The acquisition and generalisation of orthography-phonology correspondences

Rebecca Rosamond Lawrence

Submitted for the degree of Doctor of Philosophy

Royal Holloway, University of London

Department of Psychology

April 2022

Declaration of authorship

I, Rebecca Lawrence, hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Signed: Rebecca lawrence

Dated: 18.04.2022

For Gordon Gerald Lawrence (1929 – 2018)

and

Joan Margaret Hobbis (1918 – 2018)

Acknowledgements

Firstly, I would like to thank my supervisors Kathy Rastle and Jakke Tamminen for their generous advice, time and support, and for the opportunity to carry out this research which I have enjoyed so much. Thanks also to the Department of Psychology at Royal Holloway for creating such a welcoming and stimulating environment to work in. Further afield, I am grateful to Charles Yang and Elissa Newport for thought-provoking discussions, and to Napoleon Katsos for always encouraging me to pursue language research.

Thank you to the wonderful members of the Rastle lab during my time at Royal Holloway (including Ana Ulicheva, Clare Lally, Chloe Newbury, Adam Jowett, Benedetta Cevoli, Becky Crowley and Oxana Grosseck) for your help and collaboration. I would also like to thank everyone in our PhD cohort (especially Alex, Aysha, Beatrice, Clare L., Clare M., David, Giusi, Isaac, Louisa and Rachel) for your advice, friendship and many happy memories. I'm very lucky to have shared this experience with you all, and also grateful that you withstood my endless questions.

I would like to thank my parents for all their love, generosity, and belief in what I could achieve, and my sister Ellie for her endless encouragement – your own dedication has inspired me to keep going. Finally, thank you to James for your kindness, reassurance and total support (not to mention proof-reading and technical assistance) over 11 years and several degrees. I think it's my turn to cook.

Abstract

For both skilled and developing readers, reading unfamiliar words aloud requires knowledge about the correspondences between spelling and sound in their writing system. These correspondences may not be entirely consistent, making the writing system quasi-regular. This thesis explores how readers acquire such spelling-sound correspondences through text experience, and how they generalise this knowledge to words they have not encountered before.

In the literature on word reading, debate about whether readers use categorical rules or statistical information to read words aloud is unresolved. To address this issue, I apply the Tolerance Principle (TP) (Yang, 2016), a recently-proposed theory of rule-productivity in spoken language acquisition, to the field of reading. The primary aim of the thesis is to assess whether the TP can predict which spelling-sound correspondences readers use productively. Experiment 1 explores whether the TP can predict readers' productive use of familiar spelling-sound correspondences. I conduct a nonword reading aloud task with adults and children (aged 8-9), using the TP to predict which spelling-sound correspondences in the English writing system readers use to pronounce novel words. Results show that the TP predicts adults' and children's pronunciations of letter sequences more successfully than three extant models of reading.

This thesis also aims to contribute to the literature on statistical learning, and its relationship with reading. To explore the TP in this context, I conduct a series of artificial orthography learning experiments with adults and children (aged 9-10) to assess whether the TP can predict the acquisition and generalisation of novel spelling-sound correspondences. In Experiment 2, the TP is able to predict adult and child participants' generalisation beyond the effect of token frequency distributions in the input. In Experiment 3, adults' generalisation is moderated by increasing the relative frequency of irregular items during training. In Experiment 4, the TP does not successfully predict the use of contextually-conditioned pronunciation sub-rules for most adult learners. Overall, this thesis makes a novel contribution to our knowledge of skilled and developing reading, and to our understanding about how statistical information from the input is used during learning and generalisation.

OSF Locations

Data and analysis scripts are available on the Open Science Framework at the following locations:

Experiment 1: https://osf.io/t8c9x/?view_only=81d75621ae44479a95c8d912ca0ebd25

Experiment 2: https://osf.io/pbjqu/?view_only=4a5601d01cc64ca59cb741aceb3990da

Experiment 3: https://osf.io/z3y5q/?view_only=5f10575f9e1240588236381e4cff3952

Experiment 4: https://osf.io/kn3db/?view_only=ef297b4ec23b41828a27c1dd977cbc64

Contents

List of Tables

List of Figures

Chapter 1: An introduction to word reading

Productivity is a fundamental capacity of the human mind: we can make infinite use of finite means, even within quasi-regular systems such as those of spoken and written language. Learners are able to overcome exceptions to the regular patterns they see and hear in order to develop fully productive systems, enabling them to produce sentences they have not heard before and read words they have never seen. This ability to apply acquired knowledge to novel situations is known as *generalisation*. This thesis will investigate whether a theory recently posited to explain rule-learning and generalisation in spoken language, the Tolerance Principle (Yang, 2016), underlies our ability to generalise knowledge of spelling-sound correspondences in written language. As well as adding to our understanding of the processes underlying word reading, this investigation will be relevant to fundamental issues involving the nature of cognitive representations and processing more generally.

1.1 Quasi-regularity in alphabetic writing systems

Reading a word aloud requires converting its written (orthographic) form to its spoken (phonological) form. Writing systems vary in the ways in which written symbols represent spoken language. Alphabetic writing systems use a set of symbols that each represent individual sounds; small orthographic units (graphemes) correspond systematically to individual phonological units (phonemes). Whilst all alphabetic systems involve strong associations between spelling (orthography) and sound (phonology), they can differ in their *orthographic depth*, or the *transparency* with which symbols relate to sounds. For example, languages such as Greek and Serbian have *shallow* orthographies, as there is high consistency in the correspondences between graphemes and phonemes. English has a *deep* orthography, as there is some level of inconsistency in the correspondences between graphemes and phonemes. Quantitatively, the orthographic depth (i.e., the average number of pronunciations of a grapheme) of monosyllabic words in English is 1.7, whilst Serbo-Croat, for example, has an orthographic depth of 1.0 (Vousden, 2008). Because of the inconsistency that can be observed in its spelling-sound correspondences, English is an example of a *quasi-regular* alphabetic system (as will be laid out in detail below). The ways in which we learn, use and generalise correspondences between spelling and sound in quasi-regular systems are not fully understood, and will be explored in this thesis. Whilst my investigation will focus on English, the method and conclusions reached could equally be applied to other quasi-regular alphabetic writing systems.

1.1.1 Regularity and consistency in the English writing system

As an alphabetic writing system, English uses predictable mappings between graphemes and phonemes. For example, the grapheme “d” corresponds to the phoneme /d/. Written words can be read aloud by blending the phonemes that correspond to the word’s orthographic form, for example “dog” - /d - ɒ - g/ - / dɒg /. This is a powerfully productive process: knowledge of spelling-sound correspondences can be *generalised* to pronounce unfamiliar written words, e.g. “dop” - /dɒp/. Once developing readers are able to assemble knowledge of individual graphemes and phonemes to generate the phonological form of a word (Castles et al., 2019), this generalisation ability also facilitates the reading acquisition process. In this way, knowledge about the pronunciation of familiar words can be used to pronounce other words as reading experience develops.

The most frequent correspondences between graphemes and phonemes in English words (e.g. “d” - /d/) can be described as *regular*. Regularity is a binary notion: in terms of reading, a pronunciation is regular if it matches a pre-determined set of spelling-sound correspondences, such as the most frequently occurring correspondences in a given vocabulary. Pronunciation regularity can be applied to any orthographic unit (i.e., shorter or longer sequences of letters), but in the reading literature it tends to be applied to smaller orthographic units. Therefore, the regularity of a particular pronunciation is usually assessed according to the most frequent individual grapheme-phoneme correspondences (GPCs). Words which can be pronounced accurately using these GPCs are sometimes described as “regular” words. However, in English, not all words can be pronounced accurately using only GPCs; for example, “son” is pronounced /sʌn/, rather than - /sɒn/. Such words as these are sometimes described as “irregular” or “exception” words[1]. In fact, around 20% of English monosyllabic words cannot be pronounced

[1] Not all theories of word reading ascribe to the abstract notion of regularity, nor to the distinction between “regular” and “irregular” words, as will be discussed further below. However, it is worth noting how a regular or an irregular pronunciation of a word or letter string may be categorised.

using only regular GPCs (Coltheart et al., 2001). Indeed, graphemes in English orthography can often be pronounced in a variety of different ways, allowing a one-to-many mapping between graphemes and phonemes. For example, the grapheme “u” is pronounced differently in the words “push” - /pʊʃ/, “bun” - /b^n/ and “truth” - /tru:θ/. This property makes the relationship between spelling and sound *inconsistent*, even within a writing system that is broadly predictable and fully productive.

