1)	JA	Behavioural	A		24	English	X		

E022	Byrne & Carroll (1989) (Experiment 2)	JA	Behavioural	A		24	English	X		
E023	Byrne & Carroll (1989) (Experiment 3)	JA	Behavioural	A		16	English	X		
E024	Catronas, Reis, & Araújo (2020)	T (Mas)	Behavioural + eye-tracking	A		36	Portuguese	X		
E025	Chetail (2017) (Experiment 1a)	JA	Behavioural	A		35	French			X
E026	Chetail (2017) (Experiment 1b)	JA	Behavioural	A		35	French			X
E027	Chetail (2017) (Experiment 2)	JA	Behavioural	A		25	French	X		X
E028	Chetail & Sauval (2022)	JA	Behavioural	A		24	French			X
E029	Dong, Yue, Li, Gu, Su, Chen, & Mei (2022)	JA	Behavioural + fMRI	A		25	Chinese	X		
E030	Fernández-López & Perea (2023)	JA	Behavioural	A		36	Spanish			X
E031	Fernández-López, Marcet, & Perea (2021) (Experiment 1)	JA	Behavioural	A		28	Spanish	X		
E032	Fernández-López, Marcet, & Perea (2021) (Experiment 2)	JA	Behavioural	A		28	Spanish	X		
E033	Gelzheiser (1991)	JA	Behavioural	C		36	English	X		

E034	Guerra, Tijms, Tierney, Vaessen, Dick, & Bonte (2024)	JA	Behavioural (EEG data on the nonspeech sustained auditory selective attention task)	C	With/without dyslexia	110	Dutch	X		
E035	Hart & Perfetti (2005) (Thesis chapter 2-6)	T (PhD)	Behavioural (ERP data not related to artificial orthographic patterns)	A	With good/poor lexical and/or reading comprehension skills	45	English	X		
E036	He & Tong (2017) (Experiment 1)	JA	Behavioural	C		22	Cantonese			X
E037	He & Tong (2017) (Experiment 2)	JA	Behavioural	C		22	Cantonese	X		
E038	He & Tong (2017) (Experiment 3)	JA	Behavioural	C		22	Cantonese		X	
E039	Hirshorn, Wrencher, Durisko, Moore, & Fiez (2016)	JA	Behavioural + fMRI	A		27	English	X		
E040	Ise, Arnoldi, Bartling, & Schulte-Körne (2012)	JA	Behavioural	C	With/without spelling disabilities	61	German*			X
E041	Jenkins, Bausell, & Jenkins (1972) (Experiment 1)	JA	Behavioural	C		96	English*	X		
E042	Jenkins, Bausell, & Jenkins (1972) (Experiment 2)	JA	Behavioural	C		60	English*	X		

E043	Jenkins, Bausell, & Jenkins (1972) (Experiment 3)	JA	Behavioural	C		36	English*	X		
E044	Laine, Polonyi, & Abari (2014)	JA	Behavioural	A		55	Hungarian			X
E045	Lally, Taylor, Lee, & Rastle (2020)	JA	Behavioural	A		48	English	X		
E046	Law, De Vos, Vanderauwera, Wouters, Ghesquière, & Vandermosten (2018)	JA	Behavioural	C	With/without dyslexia	84	Dutch	X		
E047	Lelonkiewicz, Ktori, & Crepaldi (2020) (Experiment 1)	JA	Behavioural	A		70	Italian			X
E048	Lelonkiewicz, Ktori, & Crepaldi (2020) (Experiment 2)	JA	Behavioural	A		71	Italian			X
E049	Lelonkiewicz, Ktori, & Crepaldi (2023) (Experiment 1a)	JA	Behavioural	A		41	Italian			X
E050	Lelonkiewicz, Ktori, & Crepaldi (2023) (Experiment 1b)	JA	Behavioural	A		45	Italian			X
E051	Lelonkiewicz, Ktori, & Crepaldi (2023) (Experiment 2a)	JA	Behavioural	A		40	Italian			X
E052	Lelonkiewicz, Ktori, & Crepaldi (2023) (Experiment 2b)	JA	Behavioural	A		40	Italian			X
E053	Lelonkiewicz, Ktori, & Crepaldi (2023) (Experiment 3)	JA	Behavioural	A		48	Italian		X	X

E054	Marian, Bartolotti, Daniel, & Hayakawa (2021) (Experiment 1)	JA	Behavioural	A		20	English	X		
E055	Marian, Bartolotti, Daniel, & Hayakawa (2021) (Experiment 2)	JA	Behavioural + Eye-tracking	A		20	English	X		
E056	Martin, Durisko, Moore, Coutanche, Chen, & Fiez (2019)	JA	Behavioural + fMRI	A		12	English	X		
E057	Martin, Hirshorn, Durisko, Moore, Schwartz, Zheng, & Fiez (2019)	JA	Behavioural + fMRI	A		14	English	X		
E058	Maurer, Blau, Yoncheva, & McCandliss (2010)	JA	Behavioural + ERP	A		30	English	X		
E059	McMillan, El-Deredy, & Woollams (2017) (Thesis chapter 5-7)	T (PhD)	Behavioural + stimulation + EEG	A	With/without dyslexia	32	English*	X		
E060	Mei, Xue, Lu, He, Wei, Zhang, Dong, & Chen (2015)	JA	Behavioural + fMRI	A		85	Chinese (42 participants); English (43 participants)	X		
E061	Mei, Xue, Lu, He, Zhang, Miao, Xue, Chen, & Dong (2014)	JA	Behavioural + fMRI	A		43	English	X		
E062	Mei, Xue, Lu, He, Zhang, Xue, Chen, & Dong (2013)	JA	Behavioural + fMRI	A		44	Chinese	X		
E063	Quinn, Taylor, & Davis (2017)	JA	Behavioural + fMRI	A		19	English	X		
E064	Rastle, Lally, Davis, & Taylor (2021)	JA	Behavioural	A		48	English	X	X	

E065	Samara & Caravolas (2014)	JA	Behavioural	A,C		113; 137	English			X
E066	Samara, Singh, & Wonnacott (2019 (Experiment 1)	JA	Behavioural	C		78	English			X
E067	Samara, Singh, & Wonnacott (2019) (Experiment 2)	JA	Behavioural	C		37	Turkish			X
E068	Schmalz, Mulatti, Schulte-Körne, & Moll (2022) (Experiment 1)	JA	Behavioural	A		32	Italian	X		
E069	Schmalz, Mulatti, Schulte-Körne, & Moll (2022) (Experiment 2)	JA	Behavioural	A		32	Italian	X		
E070	Schmalz, Mulatti, Schulte-Körne, & Moll (2022) (Experiment 3)	JA	Behavioural	A		37	German	X		
E071	Schmalz, Schulte-Körne, De Simone, & Moll (2021)	JA	Behavioural	A		55	German	X		
E072	Singh (2021) (Experiment 1)	T (PhD)	Behavioural	A,C		18; 20	English*			X
E073	Singh (2021) (Experiment 2)	T (PhD)	Behavioural	A,C		27; 35	English*			X
E074	Singh (2021) (Experiment 6)	T (PhD)	Behavioural	A		36	English			X
E075	Singh (2021) (Experiment 7)	T (PhD)	Behavioural	A		41	English			X
E076	Singh, Wonnacott, & Samara (2021) (Experiment 1)	JA	Behavioural	A,C		29; 35	English			X