Unlike regularity, the *consistency* of spelling-sound relationships is a graded notion: if an orthographic unit is pronounced the same way in the majority of words, the spelling-sound relationship is more consistent; if it is pronounced in different ways in many words, it is less consistent. Further, consistency can apply to any orthographic unit (or *grain size*). For example, it can be used to describe the strength of the relationship between individual graphemes and phonemes, as described above. It can also be used to highlight very unusual spelling-sound relationships at the whole-word level, as in words such as “yacht” - /jɒt/. However, in the reading literature, consistency is often discussed in terms of multi-letter sequences. This is because in English, multi-letter sequences can sometimes offer additional pronunciation consistency to the apparent inconsistency of individual grapheme-phoneme relationships (Treiman et al. 1995). For instance, the grapheme “oo” is most frequently pronounced /u:/, as in “moon” - /mu:n/ and “boot” - /bu:t/, but can also have different pronunciations, such as in “blood” - /blʌd/, “look” - /lʊk/ and “brooch” - /brəʊtʃ/. Therefore, the pronunciation of this individual grapheme is inconsistent. However, when “oo” is followed by “k”, this longer letter sequence “ook” is almost always pronounced /ʊk/, as in the words “took” - /tʊk/ and “book” - /bʊk/. These types of pronunciation patterns are sometimes described as “context-sensitive” (e.g. Plaut et al., 1996), as the pronunciation of one grapheme is informed by the particular graphemes in the surrounding context.

Glushko (1979) first highlighted the importance of spelling-sound consistency in reading beyond the binary regularity of individual grapheme-phoneme correspondences. Distinguishing between “regular” words that can be pronounced using GPCs, and “exception” words that cannot, he found that adults pronounced nonwords (i.e. novel pseudo-word items) derived from regular words (e.g. *taze*, derived from *maze*) more quickly than those derived from exception words (e.g. *tave*, derived from *have*). Further, these exception nonwords were sometimes pronounced to rhyme with their corresponding real exception words, rather than by using GPCs. Glushko stated that these results could not be accounted for solely by an abstract system of pronunciation rules involving individual graphemes and phonemes (i.e. GPCs). Instead, he suggested that readers use existing knowledge of known words to inform pronunciations of new words by a process of analogy, or by use of multi-letter spelling patterns. Further, he argued that classifications of "regular" and "exception" words should be replaced by a system that can encode the consistency of orthography-phonology relationships.

1.1.2 The word body

Following Glushko's suggestion that readers' nonword pronunciations demonstrate orthography-phonology knowledge beyond the regularity of individual graphemes and phonemes, much research has investigated whether readers also use knowledge of spelling-sound consistency, particularly that involving larger orthographic units. These larger units can encompass the context-sensitive patterns described above involving multi-letter sequences. In particular, much attention has been paid to the orthographic unit combining the vowel and final consonant(s) of a monosyllabic word, termed the *word body* (including but not limited to Kay & Bishop, 1987; Jared et al., 1990; Treiman et al., 1995; Jared, 1997; Ziegler et al., 2001; Jared, 2002). For instance, a series of word-reading experiments by Jared (2002) found that word body consistency predicted naming latencies more successfully than GPC regularity, providing evidence that adult readers make use of orthographic bodies and are sensitive to spelling-sound consistency. Therefore, it is possible that word bodies provide additional information to readers by offering increased spelling-sound consistency beyond simpler but less reliable GPCs. However, it should also be noted that word bodies can also have inconsistent pronunciations (Vousden, 2008), for example the pronunciation of the body "all" heard in the words "ball" - /bɔːl/ and "shall" - /ʃæl/.

Importantly, this is not to say that the orthographic unit of the word body in particular holds any abstract significance for readers *per se*. Instead, it is simply that inconsistent vowels are more strongly conditioned by the coda (i.e. the following consonant(s), which together with the vowel constitutes the word body) than by the onset (i.e. the preceding consonant(s)) in English words (Treiman et al., 1995; Kessler & Treiman, 2001). For instance, Treiman et al.

(2003) demonstrated that readers are in fact also sensitive to onset-vowel pronunciation contingencies, but only in localised and relatively rare situations, concluding that readers are able to take into account orthographic units beyond the body when pronouncing the vowel. The focus on the orthographic body in the reading literature is therefore likely to be the result of statistical patterns of pronunciation contingencies in English words, rather than because readers are unable to learn a wider variety of complex spelling-sound correspondences.

1.2 Models of word reading

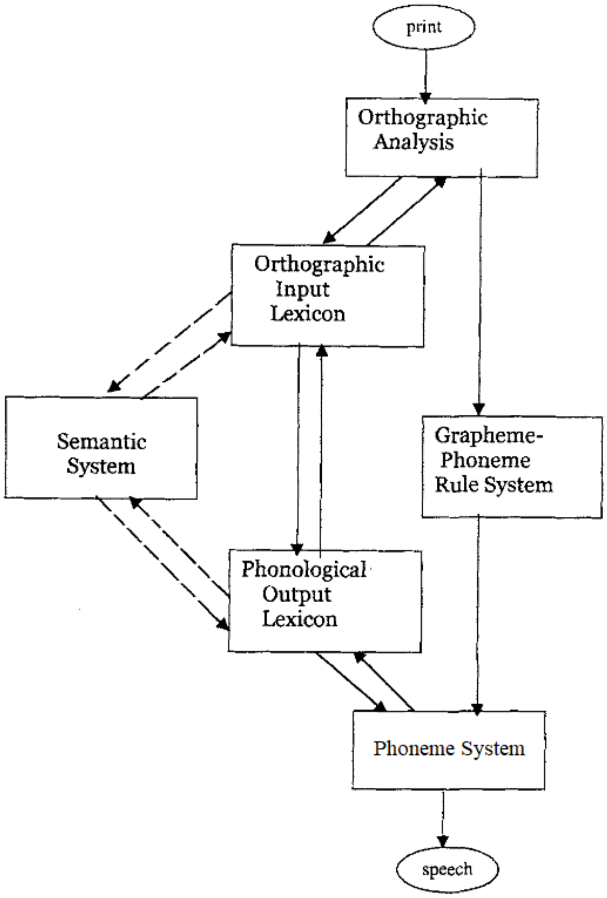
Theoretical accounts of word reading have mapped out different ways of capturing the correspondences between spelling and sound, and subsequently how readers pronounce words which can be accurately read aloud using GPCs, words which cannot, and novel words. These accounts tend to stem from two camps. Rule-based approaches categorise the *regularity* of spelling-sound correspondences, by defining a set of pronunciation rules which capture the most frequent (i.e. regular) correspondences between orthographic and phonological units - typically between individual graphemes and phonemes. Regular pronunciations can be generated using these regular correspondences; irregular pronunciations (i.e. those not using GPCs) must be derived through a separate process (e.g. Coltheart et al., 2001).

In contrast, statistical approaches do not categorically distinguish between regular and irregular pronunciations, but can encode the graded *consistency* of spelling-sound correspondences, i.e. how often a pronunciation is used for a particular spelling pattern. These spelling patterns can include longer sequences of letters (i.e. larger orthographic units such as the word body). These statistical approaches have been developed into full computational models, often using connectionist network architecture (e.g. Seidenberg & McClelland, 1989; Plaut et al., 1996; Harm & Seidenberg, 2004). These networks allow for the *context-sensitivity* of spelling-sound correspondences, as they maintain that readers are sensitive to the statistical probability of pronunciation patterns that are associated with particular orthographic contexts in their text input. For decades, research has investigated which approach (rule-based or statistical) most successfully captures human reading behaviour, and a variety of computational models working within these broad frameworks have been developed to this end.

The dominant computational model working from a rule-based approach is the Dual-Route Cascaded model (Coltheart et al., 1993; Coltheart et al., 2001). According to this model, there are two possible routes involved in generating the pronunciation of a letter string. These are a rule-based, nonlexical route, and a lexical route that involves retrieving information about the whole word from the lexicon (see Figure 1.1). The nonlexical route uses a set of the most common grapheme-phoneme correspondences in English (GPCs) to read regular words and nonwords accurately. For example, the grapheme “a” most frequently corresponds to the phoneme /æ/ in English words, so this is considered the regular pronunciation and can be assembled with other GPCs to pronounce regular words such as “mat” - /mæt/. Words that are not pronounced using regular GPCs, such as “what” - /wɒt/, are categorised as exception words, and must be pronounced via a separate system. For this process, the DRC uses an associative, lexical route that runs parallel to the nonlexical route. The lexical route is based on architecture from the McClelland and Rumelhart (1981) interactive activation model, in which lexical items are represented as nodes within a network. As the nodes correspond to individual, higher-order units (i.e. words), it is classed as a *localist* network (McClelland & Rumelhart, 1981).

Figure 1.1

The Basic Architecture of the Dual-Route Cascaded Model


print
Orthographic Analysis
Orthographic Input Lexicon
Semantic System
Grapheme-Phoneme Rule System
Phonological Output Lexicon
Phoneme System
speech

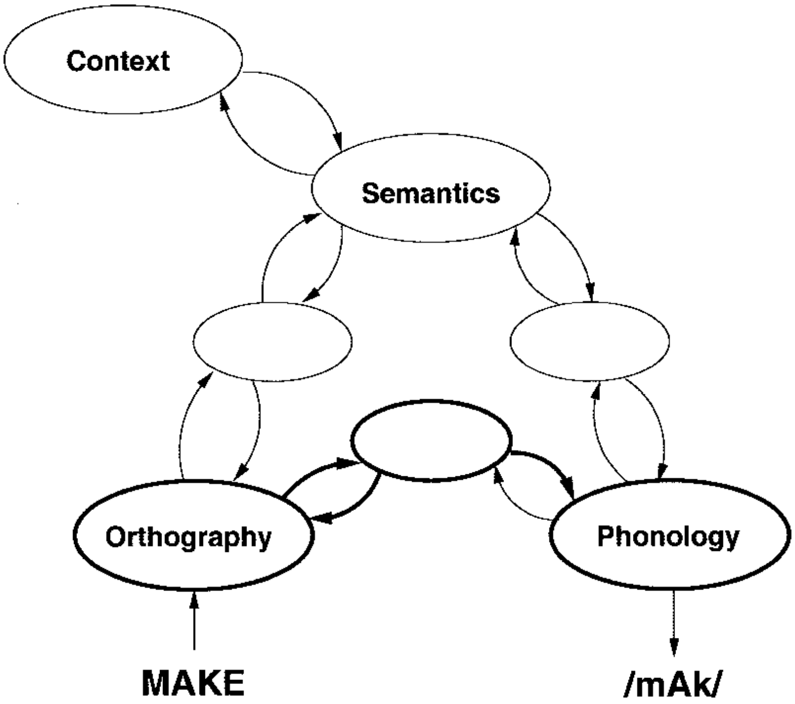
Note. Aadapted from Coltheart et al., 2001, p. 213. Note that the semantic system is unimplemented.

An early statistical model was developed by Seidenberg and McClelland (1989), the first in a series of models using the “Triangle” framework (subsequent versions have been developed by Plaut et al., (1996) and Harm and Seidenberg (2004) amongst others). As a connectionist network, it is built on the assumption that knowledge, such as that of spelling-sound correspondences, is represented by weights on connections that link processing units. Because the representation of lexical information is spread over sets of individual, sublexical units (which can overlap with those used by other lexical items), the Triangle model is known as a *parallel* *distributed processing* (PDP) model. This is in contrast to the *localist* representations of the DRC's lexical route in which nodes correspond to individual words.

The Triangle framework has two pathways by which to reach the pronunciation of a written form: a direct orthography-phonology pathway, and an indirect orthography-semantics-phonology pathway (see Figure 1.2), although only the orthography-phonology pathway was implemented by Seidenberg and McClelland in the original version of the model. The orthography-phonology pathway has a three-layer neural network: connections between a layer of phonological units and a layer of orthographic units are mediated by a layer of hidden units. These hidden units allow the network to capture more complex spelling-sound mappings, such as body-rime correspondences, by developing a layer of abstracted information which emerges through the partial activation of units in the hidden layer shared by similar words (Plaut et al., 1996).

Figure 1.2

Seidenberg and McClelland's Triangle Model (1989, p. 526)



Note. Copyright 1989 by the American Psychological Association. Only the orthography-phonology pathway (in bold) was implemented by Seidenberg and McClelland.

Besner et al. (1990) identified issues with the Seidenberg and McClelland (1989) model regarding its ability to generalise to novel forms. These issues were addressed in subsequent versions of the model, including those of Plaut et al. (1996) and Harm and Seidenberg (2004), in which the second orthography-semantics-phonology pathway was fully implemented.

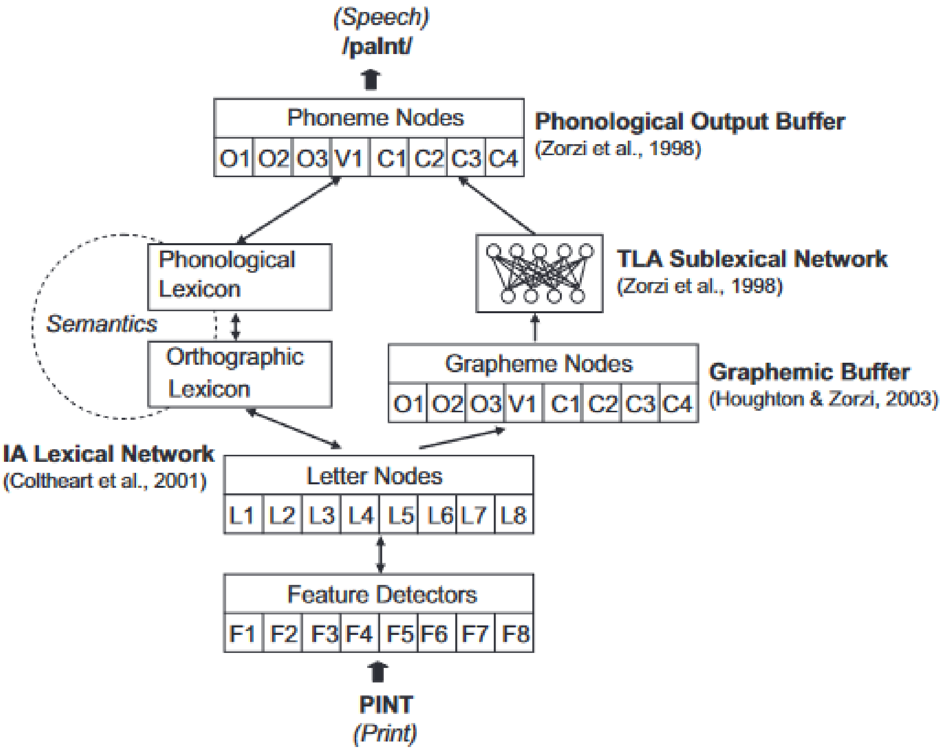
Although the Triangle models implement separate orthography-phonology and orthography-semantics pathways, they are sometimes still categorised (somewhat counter-intuitively) as *single-route models* through their use of a homogeneous processing mechanism for both pathways: namely, the spread of activation by weighted connections across distributed representations (Harm & Seidenberg 2004, p. 8). In other words, all orthographic, phonological and semantic knowledge is encoded in a single network, even though these types of knowledge can involve separate pathways through the network. This approach stands in contrast to the dual-route architecture of the DRC, where the pronunciation of a letter string is reached through two distinct types of processing mechanisms (i.e. the rule-based non-lexical route and the interactive-activation lexical route, as described above).

More recently, hybrid models of reading have been developed which use a dual-route processing system with separate lexical and non-lexical mechanisms, but which maintain aspects of distributed-connectionist processing architecture. These hybrid models include the Connectionist Dual Processing (CDP) model (Zorzi et al., 1998) and later the CDP+ model (Perry et al., 2007) and the CDP++ model (Perry et al., 2010), which was extended to include disyllabic words. The lexical system of the CDP+ model connects orthography and phonology in an interactive activation network based on the McClelland and Rumelhart (1981) architecture and similar to that of the DRC[2]. The non-lexical system maps orthography to phonology in a two-layer associative network (Ziegler et al., 2014). This non-lexical route is sensitive to sequences of graphemes which frequently occur together, meaning that it can capture the consistency of the spelling-sound relationship of multiple orthographic levels, or *grain sizes*, including the word body (Perry et al., 2010). Competing codes from these two routes interact in a phonological output buffer to produce the final pronunciation (see Figure 1.3).

[2] The lexical route of the CDP model (Zorzi et al., 1998) was not fully implemented but involved the activation of the phonological word form corresponding to a lexical entry and the spread of this activation to phoneme output nodes.

Figure 1.3

Perry et al.'s (2007) Schematic Description of the CDP+ Model



Note. O = onset; V = vowel; C = coda; TLA = two-layer assembly; IA = interactive activation, L = letter; F = feature (Perry et al., 2007, p. 280).

1.2.1 Frequency

An important distinction between rule-based and statistical accounts involves frequency counts. Rule-based models tend to use *type* frequency to determine the regular pronunciation of a grapheme: this means they use the absolute number of different word types in a corpus to measure which is the most frequent pronunciation of a grapheme. For this, the DRC's sublexical route uses type frequencies from the CELEX corpus (Baayen et al., 1995), which is a database of 7991 monosyllabic words. Meanwhile, statistical models also take the *token* frequency of words into account (Andrews & Scarratt, 1998), which means that the relative frequency of words in the input can also affect the patterns that the models produce. Specifically, distributed-connectionist architecture is based on the assumption that connection weights between units are strengthened through use, so the frequency of a word's occurrence during training affects the weighting of spelling-sound correspondences. Harm and Seidenberg's (2004) Triangle model is trained on a corpus of words presented to the model with the log frequency of the Francis and Kucera (1982) word norms. This determines the patterns of activation in the orthography-phonology route. The CDP+ model (Perry et al., 2007) uses the CELEX database as the training corpus. The lexical route is similar to that of the DRC, although it uses phonological rather than orthographic frequencies. The sublexical route was pre-trained on GPCs to simulate explicit phonics instruction, before being trained on items from the word corpus using normalised logarithmic frequency values, thereby weighting the type frequency counts by tokens.