E077	Singh, Wonnacott, & Samara (2021) (Experiment 2)	JA	Behavioural	A,C		35; 25	English			X
E078	Singh, Wonnacott, & Samara (2021) (Experiment 3)	JA	Behavioural	C		25	English			X
E079	Tamminen, Newbury, Crowley, Vinals, Cevoli, & Rastle (2020) (Experiment 1)	JA	Behavioural	A		47	English	X		
E080	Tamminen, Newbury, Crowley, Vinals, Cevoli, & Rastle (2020) (Experiment 2)	JA	Behavioural	A		46	English	X		
E081	Taylor (2010) (Experiment 2)	T (PhD)	Behavioural	A		12	English	X		
E082	Taylor (2010) (Experiment 3)	T (PhD)	Behavioural	A		16	English	X		
E083	Taylor (2010) (Experiment 4)	T (PhD)	Behavioural	A		32	English	X		
E084	Taylor, Davis, & Rastle (2017)	JA	Behavioural + fMRI	A		24	English	X		
E085	Taylor, Davis, & Rastle (2019)	JA	Behavioural + fMRI	A		24	English	X	X	
E086	Taylor, Plunkett, & Nation (2011) (Experiment 1)	JA	Behavioural	A		16	English	X		
E087	Taylor, Plunkett, & Nation (2011) (Experiment 2)	JA	Behavioural	A		32	English	X		
E088	Taylor, Rastle, & Davis (2014)	JA	Behavioural + fMRI	A		19	English	X		

E089	Thakkar, Engelhart, Khodaparast, Abadzi, & Centanni (2020)	JA	Behavioural + stimulation	A		37	English	X		
E090	Tong, Duan, Shen, Yu, & Tong (2023)	JA	Behavioural + ERP	C		97	Chinese			X
E091	Tong, Zhang, & He (2020)(Experiment 1)	JA	Behavioural	C	With/without dyslexia	78	Cantonese			X
E092	Tong, Zhang, & He (2020) (Experiment 2)	JA	Behavioural	C	With/without dyslexia	85	Cantonese	X		
E093	Tong, Zhang, & He (2020) (Experiment 3)	JA	Behavioural	C	With/without dyslexia	86	Cantonese		X	
E094	Verwimp, Snellings, Wiers, & Tijms (2023)	JA	Behavioural	C		107	Dutch	X		
E095	Vidal, Viviani, Zoccolan, & Crepaldi (2021) (Experiment 1)	JA	Behavioural	A		22	Italian			X
E096	Vidal, Viviani, Zoccolan, & Crepaldi (2021) (Experiment 2)	JA	Behavioural	A		40	Italian			X
E097	Vidal, Viviani, Zoccolan, & Crepaldi (2021) (Experiment 3)	JA	Behavioural	A		36	Italian			X
E098	Wei, Manis, & Lu (2015)	T (PhD)	Behavioural + MRI	A		144	Chinese (43 participants); English (101 participants)	X		
E099	Williams, Panayotov, & Kempe (2020) (Experiment 1)	JA	Behavioural	A		112	English	X		

E100	Williams, Panayotov, & Kempe (2020) (Experiment 2a)	JA	Behavioural	A		112	English	X		
E101	Williams, Panayotov, & Kempe (2020) (Experiment 2b)	JA	Behavioural	A		112	English	X		
E102	Williams, Panayotov, & Kempe (2020) (Experiment 3)	JA	Behavioural	A		160	English	X		
E103	Williams, Panayotov, & Kempe (2022)	JA	Behavioural	A		320	English	X		
E104	Wu, Fang, Chen, Li, & Deng (2011) (Experiment 1)	CP	Behavioural	A		28	Chinese			X
E105	Wu, Fang, Chen, Li, & Deng (2011) (Experiment 2)	CP	Behavioural	A		28	Chinese			X
E106	Wu, Fang, Chen, Li, & Deng (2011) (Experiment 3)	CP	Behavioural	A		25	Chinese			X
E107	Xue, Zhao, Wang, Dong, Chen, & Xue (2017)	JA	Behavioural + stimulation	A		46	Chinese	X		
E108	Yoncheva, Blau, Maurer, & McCandliss (2010)	JA	Behavioural + ERP	A		30	English	X		
E109	Yoncheva, Wise, & McCandliss (2015)	JA	Behavioural + ERP	A		16	English	X		
E110	Younger & Booth (2018)	JA	Behavioural + stimulation	A		63	English	X		
E111	Zhao & Rueckl (2012) (Experiment 1)	T (PhD)	Behavioural	A		64	English	X	X	

E112	Zhao & Rueckl (2012) (Experiment 2)	T (PhD)	Behavioural	A		32	English	X	X	
E113	Zhao, Li, Elliott, & Rueckl (2018) (Experiment 1)	JA	Behavioural	A		48	English	X	X	
E114	Zhao, Li, Elliott, & Rueckl (2018) (Experiment 2)	JA	Behavioural	A		48	English	X	X	

Note. Publication type: BC = book chapter; CP = conference paper; JA = journal article; T (PhD) = PhD thesis; T (Mas) = Master's thesis. Age group: A = adults; C = children. Artificial orthographic patterns: O-P = orthography-phonology mappings; O-S = orthography-semantics mappings; O = orthotactic regularities. Experiments marked with an asterisk (*) represent those that did not clearly state the native language background of participants and are based on educated guesses.

Appendix for Chapter 3 (Experiments 1 & 2)

Appendix 3A. Additional Information on Participants' Language Backgrounds in Experiments 1 and 2

In Experiments 1 and 2, all participants were native English-speaking adults. As part of the language and background questionnaire, we also asked whether participants could read or write in another language. Although some participants reported knowledge of additional languages, there was no consistent patterns in the languages reported. Therefore, we did not further examine whether knowledge of another language influenced performance on the experimental tasks. Below, we provide additional details on participants' language backgrounds.

In Experiment 1, 10 of the 37 participants in the final sample reported that they could read or write in a second language. The languages reported were French ($N = 3$), Hebrew ($N = 2$), Filipino ($N = 1$), Mandarin Chinese ($N = 1$), Serbian ($N = 1$), Spanish ($N = 1$) and Tagalog ($N = 1$). Additionally, three participants reported proficiency in a third language: French ($N = 1$), German ($N = 1$) and Korean ($N = 1$).

In Experiment 2, out of the 155 participants included in the final sample, 20 reported that they could read or write in a second language. The languages reported were French ($N = 8$), German ($N = 4$), Spanish ($N = 4$), Arabic ($N = 1$), Danish (N = 1), Dutch ($N = 1$) and Mandarin Chinese ($N = 1$). In addition, three participants reported proficiency in a third language: French ($N = 2$) and Malay ($N = 1$).

Appendix 3B. Symbols and Letters/Number Mappings

Experiment and Phase	Letter/Number	Corresponding symbol
Experiments 1 and 2	T	
Exposure phase (first consonants)	P	
	N	
	S	
	D	
	M	
	L	
	G	
Experiments 1 and 2	B	
Testing phase (first consonants)	C	
	F	
	J	
	K	
	R	
	Q	
	H	
Experiment 2	1	
Testing phase preferential trials only	2	
(first consonants)	3	
	4	
	5	
	6	
	7	
	8	
Experiments 1 and 2	X	
Exposure and testing phase (second	Y	
consonants)	W	
	Z	