1.2.2 Development of orthography-phonology knowledge

It is worth noting that as connectionist models build their representations during a training phase, they can simulate a knowledge-building process potentially akin to that of human readers. For example, Powell et al. (2006) explored specific modifications that could be made to the Plaut et al. (1996) connectionist model to bring it closer in line with children's literacy learning environment, which as a result improved the performance of the original network. These modifications included an incremental training regime and use of a training corpus based on words from children's early reading materials (see further discussion in Chapter 3). More recently, Chang et al. (2020) investigated the effects of prior knowledge and training on the learning trajectory and performance of the Triangle model (the Chang & Monaghan, 2019 version), with the aim of further understanding the links between pre-literate oral language and the specific focus of reading instruction on children's reading development. They used a vocabulary of English monosyllabic words as the training set in order to approximate children's literacy learning, and varied the initial oral language skills of the model by implementing 3 different levels of exposure to training items. Following this, the model was trained to read items from the training set through either an orthography-phonology (OP) focused or orthography-semantics (OS) focused regime. Results showed that the OP focused training model performed better on a reading aloud task than the OS focused training model. Further, there was an effect of prior oral language skills on both the reading aloud task and a written comprehension task; the effect was greater for the OS than the OP focused training model. The authors therefore conclude that poor oral language impacts reading comprehension more than it does reading aloud.

Studies such as these demonstrate the rich potential of the Triangle model to be adapted in line with children's experience of reading development, and indeed the insights that can be gained through this process. Meanwhile, GPCs in the DRC's nonlexical route are pre-set according to the corpus-based type frequency of pronunciations in English words. Therefore, this model cannot directly reflect a realistic learning process.

1.3 Nonword reading aloud

A common way to assess a computational model's ability to capture word reading and generalisation of spelling-sound knowledge is by running simulations which generate their predicted pronunciations for nonwords, and comparing these predictions with pronunciations produced by human readers for these nonword items. Participants' responses can also be analysed to reveal whether their pronunciation of each novel lexical item uses only individual GPCs (i.e., a "regular" pronunciation), or instead demonstrates knowledge of larger orthographic units (e.g. word bodies) or similar known words. These types of responses would suggest that readers are using the *consistency* of spelling-sound correspondences to inform their pronunciations rather than simply using the most frequent (or *regular*) pronunciation of each grapheme. Pronunciations of vowel graphemes in monosyllabic nonwords are particularly informative, as vowel graphemes can often be pronounced in a variety of ways in English orthography. Broadly speaking, support for a statistical account would be provided by a response that demonstrates an effect of word body consistency or consonantal context on the pronunciation of the vowel, for example pronouncing "pook" /pʊk/ to rhyme with "look". Meanwhile, a rule-based account would be supported by a response which uses the most common pronunciation of the vowel grapheme in isolation, for example "pook" - /pu:k/.

A significant body of work has collected and analysed human readers' nonword pronunciations in order to assess computational reading models in this way. For the purposes of the current investigation, these studies provide useful evidence about the way readers generalise in a quasi-regular system, and how best to capture this behaviour. For instance, Seidenberg et al. (1994) compared adult nonword pronunciations with those generated by the DRC model (Coltheart et al., 1993) and the Triangle (PDP) model (Plaut et al., 1993). They found that the Triangle model was able to predict the pronunciation produced most often by participants slightly more closely than the DRC, but concluded that both models are able to produce plausible nonword pronunciations. They suggested that adding context-sensitive rules may increase the success of the DRC, particularly by incorporating assumptions such as relative strengths or conflicts between rules. However, Seidenberg et al. did not elucidate how this could be achieved[3]. Similar themes regarding a precise hierarchy of increasingly specific rules will be developed in the current thesis. Additionally, Seidenberg et al. noted that spelling-sound consistency is likely to be the locus of pronunciation variability observed between participants; greater inconsistency gives rise to higher participant variability. This result suggests that both consistency and variability are variables which should be explored in order to provide a comprehensive understanding of human reading behaviour.

Andrews and Scarratt (1998) also carried out a nonword reading study to evaluate the competing predictions of Coltheart et al.'s (1993) DRC model and Plaut et al.'s (1996) Triangle model. Adults' nonword pronunciations were assessed according to their use of "regular" (GPC) or "analogy" (context-sensitive) strategies, and also their match against model predictions. Results revealed that participants used regular pronunciations for the majority of nonword responses, except for items with word bodies that are never pronounced regularly in English monosyllabic words, such as "beart" (i.e. they are *consistently irregular*). For these items, adults used context-sensitive pronunciations (by analogy to word neighbours with irregular bodies) approximately half of the time, and regular pronunciations less than a third of the time. Overall, the DRC model was better able to predict readers' responses than the Triangle model, although it overestimated the number of regular pronunciations participants produced, and could not predict their irregular pronunciations of nonword items. In fact, the best predictor of a regular pronunciation response was the number of regular word neighbours counted by type rather than token (contradicting the Triangle model's approach, which uses the token frequency of neighbours to determine the probability of an irregular pronunciation). This result suggests that

[3] Rastle and Coltheart (1999, Appendix B) later developed a set of context-sensitive rules for the DRC.

consistency is an important variable in the generalisation of spelling-sound correspondences, but that it should be based on type rather than token frequency counts. It is possible that a more successful account of reading will involve such an approach to consistency, and this will indeed be explored in the current thesis. In terms of the variability of pronunciations across participants, the authors found that items with inconsistent bodies produced more varied pronunciations than those with more consistent bodies, mirroring a similar finding by Seidenberg et al. (1994).

Treiman et al. (2003) examined the effect of consonantal context on vowel pronunciations in nonwords, comparing adults' pronunciations with those produced by a variety of computational models. They selected vowel graphemes with inconsistent pronunciations, such as "ea" which has a GPC pronunciation in words such as *cheap* and a context-sensitive pronunciation in words such as *head.* These vowel graphemes were used in nonword items, in both a control consonantal context in which the GPC pronunciation would be expected (e.g. *cleam*, as in *cheap*) and a critical consonantal context, in which a context-sensitive pronunciation may be expected (e.g. *clead*, as in *head*). They found that participants produced more context-sensitive pronunciations of the vowel in critical than control consonantal contexts, suggesting that adult readers were using knowledge of context-sensitive pronunciation patterns (involving the word body) in their responses. However, the authors noted that the context-sensitive pronunciations were used less often in critical contexts than would be predicted by the statistical frequency of these pronunciations in English words (see further discussion in Chapter 6). Their assessment of rule-based and connectionist models including those of Coltheart et al. (2001), Zorzi et al. (1998), Plaut et al. (1996), Plaut and McClelland (1993), Powell et al. (2001), Harm and Seidenberg (2002)[4], and Norris (1994) found that none of the models matched adults' pronunciations very well, particularly for nonwords with critical consonantal contexts. For instance, for critical nonwords such as *squant* (which may be pronounced to rhyme with *font* or *rant*), the most successful model (Norris, 1994) matched the participants' most common pronunciation only 68% of the time; for the least successful model (Coltheart et al. 2001) this was 38%. Treiman et al. suggested that an approach which combines information about the frequency of orthography-phonology mappings in different contexts may result in a more

[4] Treiman et al. refer to the paper submitted for publication by Harm and Seidenberg (2002); see also the published model (Harm & Seidenberg, 2004).

successful account of human reading behaviour. They did not specify the details of how such an approach could be developed, but one possible account will be explored in the current thesis.

Pritchard et al. (2012) compared nonword pronunciations produced by adult readers with those generated by the DRC (Coltheart et al., 2001) and CDP+ (Perry et al., 2007). Both of these models feature a dual-route architecture as discussed above, but use rule-based and connectionist non-lexical routes respectively, which can produce different nonword pronunciations. Results showed that adult participants used regular (GPC) pronunciations most often in their nonword responses, but that they did produce some alternative, irregular pronunciations (although not at the rate they occur statistically in English words). This discrepancy between the rate of alternative pronunciations produced by participants and their frequency in English vocabulary was also observed by Treiman et al. (2003). Together, these findings suggest that irrespective of whether we categorise possible pronunciations according to regularity (as did Pritchard et al.) or context-sensitivity (as did Treiman et al.), readers do not simply reproduce the statistical distributions of their input in their pronunciations.