Appendix 3C. Word Lists for Exposure Phase in Experiments 1 and 2

		Experiment 1		Experiment 2 (Version 1)		Experiment 2 (Version 2)	
Word class	Target Word	List A	List B	List A	List B	List A	List B
Noun	coat	DOX	TOX	DOX	TOX	DAZ	TAZ
	car	MOX	POX	MOX	POX	MAZ	PAZ
	vase	LOX	NOX	LOX	NOX	LAZ	NAZ
	teacher	GOX	SOX	GOX	SOX	GAZ	SAZ
	book	DEX	TEX	TEX	DEX	TUW	DUW
	forest	MEX	PEX	PEX	MEX	PUW	MUW
	desk	LEX	NEX	NEX	LEX	NUW	LUW
	dog	GEX	SEX	SEX	GEX	SUW	GUW
Verb	walk	TOY	DOY	TOY	DOY	TAZ	DAZ
	write	POY	MOY	POY	MOY	PAZ	MAZ
	listen	NOY	LOY	NOY	LOY	NAZ	LAZ
	call	SOY	GOY	SOY	GOY	SAZ	GAZ
	play	TEY	DEY	DEY	TEY	DUW	TUW
	cook	PEY	MEY	MEY	PEY	MUW	PUW
	watch	NEY	LEY	LEY	NEY	LUW	NUW
	buy	SEY	GEY	GEY	SEY	GUW	SUW
Adjective	beautiful	TAZ	DAZ	DAZ	TAZ	DOX	TOX
	careful	PAZ	MAZ	MAZ	PAZ	MOX	POX
	busy	NAZ	LAZ	LAZ	NAZ	LOX	NOX
	clever	SAZ	GAZ	GAZ	SAZ	GOX	SOX
	happy	TUW	DUW	TUW	DUW	TEX	DEX
	healthy	PUW	MUW	PUW	MUW	PEX	MEX
	funny	NUW	LUW	NUW	LUW	NEX	LEX
	helpful	SUW	GUW	SUW	GUW	SEX	GEX
Adverb	politely	DAZ	TAZ	TAZ	DAZ	TOY	DOY
	frequently	MAZ	PAZ	PAZ	MAZ	POY	MOY
	immediately	LAZ	NAZ	NAZ	LAZ	NOY	LOY
	slowly	GAZ	SAZ	SAZ	GAZ	SOY	GOY
	badly	DUW	TUW	DUW	TUW	DEY	TEY
	directly	MUW	PUW	MUW	PUW	MEY	PEY
	exactly	LUW	NUW	LUW	NUW	LEY	NEY
	relatively	GUW	SUW	GUW	SUW	GEY	SEY

Appendix 3D. Word Lists for the Fill-in-the-blank Task in Experiments 1 and 2

Trial Type	Experiment 1				Experiment 2 (Version 1 and Version 2)			
	List A		List B		List A		List B	
	Correct	Foil	Correct	Foil	Correct	Foil	Correct	Foil
Semantic	BOX	BOY	KOX	KOY	BOX	BOY	KOX	KOY
Semantic	BEX	BEY	KEX	KEY	COX	COY	ROX	ROY
Semantic	COX	COY	ROX	ROY	FOX	FOY	QOX	QOY
Semantic	CEX	CEY	REX	REY	JOX	JOY	HOX	HOY
Semantic	FOX	FOY	QOX	QOY	KEX	KEY	BEX	BEY
Semantic	FEX	FEY	QEX	QEY	REX	REY	CEX	CEY
Semantic	JOX	JOY	HOX	HOY	QEX	QEY	FEX	FEY
Semantic	JEX	JEY	HEX	HEY	HEX	HEY	JEX	JEY
Semantic	KOY	KOX	BOY	BOX	KOY	KOX	BOY	BOX
Semantic	KEY	KEX	BEY	BEX	ROY	ROX	COY	COX
Semantic	ROY	ROX	COY	COX	QOY	QOX	FOY	FOX
Semantic	REY	REX	CEY	CEX	HOY	HOX	JOY	JOX
Semantic	QOY	QOX	FOY	FOX	BEY	BEX	KEY	KEX
Semantic	QEY	QEX	FEY	FEX	CEY	CEX	REY	REX
Semantic	HOY	HOX	JOY	JOX	FEY	FEX	QEY	QEX
Semantic	HEY	HEX	JEY	JEX	JEY	JEX	HEY	HEX
Graphotactic	KAZ	KAW	BAZ	BAW	BAZ	BAW	KAZ	KAW
Graphotactic	KUW	KUZ	BUW	BUZ	CAZ	CAW	RAZ	RAW
Graphotactic	RAZ	RAW	CAZ	CAW	FAZ	FAW	QAZ	QAW
Graphotactic	RUW	RUZ	CUW	CUZ	JAZ	JAW	HAZ	HAW
Graphotactic	QAZ	QAW	FAZ	FAW	KUW	KUZ	BUW	BUZ
Graphotactic	QUW	QUZ	FUW	FUZ	RUW	RUZ	CUW	CUZ
Graphotactic	HAZ	HAW	JAZ	JAW	QUW	QUZ	FUW	FUZ
Graphotactic	HUW	HUZ	JUW	JUZ	HUW	HUZ	JUW	JUZ
Graphotactic	BAZ	BAW	KAZ	KAW	KAZ	KAW	BAZ	BAW
Graphotactic	BUW	BUZ	KUW	KUZ	RAZ	RAW	CAZ	CAW
Graphotactic	CAZ	CAW	RAZ	RAW	QAZ	QAW	FAZ	FAW
Graphotactic	CUW	CUZ	RUW	RUZ	HAZ	HAW	JAZ	JAW
Graphotactic	FAZ	FAW	QAZ	QAW	BUW	BUZ	KUW	KUZ
Graphotactic	FUW	FUZ	QUW	QUZ	CUW	CUZ	RUW	RUZ
Graphotactic	JAZ	JAW	HAZ	HAW	FUW	FUZ	QUW	QUZ
Graphotactic	JUW	JUZ	HUW	HUZ	JUW	JUZ	HUW	HUZ

Trial Type	Experiment 1				Experiment 2 (Version 1 and Version 2)			
	List A		List B		List A		List B	
	Correct	Foil	Correct	Foil	Correct	Foil	Correct	Foil
Preference	BAX	BAZ	KAX	KAZ	1AX	1AZ	5AX	5AZ
Preference	CAX	CAZ	RAX	RAZ	2AX	2AZ	6AX	6AZ
Preference	FAX	FAZ	QAX	QAZ	3AX	3AZ	7AX	7AZ
Preference	JAX	JAZ	HAX	HAZ	4AX	4AZ	8AX	8AZ
Preference	BUX	BUW	KUX	KUW	5UX	5UW	1UX	1UW
Preference	CUX	CUW	RUX	RUW	6UX	6UW	2UX	2UW
Preference	FUX	FUW	QUX	QUW	7UX	7UW	3UX	3UW
Preference	JUX	JUW	HUX	HUW	8UX	8UW	4UX	4UW
Preference	KAY	KAZ	BAY	BAZ	5AY	5AZ	1AY	1AZ
Preference	RAY	RAZ	CAY	CAZ	6AY	6AZ	2AY	2AZ
Preference	QAY	QAZ	FAY	FAZ	7AY	7AZ	3AY	3AZ
Preference	HAY	HAZ	JAY	JAZ	8AY	8AZ	4AY	4AZ
Preference	KUY	KUW	BUY	BUW	1UY	1UW	5UY	5UW
Preference	RUY	RUW	CUY	CUW	2UY	2UW	6UY	6UW
Preference	QUY	QUW	FUY	FUW	3UY	3UW	7UY	7UW
Preference	HUY	HUW	JUY	JUW	4UY	4UW	8UY	8UW

Appendix 3E. Pre-registered Hypotheses and Justifications for Estimates of H_1 in Each Model in Experiment 2

Task and Trial Type	Hypothesis	Estimates of SD for H_1	Source of x	Model Syntax
Fill-in-the-blank: Semantic	Participants learn semantic patterns successfully.	0.73	Estimate of the intercept from an equivalent model in the pilot data.	glmer(Accuracy ~ 1+ (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_semantics_Ver1)
	Participants who are aware of semantic patterns perform better than participants who are unaware of semantic patterns.	1.82	Estimate of the main effects of semantic awareness from an equivalent model in the pilot data.	glmer(Accuracy ~ Awareness_sem.ct + Awareness_gra.ct +(1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_semantics_Ver1)
	Participants who are aware of semantic patterns perform better than participants who are unaware of semantic patterns. However, both groups perform above chance.	0.73	Overall estimate of participants' accuracy in semantic trials in the pilot data, regardless of semantic awareness status.	glmer(Accuracy ~ -1 + Awareness_sem + Awareness_gra.ct +(1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_semantics_Ver1)
	Participants who are aware of graphotactic patterns may do worse in the semantic trials than those who are unaware of graphotactic patterns .	1.95	Estimate of the main effects of graphotactic awareness from Experiment 1 of Singh et al. (2021), which is a methodologically similar study.	glmer(Accuracy ~ Awareness_sem.ct + Awareness_gra.ct + (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_semantics_Ver1)
	Participants who are aware of graphotactic patterns may do worse in the semantic trials than those who are unaware of graphotactic patterns. However, both groups perform above chance.	0.73	Overall estimate of participants' accuracy in semantic trials in the pilot data, regardless of semantic awareness status.	glmer(Accuracy ~ -1 + Awareness_sem.ct + Awareness_gra + (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_semantics_Ver1)
Fill-in-the-blank: Graphotactic	Participants learn graphotactic patterns successfully.	0.24	Estimate of the intercept for participants who showed graphotactic bias in the preference trials in the pilot data.	glmer(Accuracy ~ 1+ (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_graphotactics_Ver1)