Pritchard et al. (2012) reported that the DRC was more successful than the CDP+ in matching the most frequent human pronunciation of each item, although neither model fared particularly well: the DRC predicted too many regular responses compared to the participants, and the CDP+, too many lexicalisations (i.e. pronouncing the nonword as an existing real word) and other irregular responses. The authors suggested that a rule-based model allowing multiple rules of different strengths to apply to a grapheme depending on the context, or a mechanism for rules involving larger units such as word bodies to override GPCs, could predict pronunciations more accurately. These suggestions were not developed further by the authors, but will be addressed in the current thesis. They also highlighted the wide variety of pronunciations produced by participants, noting that a successful model of reading should be able to account for such differences between individual readers. Again, these findings suggest that extant models do not successfully capture readers' productive use of spelling-sound correspondences, and that a more specific mechanism for predicting use of alternative pronunciations in different contexts is required.

Beyond this body of work on monosyllabic nonword reading, Mousikou et al. (2017) conducted a mega-study of disyllabic nonword reading with adults, and evaluated two competing models: the CDP++ model (Perry et al. 2010) and the rule-based disyllabic algorithm of Rastle and Coltheart (2000). Analysis of the similarity between each participant and the models - treating each model as an individual participant - revealed that neither model behaved typically, suggesting that neither a rule-based nor a statistical approach accounts well for the human reading data. Additionally, participants' pronunciation variability was predicted by the spelling-sound consistency of both the first and second syllables. The authors suggested this demonstrates that participants are sensitive to statistical patterns in the lexicon, and thereby concluded that their results lend more support to a statistical-learning approach overall. Again, these findings indicate that participants neither regularise all pronunciations, nor match predictions based on the statistical distribution of the input. Instead, a more nuanced account, which also allows variability between participants, is required and will be explored in the following chapters.

This range of studies investigating nonword reading behaviour reveals similar patterns of findings, although these results can be construed in different ways. For instance, many of the studies found that to pronounce nonwords, adult readers use regular pronunciations of individual graphemes (GPCs) most often, a result which can be used as evidence in support of a rule-based account. However, most also report that readers make some use of alternative pronunciation patterns; behaviour which can be characterised as demonstrating context-sensitivity or knowledge of larger orthographic grain-sizes (such as the word body). These results suggest that readers are sensitive to the consistency of orthography-phonology mappings; this finding supports a statistical account and cannot easily be accounted for by a rule-based account such as the DRC. Further, numerous studies found that the variability in pronunciation responses between participants is predicted by spelling-sound consistency, which again suggests that readers demonstrate sensitivity to the consistency of certain patterns in their input. However, several researchers have also noted that readers do not simply reproduce the statistical distribution of pronunciation patterns in the lexicon; rather, their use of context-sensitive pronunciations is often notably lower than would be predicted by corpus statistics. Taking these results together, it seems clear that a successful account of word and nonword reading must be able to explain precisely how readers make use of the input statistics they have been exposed to in their generalisations, including the instances in which readers do maximise use of the most common GPCs.

It is worth noting that whilst researchers often highlight ways in which models fail to capture human reading behaviour and make suggestions for improvement (e.g. Seidenberg et al., 1994; Treiman et al., 2003; Pritchard et al., 2012, Mousikou et al., 2017), they rarely develop a detailed account of precisely how this could be achieved. In this thesis, one possible solution that makes quantitative predictions on the basis of input statistics will be applied to this issue and explored in detail.

1.4 Unresolved issues and the next steps

Overall, the research discussed above suggests that extant models of reading are unable to fully or precisely capture human nonword reading behaviour: rule-based models are able to successfully predict the majority of pronunciations which use the most common grapheme-phoneme mappings, but cannot account for those instances in which readers use alternative pronunciation patterns. Meanwhile, statistical models are to able predict context-sensitive pronunciations (e.g. those involving the word body), but often do so more frequently than human readers actually produce them. Therefore, there seems to exist a gap in these models' capacity to capture readers' behaviour, in which readers' nonword responses are either less regular or less context-sensitive than each approach would predict. More broadly, important questions remain regarding the significance of graphemes versus word bodies, and the conflict between the use of rules or statistics to characterise generalisation of spelling-sound knowledge. What is required is an approach that can capture the effects of both regularity and consistency, perhaps by taking into account the statistical properties of orthography-phonology correspondences in English words. To be successful, it should also be able to predict when information from different orthographic grain sizes or contexts will be used productively by readers, and when pronunciations may vary across readers.

In this thesis, I will investigate whether the Tolerance Principle (Yang, 2016) can be applied to reading in order to fulfil these requirements and to address the related long-standing questions in the word reading literature. As will be explored in the following chapters, the Tolerance Principle is a rule-based account which uses statistical information to form categorical predictions for generalisation. In this way, it is able to capture both the regularity and consistency of a pattern, and make predictions about its productivity. I will investigate whether the Tolerance Principle can shed light on decades-old debate in the field of reading such as readers’ use of either graphemes or word bodies in nonword reading aloud; whether orthography-phonology knowledge is better characterised by rules or statistics; and precisely how readers make use of statistical information in their text input to form generalisations.

Chapter 2: Introducing the Tolerance Principle

2.1 The power of productivity

As introduced in Chapter 1, a central property of human cognition is productivity: we can generalise beyond knowledge gleaned from past experience to produce and understand an infinite array of structures. In the previous chapter, this property was discussed within the context of reading, specifically regarding a decades-old undertaking to understand the ways in which readers generalise spelling-sound correspondences to pronounce unfamiliar written items. Despite a large body of research, several outstanding issues remain. In what instances are readers either categorical or sensitive to graded consistency in their generalisations? When do readers use more general context-independent or more specific context-dependent spelling-sound correspondences productively? Does the type or token frequency of items in our reading experience determine the way we learn and use pronunciation patterns? What is the best way to characterise and model human reading behaviour overall?

Meanwhile, productivity is also strikingly apparent in child spoken language acquisition: children do not simply memorise and reproduce the sentences they have been exposed to, but can produce novel linguistic structures by abstracting information from their input and applying the patterns they have identified to new situations.[5] For example, after learning that many verbs in English form the past tense by adding *–ed*, by the age of three children can apply this pattern to other verbs that they have not heard in the past tense before (Marcus et al., 1992). Experimentally, Berko (1958) demonstrated that children aged five could apply the *-ed* past tense morpheme when presented with novel verbs, e.g. *gling – glinged*, suggesting that they are able to generalise this grammatical knowledge.

Running parallel to the debate between rule-based vs. statistical models of reading discussed in the previous chapter, there have been similar developments in research setting out to capture generalisation in spoken language. Much of this research on productivity, particularly in formal linguistics, has focused on identifying abstract *productive rules*: a regular pattern that

[5] It should be noted that accounts of language acquisition differ in the stage at which they attribute combinatorial productivity to the language of developing speakers. For example, the usage-based account (Tomasello, 2000) emphasises the role of memorisation and maintains that children's earliest multi-word utterances are instantiations of item-based schemas rather than systematic rules.

extends over a majority of forms and can be applied to novel items. Evidence of generalisation from studies such as Berko's (1958), as well as the developmental trajectory of the English past tense (discussed further below), have traditionally been used to support such rule-based accounts (e.g. Pinker, 1999; Plunkett, 1991). Alternative accounts have subsequently been proposed which instead highlight the role of input statistics, including distributed-connectionist models (e.g. Rumelhart & McClelland,1986) which invoke domain-general, probabilistic learning mechanisms. They maintain that children learn to form the past tense of a verb, for example, through a pattern association mechanism using the statistical distributions of their input, rather than by forming and applying abstract rules.

However, no matter how productivity is characterised, there remains an elephant in the room: whether we are dealing with spoken or written language systems, we find exceptions which do not conform to the majority pattern. To take another example from English past-tense morphology (which has been the subject of a large amount of research on productivity and rule-learning), the past tense form of *swim* is *swam*, not *swimmed.* As seen in the field of reading, the presence of exceptions amongst regular patterns (i.e., a quasi-regular system) has prompted swathes of research asking how regular versus exceptional grammatical forms are learned, accessed and generalised.

In the current chapter I will introduce the Tolerance Principle (Yang, 2016), a recently-proposed account of generalisation in spoken language acquisition. This theory sets out to address long-standing issues including how learners extract useful information from their input, how they overcome exceptions to form productive patterns within quasi-regular systems, and why they restrict generalisation in certain instances. The TP is a rule-based approach which incorporates statistical information, enabling it to make categorical predictions about the use of productive rules on the basis of a type-based consistency metric. After examining the ways in which this novel approach claims to resolve the range of issues surrounding spoken language acquisition outlined above, I will consider its application to reading and lay out my intentions to explore its applicability for solving similar outstanding questions in this field.

2.2 Rules and exceptions

2.2.1 Rule-based models

Children's novel generalisation of plural and past tense forms in Berko's (1958) famous study has traditionally been employed to support the existence of productive grammatical rules. For instance, children's ability to apply the past tense *–ed* pattern to unfamiliar verbs has been used as evidence that they operate an abstract rule-based mechanism, formalised as something like "add *–(e)d*", in order to produce the past tense of a regular verb (e.g. Pinker, 1989). In some rule-based theories, irregulars are memorised in word pairs such as *sleep-slept*, whilst others exclusively use a system of rules to generate both regular and irregular forms.