	Participants who are aware of graphotactic patterns perform better than participants who are unaware of graphotactic patterns.	1.95	Estimate of the main effects of graphotactic awareness from Experiment 1 of Singh et al. (2021), which is a methodologically similar study.	glmer(Accuracy ~ Awareness_sem.ct + Awareness_gra.ct + (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_graphotactics_Ver1)
	Participants who are aware of graphotactic patterns perform better than participants who are unaware of graphotactic patterns. However, both groups perform above chance.	0.24	Estimate of the intercept for participants who showed a preference for graphotactic patterns in the preference trials in the pilot data.	glmer(Accuracy ~ -1 + Awareness_sem.ct + Awareness_gra + (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_graphotactics_Ver1)
	Participants who are aware of semantic patterns may do worse in the graphotactic trials than those who are unaware of semantic patterns.	1.69	Estimate of the main effects of semantic awareness for preference trials in the pilot data.	glmer(Accuracy ~ Awareness_gra.ct + Awareness_sem.ct + (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_graphotactics_Ver1)
	Participants who are aware of semantic patterns may do worse in the graphotactic trials than those who are unaware of semantic patterns. However, both groups perform above chance.	0.24	Estimate of the intercept for participants who showed a preference for graphotactic patterns in the preference trials in the pilot data.	glmer(Accuracy ~ -1 + Awareness_gra.ct + Awareness_sem + (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_graphotactics_Ver1)
Fill-in-the-blank: Preference	Overall, participants show a stronger bias for semantic patterns than graphotactic patterns in the preference trials.	0.44	Estimate of the intercept from an equivalent model in the pilot data.	glmer(Accuracy ~ 1+ (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_preference_Ver1)
	Participants who are aware of semantic patterns show a stronger semantic bias than participants who are unaware of semantic patterns. This is because we expect awareness to be reflected in performance at test.	1.69	Estimate of the main effects of semantic awareness for preference trials in the pilot data.	glmer(Accuracy ~ Awareness_sem.ct + Awareness_gra.ct + (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_preference_Ver1)
	Participants who are aware of semantic patterns show semantic bias whereas participants who are unaware of semantic patterns show graphotactic bias.	0.44	Overall estimate of participants' bias score in the preference trials in the pilot data, regardless of	glmer(Accuracy ~ -1 + Awareness_sem + Awareness_gra.ct + (1|participantCode), control=glmerControl(optimizer =

			semantic awareness status. Note that we will be testing for effects in the opposite direction in each case.	"bobyqa"), family = binomial, data = Fill_preference_Ver1)
	Participants who are aware of graphotactic patterns will show a stronger graphotactic bias than those who are unaware of graphotactic patterns.	1.95	Estimate of the main effects of graphotactic awareness from Experiment 1 of Singh et al. (2021), which is a methodologically similar study.	glmer(Accuracy ~ Awareness_sem.ct + Awareness_gra.ct + (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_preference_Ver1)
	Participants who are aware of graphotactic patterns will show a graphotactic bias, participants who are unaware will show a semantic bias.	0.44	Estimate of participants' bias score in the preference trials in the pilot data, regardless of semantic awareness status.	glmer(Accuracy ~ -1 + Awareness_sem.ct + Awareness_gra + (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_preference_Ver1)
Fill-in-the-blank: Semantic and graphotactic	Participants learn the semantic patterns better than the graphotactic patterns.	0.59	Estimate of the main effects of condition in the pilot data.	glmer(Accuracy ~ Condition.ct + (1|participantCode), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Fill_combined_Ver1)
Word category task	Participants' accuracy in the word category task correlates with their performance in the fill-in-the-blank task.	--	Coefficients from the same correlation analyses in Experiment 1 are used, if they were significant.	--
Spelling task	Participants' spelling accuracy correlates with their performance in the fill-in-the-blank and word category task.	0.46	Coefficients of the correlation between spelling performance and the legality judgement task from Singh et al. (2021)'s Experiment 3	--

Note. Due to oversight, we did not specify (1) the estimates for H_1 for the word category and spelling task and (2) our interests in the relationship between performance in the spelling task and word category task.

Appendix 3F. Computation of Bayes Factors for Correlations in Word Category Task and Spelling Task

To further quantify evidence for our correlation analyses in the word category task and spelling task, we reported Bayes Factors for these analyses, although this was not pre-registered due to an oversight. Here we provide detailed explanations how we computed the Bayes Factors for these correlation analyses:

To compute the Bayes Factors, we first need an estimate of the predicted effect size for H_1. For the word category task, the estimates were taken from Experiment 1 wherever the relevant effect had been found in that dataset (i.e., Spearman's was significant). These estimates were from the correlation between performance on the semantic trials and word category accuracy (z_r = .69) and the correlation between semantic bias and word category accuracy (z_r = .83). For the spelling task, we used the estimate from Singh et al.'s (2021) Experiment 3 where they observed a significant, positive correlation between performance in the legality judgement task and spelling scores (z_r = .46).

We then followed the step suggested by Dienes (n.d.) where we computed a Pearson's correlation *r* for each correlation analysis (we first ranked the data of each variable to meet normality assumptions) and transformed it with Fisher's *z* transform and obtained the corresponding standard errors. These Fisher's z transformed values and the standard errors were then used in each Bayes Factor calculation.

Appendix for Chapter 4 (Experiments 3 & 4)

Appendix 4A. Additional Information on Participants' Language Backgrounds in Experiments 3 and 4

As in Chapter 3, all participants recruited for Experiments 3 and 4 were native English-speaking adults. These participants also completed the language and background questionnaire, which included questions about their ability to read and write in another language. Once again, we did not observe any consistent patterns in the languages reported, so we did not further examine whether knowledge of another language influenced performance on the experimental tasks. Further details on participants' language backgrounds are provided below.

In Experiment 3, 14 of the 93 participants in the final sample reported that they could read or write in a second language. The languages reported were German ($N = 3$), Bengali ($N = 2$), Hindi ($N = 2$), Mandarin Chinese ($N = 2$), Arabic ($N = 1$), Greek ($N = 1$), Gujarati ($N = 1$), Spanish ($N = 1$) and Yoruba ($N = 1$). Additionally, three participants reported proficiency in a third language: Cantonese ($N = 1$), German ($N = 1$) and Mandarin Chinese ($N = 1$).

In Experiment 4, only 3 out of 50 participants in the final sample reported that they could read or write in a second language. The languages reported were Italian ($N = 1$), Swedish ($N = 1$) and Yoruba ($N = 1$). No participants reported proficiency in a third language.

Appendix 4B. Stimuli Used in the Exposure Phase in Experiments 3 and 4

Pattern	Frequency	Consistency	Singular noun phrase		Plural noun phrase				
			Spoken	Written	Spoken	Written (Word list 1)	Written (Word list 2)	Written (Word list 3)	Written (Word list 4)
Ending in -a	High	Consistent	/jɪm blɪgmə/	yim bligma	/vɒp blɪgmə/	vop bligmax	vop bligman	vop bligmat	vop bligmak
Ending in -a	High	Consistent	/jɪm tʃuːsə/	yim choosa	/vɒp tʃuːsə/	vop choosax	vop choosan	vop choosat	vop choosak
Ending in -a	High	Consistent	/jɪm dɜːkə/	yim dirca	/vɒp dɜːkə/	vop dircax	vop dircan	vop dircat	vop dircak
Ending in -a	High	Consistent	/jɪm dɹuːbə/	yim druba	/vɒp dɹuːbə/	vop drubax	vop druban	vop drubat	vop drubak
Ending in -a	High	Consistent	/jɪm fɑːlə/	yim fala	/vɒp fɑːlə/	vop falax	vop falan	vop falat	vop falak
Ending in -a	High	Consistent	/jɪm fɒgmə/	yim fogma	/vɒp fɒgmə/	vop fogmax	vop fogman	vop fogmat	vop fogmak
Ending in -a	High	Consistent	/jɪm hɛbɹə/	yim hebra	/vɒp hɛbɹə/	vop hebrax	vop hebran	vop hebrat	vop hebrak
Ending in -a	High	Consistent	/jɪm lɛmtə/	yim lemta	/vɒp lɛmtə/	vop lemtax	vop lemtan	vop lemtat	vop lemtak
Ending in -a	High	Consistent	/jɪm lɪθə/	yim litha	/vɒp lɪθə/	vop lithax	vop lithan	vop lithat	vop lithak
Ending in -a	High	Consistent	/jɪm mæmpə/	yim mampa	/vɒp mæmpə/	vop mampax	vop mampan	vop mampat	vop mampak
Ending in -a	High	Consistent	/jɪm hɔːnə/	yim hauna	/vɒp hɔːnə/	vop haunax	vop haunan	vop haunat	vop haunak
Ending in -a	High	Consistent	/jɪm pʌndɹə/	yim pundra	/vɒp pʌndɹə/	vop pundrax	vop pundran	vop pundrat	vop pundrak
Ending in -a	High	Consistent	/jɪm ɹo͡ʊfə/	yim rofa	/vɒp ɹo͡ʊfə/	vop rofax	vop rofan	vop rofat	vop rofak
Ending in -a	High	Consistent	/jɪm ʃɹɑːtə/	yim shrata	/vɒp ʃɹɑːtə/	vop shratax	vop shratan	vop shratat	vop shratak
Ending in -a	High	Consistent	/jɪm tɪθə/	yim titha	/vɒp tɪθə/	vop tithax	vop tithan	vop tithat	vop tithak
Ending in -a	High	Consistent	/jɪm tɹʌblə/	yim troubla	/vɒp tɹʌblə/	vop troublax	vop troublan	vop troublat	vop troublak
Ending in -i	Low	Consistent	/jɪm kɑːviː/	yim kavi	/vɒp kɑːviː/	vop kaviv	vop kavix	vop kavin	vop kavit
Ending in -i	Low	Consistent	/jɪm lo͡ʊmiː/	yim lomi	/vɒp lo͡ʊmiː/	vop lomiv	vop lomix	vop lomin	vop lomit
Ending in -i	Low	Consistent	/jɪm ɹeɪtiː/	yim raiti	/vɒp ɹeɪtiː/	vop raitiv	vop raitix	vop raitin	vop raitit
Ending in -i	Low	Consistent	/jɪm hɪdʒiː/	yim hiji	/vɒp hɪdʒiː/	vop hijiv	vop hijix	vop hijin	vop hijit
Ending in -i	Low	Consistent	/jɪm luːkiː/	yim luki	/vɒp luːkiː/	vop lukiv	vop lukix	vop lukin	vop lukit
Ending in -i	Low	Consistent	/jɪm pɪndiː/	yim pindi	/vɒp pɪndiː/	vop pindiv	vop pindix	vop pindin	vop pindit
Ending in -i	Low	Consistent	/jɪm spɑːɹiː/	yim spari	/vɒp spɑːɹiː/	vop spariv	vop sparix	vop sparin	vop sparit
Ending in -i	Low	Consistent	/jɪm bɜːgliː/	yim burgli	/vɒp bɜːgliː/	vop burgliv	vop burglix	vop burglin	vop burglit
Ending in -o	High	Inconsistent	/jɪm beɪgɹəʊ/	yim begro	/vɒp beɪgɹəʊ/	vop begrod	vop begrod	vop begrov	vop begrox

Ending in -o	High	Inconsistent	/jɪm pɪktəʊ/	yim pikto	/vɒp pɪktəʊ/	vop piktod	vop piktod	vop piktov	vop piktox
Ending in -o	High	Inconsistent	/jɪm hiːsəʊ/	yim heso	/vɒp hiːsəʊ/	vop hesok	vop hesov	vop hesov	vop hesox
Ending in -o	High	Inconsistent	/jɪm kluːkəʊ/	yim clucco	/vɒp kluːkəʊ/	vop cluccok	vop cluccov	vop cluccov	vop cluccox
Ending in -o	High	Inconsistent	/jɪm dɪpəʊ/	yim dippo	/vɒp dɪpəʊ/	vop dippok	vop dippod	vop dippox	vop dippox
Ending in -o	High	Inconsistent	/jɪm fɑːməʊ/	yim farmo	/vɒp fɑːməʊ/	vop farmok	vop farmod	vop farmox	vop farmox
Ending in -o	High	Inconsistent	/jɪm fluːkəʊ/	yim fluko	/vɒp fluːkəʊ/	vop flukok	vop flukod	vop flukov	vop flukon
Ending in -o	High	Inconsistent	/jɪm gɑːgləʊ/	yim garglo	/vɒp gɑːgləʊ/	vop garglok	vop garglod	vop garglov	vop garglon
Ending in -o	High	Inconsistent	/jɪm dʒɒŋgəʊ/	yim jongo	/vɒp dʒɒŋgəʊ/	vop jongok	vop jongod	vop jongov	vop jongon
Ending in -o	High	Inconsistent	/jɪm blɪtəʊ/	yim blitto	/vɒp blɪtəʊ/	vop blittok	vop blittod	vop blittov	vop blitton
Ending in -o	High	Inconsistent	/jɪm kɹiːsəʊ/	yim creaso	/vɒp kɹiːsəʊ/	vop creasok	vop creasod	vop creasox	vop creasox
Ending in -o	High	Inconsistent	/jɪm pɒmbəʊ/	yim pombo	/vɒp pɒmbəʊ/	vop pombok	vop pombod	vop pombox	vop pombox
Ending in -o	High	Inconsistent	/jɪm ɹænθəʊ/	yim rantho	/vɒp ɹænθəʊ/	vop ranthok	vop ranthov	vop ranthov	vop ranthox
Ending in -o	High	Inconsistent	/jɪm sæŋkləʊ/	yim sanclo	/vɒp sæŋkləʊ/	vop sanclok	vop sanclov	vop sanclov	vop sanclox
Ending in -o	High	Inconsistent	/jɪm sɔːgəʊ/	yim sorgo	/vɒp sɔːgəʊ/	vop sorgod	vop sorgod	vop sorgov	vop sorgox
Ending in -o	High	Inconsistent	/jɪm weɪpəʊ/	yim waypo	/vɒp weɪpəʊ/	vop waypod	vop waypod	vop waypov	vop waypox
Ending in -u	Low	Inconsistent	/jɪm bo͡ʊfuː/	yim bofu	/vɒp bo͡ʊfuː/	vop bofut	vop bofut	vop bofuk	vop bofud
Ending in -u	Low	Inconsistent	/jɪm nɒbluː/	yim knobblu	/vɒp nɒbluː/	vop knobblut	vop knobblut	vop knobbluk	vop knobblud
Ending in -u	Low	Inconsistent	/jɪm mɑːzuː/	yim mazu	/vɒp mɑːzuː/	vop mazun	vop mazuk	vop mazuk	vop mazud
Ending in -u	Low	Inconsistent	/jɪm ɹɒbuː/	yim robbu	/vɒp ɹɒbuː/	vop robbun	vop robbuk	vop robbuk	vop robbud
Ending in -u	Low	Inconsistent	/jɪm tɪmuː/	yim timmu	/vɒp tɪmuː/	vop timmun	vop timmut	vop timmud	vop timmud
Ending in -u	Low	Inconsistent	/jɪm lɪnduː/	yim lindu	/vɒp lɪnduː/	vop lindun	vop lindut	vop lindud	vop lindud
Ending in -u	Low	Inconsistent	/jɪm kaɪɹuː/	yim cairu	/vɒp kaɪɹuː/	vop cairun	vop cairut	vop cairuk	vop cairuv
Ending in -u	Low	Inconsistent	/jɪm spo͡ʊkuː/	yim spoku	/vɒp spo͡ʊkuː/	vop spokun	vop spokut	vop spokuk	vop spokuv

Note. Spoken stimuli were included only in Experiment 3. All written stimuli were identical across both experiments but were presented in Latin alphabet in Experiment 3 and in BACS-2 font.