For example, according to Chomsky and Halle's (1968) Sound Patterns of English, the past tense form of a verb is created by using either the *–ed* rule or a small selection of additional minor rules; in Halle and Marantz's (1993) Distributed Morphology theory, the past tense form of a verb is generated according to either the regular allomorph or a set of "readjustment rules" (Halle, 1990). According to Kiparsky's (1982) Lexical Phonology and Morphology, verb inflections are derived through a series of levels each associated with a set of phonological rules, whilst the Yip-Sussman model (Yip & Sussman, 1997) uses an inductive learning process which creates a default rule plus phonologically specified rules. However, the dominance of rule-based approaches such as these was challenged when a new, associative model of the past tense was developed using connectionist parallel distributed processing architecture.

2.2.2 Connectionist models

In distributed-connectionist models, knowledge is represented by patterns of activity across a network of processing units which represent the statistical structure of the input. These patterns are not distinguished according to regularity[6]; both the majority and alternative patterns are captured within a single network, without the need for separate mechanisms for rules and exceptions (Joanisse & McClelland, 2015). These patterns can also be used to generate novel forms. For example, Rumelhart and McClelland (1986) designed a revolutionary model which

[6] Connectionist models do not invoke abstract rules so a distinction between regular and irregular patterns is not applicable.

could learn the mapping between the present and past tense form of a verb using a learning algorithm which adjusts the patterns of activation according to exposure to those forms. The network demonstrated successful acquisition of both regular and irregular forms, and also produced generalisations to novel forms, all within a single architecture. However, as noted by Joanisse and McClelland, this generalisation does not always accurately match human behaviour; a common criticism of connectionist models is that they overgeneralise irregular forms (e.g. producing the past tense *glang* for the novel verb *gling*) much more often than children do (Marcus 1995; Yang, 2016; Schuler et al., 2021). For instance, the Rumelhart and McClelland model developed sensitivity to a subset of irregular verbs which end in word-final *–d* or *–t* and do not mark a change between present and past tense (e.g. *cut – cut*); the model incorrectly overgeneralised this no-change pattern to regular verbs and other irregulars ending in *–d* or *–t* (Marcus et al., 1992). Indeed, in contrast to the rule-based model that will form the basis of investigation in this thesis, connectionist theories maintain that “there is no dichotomous distinction between productive and unproductive phenomena; rather, there are only degrees of productivity” (McClelland & Bybee, 2007, p. 439).

2.2.3 Dual-route models

In response to the development of single-route connectionist models there emerged a new approach which incorporated elements of both previous sides of the past-tense debate: dual-route models (Pinker & Prince, 1988; Prasada & Pinker, 1993; Pinker & Ullman, 2002). These models proposed that regular inflections are generated using productive rules, whilst irregular forms are memorised and accessed through a separate associative system. In contrast to connectionist models, they maintain that a qualitative distinction between regular and irregular forms is necessary to account for a range of behavioural findings. For instance, Prasada and Pinker (1993) investigated adults’ willingness to generalise regular (*walk-walked*) and irregular (*swing-swung*) past tense patterns to novel verbs. They found that participants’ willingness to generalise from known irregular verbs to novel verbs (e.g. *spling-splung*) depended on the similarity between the verb forms. This was not the case for regular verbs, where the rating and production of generalisations was not associated with similarity between the existing and novel forms. These results replicated similar findings from Bybee and Moder (1983). In contrast, simulations from

Rulmelhart and McClelland’s (1986) connectionist model demonstrated generalisation of both regular and irregular patterns to novel forms as a function of similarity to trained forms. Therefore, Prasada and Pinker suggested that their results did not support a single-network theory such as Rumelhart and McClelland’s; neither did they support a rule-only theory which cannot account for the observed patterns of irregular generalisations. Instead, they proposed a hybrid model in which regular inflections are generated by abstract rule, and irregular forms are stored separately within an associative memory system and can be generalised by a process of analogy.

2.2.4 A return to a rule-based approach?

Dual-route models embrace aspects of both rule-based and associative approaches, and thus offer considerable explanatory power. For instance, they account for frequency and similarity effects in irregular but not regular past tense verb productions (Prasada & Pinker, 1993), and for difficulty producing irregular but not regular forms in children with Specific Language Impairment (Gopnik, 1990). However, there is some evidence that dual-route models do not always capture behavioural findings. Instead, there has been a return by some researchers to rule-based approaches that involve a hierarchy of rules (e.g. Albright & Hayes, 2003; Ambridge, 2010). For instance, Ambridge (2010) carried out an acceptability judgement task of the English past tense with children aged 6-7 and 9-10. Participants rated the acceptability of novel verbs using either the regular or an irregular past tense form. The novel verbs differed in their phonological similarity to real regular and irregular past tense forms, for example *nace* (similar to an existing class of regulars inclduing *race* - *raced*); *fleep* (similar to an existing class of irregulars including *sleep* - *slept*); and *gude* (not similar to either an existing class of regulars or irregulars). Results found that the acceptability of novel irregulars increased with similarity to existing irregular forms, with no interaction with age of the participant. For novel regulars, acceptability increased with similarity to existing regular forms for the older but not the younger age group. Ambridge proposed that the developmental effect between age groups was observed for regulars but not irregulars because irregular forms are acquired early and would be known by all participants, whilst the younger children may still be acquiring knowledge of regular forms. Overall, Ambridge (2010) suggested that these findings support models which either allow generalisation by analogy to operate across stored regulars, such as single-route models (Bybee & Moder 1983), or those that involve a hierarchy of increasingly specific rules, such as multiple-rule models (Albright & Hayes, 2003). He argued that the findings do not support dual-route models (e.g. Prasada & Pinker 1993), unless the models allow for a substantial effect of analogical generalisation over stored regular forms. However, these results diverge from those of Prasada and Pinker's (1993) study, where no function of similarity was observed for generalisation of the regular pattern. It is possible that acceptability judgement tasks are not reliable assessments of generalisation, and instead production tasks may be more valuable in order to discriminate between processing models.

2.3 When does a rule become productive?

Beyond this discussion of the most successful way to model processing of rules and exceptions is the additional (and perhaps more pertinent) consideration of precisely *when* a rule actually becomes productive. Not all linguistic patterns are productive; for example the *sing – sang*, *ring – rang* past tense forms do not extend beyond a restricted subset (Schuler et al., 2021). Such examples prompt the question of at what point should a pattern be generalised to novel instances, rather than being restricted only to items attested during linguistic experience.

Looking at evidence from child language data, it is well-attested that children overgeneralise regular patterns during the course of language acquisition (Marcus et al., 1992; Pinker, 1999); for example, producing *eated* rather than *ate*. Notably, the acquisition of English past tense forms follows a U-shaped curve (Ervin & Miller, 1963; Cazden, 1968; Pinker & Prince, 1988; Marcus et al., 1992), in which young children produce irregular forms correctly before entering a stage in which overregularisations are common, then subsequently approach adult-like behaviour. Additionally, Berko's (1958) Wug test demonstrated young children's readiness to generalise regular grammatical patterns to novel items. However, it is uncommon for children to overgeneralise *irregular* forms (i.e. apply an irregular pattern productively), and the occurrence of such errors has sometimes been overestimated by both rule-based and connectionist theories (Marcus, 1995; Xu & Pinker, 1995; O'Donnell, 2015). In fact, Yang (2016, p. 33) claims that irregular analogical errors are "almost completely anecdotal". This propensity to generalise regular but not irregular forms has in fact been demonstrated crosslinguistically (e.g. Clahsen & Penke, 1992; Allen, 1996, Caprin & Guasti, 2009, discussed further by Yang, 2016). Critically, if a categorical distinction between productive regular forms and unproductive irregular forms is a universal characteristic of children's linguistic behaviour, then it is crucial to understand which patterns qualify as productive, and at what stage, during the language acquisition process.

Some researchers do not go as far as setting out clear predictions about when a rule may become productive for a developing speaker. For instance, Marcus et al. (1992) rejected the possibility that children require a pattern to apply to the majority of tokens, or a relative or absolute number of types, in order for it to become productive. They did note the possibility that children perhaps require little input in order to establish a regular pattern, but the authors stated they "lack evidence that would allow us to identify which cues children actually use to acquire a regular rule" (1992, p. 133).

As highlighted by Schuler et al. (2021, p. 7), many approaches which do make predictions about productivity invoke the idea of "statistical dominance" to determine which patterns in the input will be extended. According to the individual approach, the majority form may be identified on the basis of type frequency (Bybee, 1995), the proportion of type frequency across a number of tokens (Baayen, 1989; Baayen & Lieber, 1991), or type counts weighted by token frequency (as in distributed-connectionist models, such as Rumelhart & McClelland, 1986). Other accounts predict which pattern will be generalised according to measures of efficiency or data optimisation (Taatgen & Anderson, 2002, O'Donnell, 2011; 2015). But as Schuler et al. (2021) note, none offer a precise prediction of productivity involving input data and an evaluation metric in a way that children could be expected to undertake during language acquisition.