Appendix 4C. Stimuli Used in the Testing Phase in Experiments 3 and 4

Table 4C.1

Stimuli used in the visual and auditory recognition tasks in Experiment 3

Type	Pattern	Visual recognition	Auditory recognition	
		Written	Spoken	Written
Exposure	Singular noun ending in -a	yim dirca	/jɪm ɹo͡ʊfə/	yim rofa
Exposure	Singular noun ending in -a	yim hauna	/jɪm dɹuːbə/	yim druba
Exposure	Singular noun ending in -a	yim pundra	/jɪm lɛmtə/	yim lemta
Exposure	Singular noun ending in -i	yim lomi	/jɪm pɪndiː/	yim pindi
Exposure	Singular noun ending in -i	yim luki	/jɪm bɜːgliː/	yim burgli
Exposure	Singular noun ending in -i	yim spari	/jɪm ɹeɪtiː/	yim raiti
Exposure	Singular noun ending in -o	yim heso	/jɪm ɹænθəʊ/	yim rantho
Exposure	Singular noun ending in -o	yim jongo	/jɪm weɪpəʊ/	yim waypo
Exposure	Singular noun ending in -o	yim pombo	/jɪm fluːkəʊ/	yim fluko
Exposure	Singular noun ending in -u	yim mazu	/jɪm spo͡ʊkuː/	yim spoku
Exposure	Singular noun ending in -u	yim lindu	/jɪm ɹɒbuː/	yim robbu
Exposure	Singular noun ending in -u	yim cairu	/jɪm tɪmuː/	yim timmu
Foil	Novel article + novel noun with learned suffix	bes homa	/tæv ɛmkə/	tav emka
Foil	Novel article + novel noun with learned suffix	bes grati	/tæv nɔːmɪ/	tav normi
Foil	Novel article + novel noun with learned suffix	bes lorro	/tæv nɛləʊ/	tav nello
Foil	Novel article + novel noun with learned suffix	bes daiku	/tæv tɛstuː/	tav testu
Foil	Learned article + novel noun with novel suffix	yim cosple	/jɪm mæŋkbɪl/	yim mankbill
Foil	Learned article + novel noun with novel suffix	yim thoosy	/jɪm lɛndəm/	yim lendem
Foil	Learned article + novel noun with novel suffix	yim aining	/jɪm dɒnfɛs/	yim donfess
Foil	Learned article + novel noun with novel suffix	yim enmosh	/jɪm pɹæpəl/	yim prapple
Foil	Novel article + novel noun with novel suffix	cen admusts	/sæf gʌmdɹɛp/	saf gumdrep
Foil	Novel article + novel noun with novel suffix	kad docab	/bʌk snæppɪθ/	buk snappith
Foil	Novel article + novel noun with novel suffix	fas seginner	/næks wʊmədɪʃ/	nax womadish
Foil	Novel article + novel noun with novel suffix	dof mestowal	/ko͡ʊv æmpəsæŋ/	cov ampersang

Note. Stimuli for the auditory recognition task were presented in audio form only.

Table 4C.2

Stimuli used in the visual recognition task in Experiment 4

Type	Pattern	Stimuli
Exposure	Singular noun ending in -a	yim dirca
Exposure	Singular noun ending in -a	yim hauna
Exposure	Singular noun ending in -a	yim pundra
Exposure	Singular noun ending in -i	yim lomi
Exposure	Singular noun ending in -i	yim luki
Exposure	Singular noun ending in -i	yim spari
Exposure	Singular noun ending in -o	yim heso
Exposure	Singular noun ending in -o	yim jongo
Exposure	Singular noun ending in -o	yim pombo
Exposure	Singular noun ending in -u	yim mazu
Exposure	Singular noun ending in -u	yim lindu
Exposure	Singular noun ending in -u	yim cairu
Foil	Novel article (five-letter) + novel noun with learned suffix	coreb homa
Foil	Novel article (five-letter) + novel noun with learned suffix	coreb grati
Foil	Novel article (five-letter) + novel noun with learned suffix	coreb lorro
Foil	Novel article (five-letter) + novel noun with learned suffix	coreb daiku
Foil	Learned article in the wrong position + novel noun with novel suffix	cosple yim
Foil	Learned article in the wrong position + novel noun with novel suffix	thoosy yim
Foil	Learned article in the wrong position + novel noun with novel suffix	aining yim
Foil	Learned article in the wrong position + novel noun with novel suffix	enmosh yim
Foil	Novel article + novel noun with novel suffix (presented in BACS-1 font)	cen admusts
Foil	Novel article + novel noun with novel suffix (presented in BACS-1 font)	kad docab
Foil	Novel article + novel noun with novel suffix (presented in BACS-1 font)	fas seginner
Foil	Novel article + novel noun with novel suffix (presented in BACS-1 font)	dof tarnival

Note. Unless otherwise specified, all stimuli were presented in BACS-2 font.

Table 4C.3

Stimuli used in the fill-in-the-blank task in Experiments 3 and 4

Pattern	Singular noun phrase (written)	Singular noun phrase (spoken)	Plural noun phrase (spoken)	Option 1	Option 2	Option 3	Option 4	Option 5	Option 6
Ending in -a	yim claza	/jɪm klɑːzə/	/vɒp klɑːzə/	clazax	clazav	clazad	clazak	clazat	clazan
Ending in -a	yim hubna	/jɪm hʌbnə/	/vɒp hʌbnə/	hubnax	hubnav	hubnad	hubnak	hubnat	hubnan
Ending in -a	yim binga	/jɪm bɪŋgə/	/vɒp bɪŋgə/	bingax	bingav	bingad	bingak	bingat	bingan
Ending in -a	yim trampla	/jɪm tɹæmplə/	/vɒp tɹæmplə/	tramplax	tramplav	tramplad	tramplak	tramplat	tramplan
Ending in -a	yim grova	/jɪm gɹo͡ʊvə/	/vɒp gɹo͡ʊvə/	grovax	grovav	grovad	grovak	grovat	grovan
Ending in -a	yim mecra	/jɪm mɛkɹə/	/vɒp mɛkɹə/	mecrax	mecrav	mecrad	mecrak	mecrat	mecran
Ending in -a	yim felma	/jɪm fɛlmə/	/vɒp fɛlmə/	felmax	felmav	felmad	felmak	felmat	felman
Ending in -a	yim prandpa	/jɪm pɹændpə/	/vɒp pɹændpə/	prandpax	prandpav	prandpad	prandpak	prandpat	prandpan
Ending in -o	yim vitbo	/jɪm vɪtbəʊ/	/vɒp vɪtbəʊ/	vitbox	vitbov	vitbod	vitbok	vitbot	vitbon
Ending in -o	yim tinklo	/jɪm tɪŋkləʊ/	/vɒp tɪŋkləʊ/	tinklox	tinklov	tinklod	tinklok	tinklot	tinklon
Ending in -o	yim hodo	/jɪm ho͡ʊdəʊ/	/vɒp ho͡ʊdəʊ/	hodox	hodov	hodod	hodok	hodot	hodon
Ending in -o	yim cacho	/jɪm kætʃəʊ/	/vɒp kætʃəʊ/	cachox	cachov	cachod	cachok	cachot	cachon
Ending in -o	yim forepo	/jɪm fɔːpəʊ/	/vɒp fɔːpəʊ/	forepox	forepov	forepod	forepok	forepot	forepon
Ending in -o	yim borno	/jɪm bɔːnəʊ/	/vɒp bɔːnəʊ/	bornox	bornov	bornod	bornok	bornot	bornon
Ending in -o	yim lanjo	/jɪm lændʒəʊ/	/vɒp lændʒəʊ/	lanjox	lanjov	lanjod	lanjok	lanjot	lanjon
Ending in -o	yim gezzo	/jɪm gɛzəʊ/	/vɒp gɛzəʊ/	gezzox	gezzov	gezzod	gezzok	gezzot	gezzon
Ending in -i	yim angli	/jɪm æŋgliː/	/vɒp æŋgliː/	anglix	angliv	anglid	anglik	anglit	anglin
Ending in -i	yim bemi	/jɪm bɛmiː/	/vɒp bɛmiː/	bemix	bemiv	bemid	bemik	bemit	bemin
Ending in -i	yim chaki	/jɪm tʃɑːkiː/	/vɒp tʃɑːkiː/	chakix	chakiv	chakid	chakik	chakit	chakin
Ending in -i	yim funti	/jɪm fʌntiː/	/vɒp fʌntiː/	funtix	funtiv	funtid	funtik	funtit	funtin
Ending in -i	yim delzi	/jɪm dɛlziː/	/vɒp dɛlziː/	delzix	delziv	delzid	delzik	delzit	delzin
Ending in -i	yim gronchi	/jɪm gɹɒntʃiː/	/vɒp gɹɒntʃiː/	gronchix	gronchiv	gronchid	gronchik	gronchit	gronchin
Ending in -i	yim habbi	/jɪm hæbiː/	/vɒp hæbiː/	habbix	habbiv	habbid	habbik	habbit	habbin
Ending in -i	yim passi	/jɪm pæsiː/	/vɒp pæsiː/	passix	passiv	passid	passik	passit	passin
Ending in -u	yim bingu	/jɪm bɪŋguː/	/vɒp bɪŋguː/	bingux	binguv	bingud	binguk	bingut	bingun

Ending in -u	yim huru	/jɪm hjʊəɹuː/	/vɒp hjʊəɹuː/	hurux	huruv	hurud	huruk	hurut	hurun
Ending in -u	yim algu	/jɪm ælguː/	/vɒp ælguː/	algux	alguv	algud	alguk	algut	algun
Ending in -u	yim razzlu	/jɪm ɹæzluː/	/vɒp ɹæzluː/	razzlux	razzluv	razzlud	razzluk	razzlut	razzlun
Ending in -u	yim mayu	/jɪm mɑːjuː/	/vɒp mɑːjuː/	mayux	mayuv	mayud	mayuk	mayut	mayun
Ending in -u	yim faunu	/jɪm fɔːnuː/	/vɒp fɔːnuː/	faunux	faunuv	faunud	faunuk	faunut	faunun
Ending in -u	yim thelmu	/jɪm θɛlmuː/	/vɒp θɛlmuː/	thelmux	thelmuv	thelmud	thelmuk	thelmut	thelmun
Ending in -u	yim hordu	/jɪm hɔːduː/	/vɒp hɔːduː/	hordux	horduv	hordud	horduk	hordut	hordun

Note. Spoken stimuli were included only in Experiment 3. All written stimuli were identical across both experiments but were presented in Latin alphabet in Experiment 3 and in BACS-2 font.

Appendix 4D. Pre-registered Hypotheses and Justifications for Estimates of H_1 in Each Model in Experiments 3 and 4

Task and Data	Model syntax	Hypothesis	Coefficient from which estimate and SD obtained as model of the data	Estimates of SD for H_1	Source of x
Fill-in-the-blank task: Consistent/ inconsistent item accuracy	*Separate analyses were conducted on consistent and inconsistent item data.* glmer(Accuracy ~ Frequency.ct*Condition.ct+ (Frequency.ct|Participant_ID), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Consistent_accuracy)	Participants learn the graphotactic patterns successfully.	Intercept	1.81 (CON) / 0.89 (INCON)	0.2 is the estimate of the intercept for graphotactic learning in the 2AFC task with adults in Singh et al. (2021)'s Experiment 2. However, as our chance level here is 1/6 (i.e., log-odds = -1.61) for consistent items and 1/3 for inconsistent items (i.e., log-odds = -0.69), the estimate of SD is set as intercept 0.2 + 1.61 (consistent items) and 0.2 + 0.69 (inconsistent items).
		Participants perform better in high-frequency items than low-frequency items.	Main effect of frequency	0.91 (CON) / 0.45 (INCON)	As there is no methodologically similar experiment from which we could extract an informative prior, we used the maximum approach and estimated the main effect of frequency as half of the intercept (consistent items: 1.81/2 = 0.91; inconsistent items = 0.89/2 = 0.45).
		Participants in the SE condition perform better than those in the GR condition.	Main effect of condition	0.41	0.41 is the main effect of trial types from a methodologically similar experiment in Law et al. (2025, Experiment 2 Version 1) which compared participants' semantic and graphotactic learning.
		As an exploratory analysis, we examine whether there is an interaction between condition and frequency.	Interaction between condition and frequency	0.20	As there is no methodologically similar experiment from which we could extract an informative prior, we used the maximum approach and estimated the interaction effect between condition and frequency as half of the main effects of condition (0.41/2 = 0.20).
Fill-in-the-blank task: Consistent/ inconsistent item accuracy	*Separate analyses were conducted on consistent and inconsistent item data, by SE/GR conditions.* glmer(Accuracy ~ Frequency.ct*Awareness.ct + (Frequency.ct|Participant_ID), control=glmerControl(optimizer =	Within each condition (SE/GR), participants who can accurately describe the patterns (i.e., showing explicit awareness of the patterns) perform better than those who cannot.	Main effect of awareness	1.23	1.23 is the main effect of semantic awareness from a methodologically similar experiment in Law et al. (2025, Experiment 2 Version 1).

	"bobyqa"), family = binomial, data = Consistent_accuracy_SE)	As an exploratory analysis, we examine whether there is an interaction between awareness and frequency.	Interaction between awareness and frequency	0.62	As there is no methodologically similar experiment from which we could extract an informative prior, we used the maximum approach and estimated the interaction effect between awareness and frequency as half of the main effects of awareness (1.23/2 = 0.62).
Fill-in-the-blank task: Consistent/ inconsistent item accuracy	*Separate analyses were conducted on consistent and inconsistent item data.* glmer(Accuracy ~ Awareness.ct*Frequency.ct*Condition.ct + (Frequency.ct|Participant_ID), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Consistent_accuracy)	As an exploratory analysis, we examine whether there is an interaction between awareness and condition.	Interaction between awareness and condition	0.62	As there is no methodologically similar experiment from which we could extract an informative prior, we used the maximum approach and estimated the interaction effect between awareness and condition as half of the main effects of awareness (1.23/2 = 0.62).
Fill-in-the-blank task: Consistent/ inconsistent item accuracy	*Separate analyses were conducted on consistent and inconsistent item data, by SE/GR conditions.* glmer(Accuracy ~ 1 + Frequency.ct + (Frequency.ct|Participant_ID), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Consistent_accuracy_SE))	Participants in both SE and GR conditions perform above chance.	Intercept	1.81 (CON) / 0.89 (INCON)	0.2 is the estimate of the intercept for graphotactic learning in the 2AFC task with adults in Singh et al. (2021)'s Experiment 2. However, as our chance level here is 1/6 (i.e., log-odds = -1.61) for consistent items and 1/3 for inconsistent items (i.e., log-odds = -0.69), the estimate of SD is set as intercept 0.2 + 1.61 (consistent items) and 0.2 + 0.69 (inconsistent items).
Fill-in-the-blank task: Consistent/ inconsistent item accuracy	*Separate analyses were conducted on consistent and inconsistent item data, by conditions (SE/GR) and awareness status (aware/unaware).* glmer(Accuracy ~ 1 + Frequency.ct + (Frequency.ct|Participant_ID), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Consistent_accuracy_SE_unaware)	Within each condition (SE/GR), participants perform above chance no matter they are aware or unaware of the patterns.	Intercept	1.81 (CON) / 0.89 (INCON)	0.2 is the estimate of the intercept for graphotactic learning in the 2AFC task with adults in Singh et al. (2021)'s Experiment 2. However, as our chance level here is 1/6 (i.e., log-odds = -1.61) for consistent items and 1/3 for inconsistent items (i.e., log-odds = -0.69), the estimate of SD is set as intercept 0.2 + 1.61 (consistent items) and 0.2 + 0.69 (inconsistent items).