2.3.1 Overcoming exceptions in other linguistic domains

There is evidence that children can overcome exceptions to extend certain patterns on the basis of type frequency in domains beyond grammatical generalisation. This research highlights the need for productivity metrics that can apply beyond the grammatical patterns that are usually described in the literature on rule-learning and generalisation - most often the English past tense.

Further, there is evidence to suggest that type frequency may not be the only determiner of generalisation in these contexts. For instance, Lazaridou-Chatzigoga et al. (2019) investigated adults' and 4-5-year-old children's generalisation of properties about novel objects. Experiment 1 found that both adults and children generalised striking properties (e.g. "play with fire") about novel objects (e.g. "glippets") less often than neutral properties (e.g. "play with toys"). This result suggests that generalisation of consistent patterns can be moderated by specific (and non-linguistic) properties of the items involved.

In their Experiment 2, varying numbers of exceptions to the properties of the objects were made: after being introduced to two instances of a novel object with a common property (e.g. "These are glippets. Glippets like to play with toys/fire"), either one or three more objects of the same kind were introduced without this common property (e.g. "This/these glippet(s) don't like to play with toys/fire"). Participants were then asked whether the property applied to new objects of the same kind. Both adults and children generalised properties to further objects at a lower rate than in Experiment 1, where there were no exceptions. Further, the greater the number of exceptions, the lower the rate of adults' generalisation; for children, this was only the case for striking properties. The authors suggest that for neutral properties, children may need exposure to more exemplars in order to demonstrate sensitivity to the number of exceptions. Indeed, the total number of exemplars of each object was very small; more exemplars may have allowed further generalisation patterns to emerge. Nevertheless, this study highlights the potential for further research investigating the generalisation of non-grammatical patterns in quasi-regular systems. Additionally, it suggests an interaction between the type frequency of exceptions and property salience in children's generalisations.

2.4 Computational efficiency

With many possible ways in which in the data in our input could be encoded to produce productive patterns, one possible factor steering this process is computational efficiency. Research in cognitive science has long explored how principles of efficiency may shape our behaviour, or even cognitive architecture. For example, Zipf's (1949) Principle of Least Effort proposes that human behaviour will expend the least amount of effort to accomplish a task; behaviours that are useful will be performed frequently, and thus become still more efficient to perform. The theory of Rational Analysis (Anderson, 1990) states that our cognition reflects the statistical structure of the input in a maximally efficient way. Other approaches emphasise the role of efficiency in the learning process, under the assumption that learning involves identifying patterns within data (Chater & Vityani, 2003). They maintain that in order to determine the optimal pattern that captures a dataset amongst an infinite number of possible patterns, the cognitive system will choose the simplest explanation of the data. Here, "simplicity" is defined according to the shortest description of the data (Mach, 1959; see also the Minimum Description Length principle (Rissanen, 1978) in the machine learning literature) and is even posited as a unifying principle across cognitive science (Chater & Vityani, 2003).

The Simplicity Principle (Chater & Vityani, 2003) has been applied specifically to language (Chater et al., 2015), where it states that the briefest representation of the linguistic data will be sought by the cognitive system. Others have argued that language learners are guided by an overarching simplicity bias, but that additional competing biases may be at play (Culbertson & Kirby, 2016). Alternatively, some have explored how Bayesian frameworks can capture the efficiency of language learning (Xu & Tenenbaum, 2007). Meanwhile, according to a usage-based, information theory view of language, communicative efficiency entails that messages be successfully transmitted between speakers with minimal effort (Gibson et al., 2019). Crucially, this definition acknowledges that language involves communication as well as computation; the efficiency of learners' encoding must not compromise successful communication between speakers. Gibson et al. note that a breadth of work suggests that insights into how language is optimised for processing, learning and communication can be derived by bringing together findings from linguistics, cognitive psychology, and mathematical and computational theories of inference and learning (e.g. Clark, 2001; Christansen & Chater, 2008; Jaeger & Tily, 2011; Fedzechkina et al., 2012). In short, an efficient language system requires cognitive effort to be minimised without communication being hampered.

Some studies explore this process experimentally, investigating how cognitive biases influence the way learners acquire patterns from their input. This process can include regularisation, whereby learners reduce the variability observed in their input by adopting a general rule that captures the majority pattern. In an artificial language learning study, Ferdinand, Kirby and Smith (2019) manipulated cognitive load, variability and task domain (linguistic vs.

non-linguistic) to investigate adult learners' regularisation, which was formalised as the reduction of entropy in the input dataset. They found that increased cognitive load for linguistic stimuli elicited regularisations in adults' productions (naming objects they had previously observed with inconsistent labels), but not in their explicit encoding of the input data (estimating the ratios they had observed). The authors conclude that linguistic regularisation is a result of both domain-general and domain-specific biases on both learning and production.

Considered within the context of the role computational efficiency plays in language, these findings indicate that there may be a distinction between the effect principles of efficiency have on either the learning or the generalisation of patterns in the input. For adults at least, whilst the most efficient *learning* of patterns from an exposure might involve reproducing variability in the input, *generalising* the information that has been gleaned to new instances might involve a more active process to be undertaken by the learner. This could include imposing structure on the input data by reducing variability (i.e. regularisation), which may take place under the pressure of both computational and communicatory constraints. Overall, it seems that our linguistic systems may be shaped in different ways by cognitive process that are driven by principles of efficiency, not only in terms of encoding the data to which we are exposed, but also in learning, producing and communicating this information.

2.5 Introducing the Tolerance Principle

A recent theory of language acquisition proposes a novel solution to a number of the issues raised above. Yang's (2016) Tolerance Principle offers an account of linguistic generalisation that makes quantitative predictions about rule productivity on the basis of input data and according to measures of computational efficiency. Crucially, it is a rule-based approach that incorporates statistical information about the consistency of a pattern, thereby offering a new middle ground between previous rule-based and statistical approaches to productivity.

As a theory of language acquisition and processing, the Tolerance Principle (TP) determines the productivity of a linguistic rules according to a quantitative balance between regulars and exceptions. By doing so, it aims to capture the way children are able to form productive rules on the basis of their linguistic input, whilst dealing with exceptions that do not conform to regular patterns. Specifically, the TP states that the learner postulates a productive rule only if it results in more efficient organisation of the language than listing every item in lexical storage. Thus, the number of exceptions to a rule must fall below a critical threshold for the rule to be productive. According to Yang (2016, p. 8-9), this threshold is calculated as follows[7] (see *Section 2.5.1.* for further details):

> If R is a productive rule applicable to *N* candidates, then the following relation holds between *N* and *e*, the number of exceptions that do not follow *R*:

$$e \leq \theta_N \text{ where } \theta_N := \frac{N}{\ln N}$$

Effectively, this threshold provides a categorical assessment of consistency: a rule that is consistent enough to pass this tolerance threshold is valuable enough to be used as a productive rule for novel items. Indeed, Yang defines the process of language acquisition itself as a search for productive generalisations, arguing that learners follow the general learning strategy "pursue rules that maximise productivity", termed the *Maximise Productivity* principle (2016, p. 72). This strategy characterises the TP as a mechanism by which learners are guided to extract productive rules from their linguistic input. Additionally, the TP allows for subtle differences between each individual's system of productive rules developed throughout their learning trajectory, as "the execution of the TP should be based on an individual learner's vocabulary" (2016, p. 70). This means that the balance between regular and irregular items for each learner is determined by the specific linguistic input they have directly received.

2.5.1 Computational efficiency and motivation for the tolerance threshold

The TP is motivated by computational efficiency in terms of real-time language processing. It provides a tipping point for productivity beyond which memorising every item,

[7] See Appendix A for the full derivation of the Tolerance Principle according to Yang (2016, p. 60 – 66).

including regular past tense forms like *walked* or irregular forms like *ran,* is a faster processing option than using a grammatical rule for the regular items, such as "add *–(e)d* to form the past tense of a verb", and storing irregular items separately. Specifically, it identifies a precise balance between storing all regular and irregular lexical forms individually in a frequency-ranked list and searching the list every time a form (such as the past tense) is needed, versus forming a productive rule ("add *–(e)d* ") for regular items (like *walked*) and storing only the exceptions (such as *ran*) in a frequency-ranked list. If the target is not among the list of exceptions, the learner applies the rule (2016, p. 9). This approach to the use and storage of rules and exceptions is based on the *Elsewhere Condition* (Anderson, 1969), according to which exceptions are handled by a more specific process than the general rule. The *Elsewhere Condition* is implemented here as a serial search procedure. When an item *w* is eligible for application of the rule, a frequency-ranked list of exceptions is first searched for a match to *w*. If a match is found amongst the exceptions, this form is used. If a match is not found, the rule is applied to item *w* (Yang 2016, p. 50). Critically, and as discussed at length by Yang (2016, p. 50-65), exceptions are searched prior to the application of the rule, and therefore a larger number of exceptions contributes to the rising cost of online processing.