Fill-in-the-blank task: Inconsistent item (dominance index)	*Separate analyses were conducted on consistent and inconsistent item data.* glmer(Dominant ~ Frequency.ct*Condition.ct + (Frequency.ct|Participant_ID), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Inconsistent_dominance)	Participants learn that one suffix is more dominant than the other among the two plausible suffixes.	Intercept	1.10	We expect that participants' answers mirror the input statistics, where they choose the dominant suffix in 75% of the items and the non-dominant suffix for 25% of the items. For a probability of 0.75, the log-odds is 1.10.
		Participants choose more of the dominant option in high-frequency items than low-frequency items.	Main effect of frequency	0.55	As there is no methodologically similar experiment from which we could extract an informative prior, we used the maximum approach and estimated the main effect of frequency as half of the intercept (1.10/2 = 0.55).
		As an exploratory analysis, we examine whether there is a main effect of condition.	Main effect of condition	0.41	0.41 is the main effect of trial types from a methodologically similar experiment in Law et al. (2025, Experiment 2 Version 1) which compared participants' semantic and graphotactic learning.
		As an exploratory analysis, we examine whether there is an interaction between condition and frequency.	Interaction between condition and frequency	0.20	As there is no methodologically similar experiment from which we could extract an informative prior, we used the maximum approach and estimated the interaction effect between condition and frequency as half of the main effects of condition (0.41/2 = 0.20).
Word meaning task	*Separate analyses were conducted on consistent and inconsistent item data.* cor(full_data$word_meaning_accuracy, full_data$fill_accuracy, method = 'pearson')	Within the SE condition, there is a correlation between performance in word meaning and accuracy in the fill-in-the-blank task.	Pearson's correlations	0.57	0.57 is the significant correlation between semantic pattern learning and word category knowledge from Law et al. (2025).
Fill-in-the-blank task: Consistent/ inconsistent item accuracy	*Separate analyses were conducted on consistent and inconsistent item data from the combined datasets of Experiment 3 and Experiment 4.* glmer(Accuracy ~ Experiment.ct + (1|Participant_ID), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Consistent_accuracy_all)	As an exploratory analysis, we examine whether participants in Experiment 3 perform better than those in Experiment 4.	Main effect of experiment	0.40	0.40 is the main effect of versions from a methodologically similar experiment in Law et al. (2025, Experiment 2) which compared participants' learning of semantic regularities when assigned to nouns and verbs (Version 1) and adjectives and adverbs (Version 2).
Fill-in-the-blank task:	*This analysis was conducted on inconsistent item data from the combined*	As an exploratory analysis, we examine whether there is an	Interaction between condition and experiment	0.20	As there is no methodologically similar experiment from which we could extract an informative prior, we used the maximum approach and estimated the

Inconsistent item accuracy	*datasets of Experiment 3 and Experiment 4.* glmer(Accuracy ~ Condition.ct*Experiment.ct*Frequency.ct + (Frequency.ct|Participant_ID), control=glmerControl(optimizer = "bobyqa"), family = binomial, data = Inconsistent_accuracy_all)	interaction effect between condition and experiment in inconsistent items.	interaction effect between condition and experiment as half of the main effects of condition (0.41/2 = 0.20).

Note. With the exception of the two analyses examining the main effect of experiment and the interaction between condition and experiment, all other analyses were conducted separately within each experiment. As noted in the main text, the final analyses differ slightly from the pre-registered analysis plan, as it became clear that including all experimented variables in a single model and the frequency variable in the random slope structure better represents the design of our experiment. CON = consistent items; INCON = inconsistent items.

Appendix 4E. Results of Exploratory Analyses in Experiments 3 and 4

Condition	Item type	Effect	Results	
			Experiment 3	Experiment 4
NA	Inconsistent (dominance)	Condition*	$BF_{(0, 0.41)} = 0.79$, *RR* [0, 1.49], β = 0.26, *SE* = 0.44, *p* = .56	$BF_{(0, 0.41)} = 1.10$, *RR* [0, 4.22], β = 0.79, *SE* = 0.59, *p* = .18
NA	Inconsistent (dominance)	Condition × Frequency*	$BF_{(0, 0.20)} = 0.97$, *RR* [0, 2.43], β = 0.30, *SE* = 0.80, *p* = .71	$BF_{(0, 0.20)} = 0.98$, *RR* [0, 2.68], β = -0.47, *SE* = 0.80, *p* = .56
NA	Consistent	Condition × Frequency	$BF_{(0, 0.20)} = 1.01$, *RR* [0, 4.08], β = -0.84, *SE* = 0.75, *p* = .26	$BF_{(0, 0.20)} = 0.98$, *RR* [0, 2.30], β = 0.46, *SE* = 0.61, *p* = .45
NA	Inconsistent	Condition × Frequency	$BF_{(0, 0.20)} = 1.07$, *RR* [0, 1.83], β = 1.11, *SE* = 0.68, *p* = .10	$BF_{(0, 0.20)} = 0.97$, *RR* [0, 1.60], β = -0.34, *SE* = 0.37, *p* = .36
NA	Consistent	Condition × Awareness*	$BF_{(0, 0.62)} = 1.08$, *RR* [0, > 4.59], β = -1.65, *SE* = 1.22, *p* = .17	This analysis was not performed as no participants in the GR condition were aware of the patterns.
NA	Inconsistent	Condition × Awareness*	$BF_{(0, 0.62)} = 1.15$, *RR* [0, > 4.59] β = -1.27, *SE* = 0.88, *p* = .15	This analysis was not performed as no participants in the GR condition were aware of the patterns.
SE	Consistent	Awareness × Frequency	$BF_{(0, 0.62)} = 0.98$, *RR* [0, > 4.59], β = -1.36, *SE* = 1.56, *p* = .38	$BF_{(0, 0.62)} = 0.95$, *RR* [0, > 4.59], β = -0.71, *SE* = 1.63, *p* = .66
SE	Inconsistent	Awareness × Frequency	$BF_{(0, 0.62)} = 0.96$, *RR* [0, > 4.59], β = -1.07, *SE* = 1.24, *p* = .39	$BF_{(0, 0.62)} = 0.82$, *RR* [0, 2.54], β = 0.27, *SE* = 0.85, *p* = .75
GR	Consistent	Awareness × Frequency	$BF_{(0, 0.62)} = 0.98$, *RR* [0, > 4.59] β = -1.48, *SE* = 1.90, *p* = .44	This analysis was not performed as no participants in the GR condition were aware of the patterns.
GR	Inconsistent	Awareness × Frequency	$BF_{(0, 0.62)} = 0.89$, *RR* [0, 3.51], β = -0.14, *SE* = 1.23, *p* = .91	This analysis was not performed as no participants in the GR condition were aware of the patterns.

Note. As these analyses are exploratory with no predetermined directions for the interaction effect, all Bayes factors were tested with two-tailed predictions. Analyses marked with an asterisk (*) were not pre-registered but are reported for completeness. SE = semantics cues condition; GR = graphotactic constraints only condition.