In this way, the serial search procedure is central to the TP theory and crucial for determining the tolerance threshold, as the threshold itself is the point at which two alternative routes to access a target item take an equal amount of processing time. If the number of frequency-ranked exceptions to be rejected before applying the rule pushes the access time for a target item beyond that involved for identifying the target amongst full lexical listing, then the threshold is breached and the rule is not productive. Yang formalises the processing time for the exceptions-plus-rule route as *T(N,e)*. This represents the weighted average time for accessing the target when it is a regular item (which will take *e* search steps, as every exception must be assessed and rejected before applying the rule) and for accessing the target when it is an exception (which is determined by its position on the frequency-ranked list). Meanwhile, the expected time of access for the full-lexical-listing route is formalised as *T(N,N),* as all *N* items are stored in a frequency-ranked list. Therefore, the analytical solution to the equation *T(N,N) = T(N,e)* provides the number of exceptions that can be tolerated (see Appendix A for the closed-form solution provided by Yang.)

As noted by Yang (2016, p. 61), a greater number of exceptions (which must be searched through before the rule is applied) increases the time required to access a rule-following item, which may mean that a high-frequency regular could incur a relatively slow processing time. At a certain point this exception-plus-rule route will become less efficient than using a full lexical listing; Yang claims that the TP identifies this specific tipping point.

2.5.2 Type or token frequency?

Frequency is a central variable in the TP theory. The tolerance algorithm itself uses the type frequencies of *N* and *e* (2016, p. 67); that is how many different types of items do not follow a majority pattern (*e*) out of a set of (N) items. This stems from the position that rules become productive after the accumulation of evidence from the input and must be supported by a sufficiently large number of distinct types (2016, p. 67). According to Yang, the number of times each item is encountered in the input (the token frequency) does not directly affect acquisition of productive rules. However, this token (or summed) frequency does feature in the calculation of expected online processing time required for accessing items which is used in the underlying derivation of the TP. This time complexity is approximated by Zipf's law (1949), according to which the frequency of a word is inversely proportional to its rank, meaning that the most frequent word will occur about twice as often as the second most frequent word, and three times as often as the third most frequent word. Using this assessment of word frequency, the algorithm is able to approximate the probability of any item amongst *N* being the target item, and consequently compute its access time according to a frequency-ranked list in the serial search procedure.

2.5.3 Smaller is better

Another important facet of the Tolerance Principle is that the number of exceptions tolerated by a productive rule is relatively low (Yang, 2016, p. 66). It is certainly not simply the case that majority rules; according to Yang and his discussion of cross-linguistic evidence, only a critical number of exceptions can be tolerated, and this decreases as a proportion of *N* as *N* increases. For example, where *N* = 10, 4 exceptions can be tolerated by a productive rule (40%);

however, where N = 100, 23 exceptions can be tolerated (23%). The sublinear growth of θ_N (the tolerance threshold) as a function of N is an outcome of the solution Yang provides to the equation $T(N,N) = T(N,e)$ (see Appendix A for the full derivation). Yang argues that as a result of this feature of the TP, forming productive rules within smaller vocabularies is easier than within larger vocabularies, because a greater proportion of exceptions can be tolerated by a smaller than a larger set of items. He suggests this may in fact be the reason that children are superior language learners; it is an easier task for a child learner with a limited vocabulary to acquire the rules of a language than for an adult with a large vocabulary (2016, p. 67) because productive rules can be formed more easily over smaller sets of words.

Further, Yang suggests that at some point during the acquisition process, children freeze their productive rules on the basis of a small set of known items; they do not wait until they have acquired a full adult vocabulary. According to Yang, this is by necessity the only way to acquire productive rules, firstly because the data sparsity of the linguistic input means that many items will be encountered so infrequently that full integration into the rule-learning system would take too long, and secondly because the threshold for productivity is proportionally higher for smaller sets of items, as discussed above. In fact, Yang suggests it is possible that rules become fixed at quite an early stage, meaning that items encountered later are not entered into the data set (i.e. the effective vocabulary) involved in calculating the balance of productivity (2016, p. 106; 2018a, p. 4). Yang's discussion does not make explicit at precisely which point during the accumulation of evidence should children freeze the rules they have formed. However, this rule-freezing should occur as an (undetermined) function of N, rather than the result of a developmental shift. Indeed, Yang proposes that the TP operates for adult learners as well as children (2018b, p. 801).

2.5.4 Recursion

An important feature of the TP's formalisation is its recursive application. If the number of exceptions to a rule is too high for the rule to be productive (i.e. above the tolerance threshold), then a revised, more specific rule is sought and tested using new N and e values from a subset of the original items. Therefore the TP applies recursively, allowing it to detect productivity in smaller subclasses when the original test fails over a larger class of items. This is motivated by the *Maximise Productivity* principle described above. According to Yang, this recursive procedure is able to capture the subset regularities we find regularly in natural language, such as German noun plural forms (2016, p. 123). Therefore, the recursive application of the TP provides a mechanism that children could use to acquire these hierarchical structures of regularity found in the world's languages, offering a promising account of a complex learning problem.

2.6 Alignment with extant theories of language acquisition

Together with collaborators, Yang aligns the TP with the theory of Universal Grammar (UG) in a recent integrated approach to language acquisition (Yang et al., 2017). Specifically, the TP is listed as an inductive general learning mechanism stemming from principles of efficient computation, which is said to interplay with both domain-specific principles of language and external experience to produce a child's linguistic system. Indeed, in Yang's original exposition of the TP (2016, p. 1-2), he sets out to "shift the explanatory burden" away from an innate language acquisition device as much as possible, whilst relying on UG's recursive architecture of hierarchical structures (i.e. "Merge"; Chomsky, 1995) to at least partially explain how children acquire language on the basis of an underdetermined input.

2.7 Critical reception of the TP theory

A recent issue of *Linguistic Approaches to Bilingualism* features the keynote article "A formalist perspective on language acquisition" (Yang, 2018a), laying out the Tolerance Principle, followed by a series of commentaries in response to this keynote. Whilst lauding Yang's aim to develop a mechanistic account of language acquisition based on a quantitative assessment of the input, these commentaries highlight a number of important considerations and criticisms of the TP theory. For instance, Kapatsinki (2018) argues that whilst the serial search mechanism is a crucial assumption for the calculation of time processing according to the TP, the model of serial search has not been widely accepted in the field of psychology. He suggests that this is largely due to its incompatibility with our current understanding of the distributed representations and parallel processing in the brain. Yang's response (2018b) is that we should develop a neural theory for the “parallel brain” that can accommodate serial behavioural effects, rather than disregard these findings.

Similarly, Wittenberg and Jackendoff (2018) note the implausibility of serial search during lexical access. Additionally, they question the process whereby the output of a productive rule (e.g. *walked*) is not stored in the lexicon, yet at some point a number of rule outputs must have been listed in order to motivate the rule in the first place. However, this challenge is based on a mistaken notion: these initial exemplars are not rule outputs but individually stored lexical items. Indeed, it could be argued that this very process drives rule-learning according to the TP: when the cost of lexically listing individual forms becomes too great, a rule is created to reduce the processing burden. Therefore, there is a qualitative change in the way this information is stored and accessed. Furthermore, one could consider the well-attested U-shaped curve of English past tense development (Ervin & Miller, 1963; Cazden, 1968; Pinker & Prince, 1988; Marcus et al., 1992). During this developmental pattern, children reach a stage in which they are able to produce correct forms of irregular past tense verbs, but later begin to overregularise these verbs, producing incorrect regularisations such as *goed*. In this way, the suggestion that individually stored forms can subsequently be lost when a productive rule comes into play is not inconceivable.

Meanwhile, De Cat (2018) asks why the number of items used in the calculation of the threshold is based on the learner’s own vocabulary size, rather than the frequencies they have been exposed to in the input. I will return to this particular issue in Chapter 5, and to further discussion of the themes raised above in Chapter 7.

2.8 Experimental work on the Tolerance Principle

A small amount of published work has begun to investigate whether the TP’s theoretical predictions hold true in empirical settings. Schuler et al. (2021) present behavioural findings from a n artificial language learning study with children and adults that go some way to support the TP’s predictions for rule learning and generalisation. Participants were exposed to sentences with novel nouns that used inconsistent plural morphological markers. In a generalisation production test, children formed and used a productive rule using the plural marker when the number of exception markers (by type) in their input did not cross the tolerance threshold. In contrast, when the number of exception markers exceeded the threshold during training, children did not generalise the most common plural marker to novel items, as the TP would predict. Moreover, the fact that the token frequency of the majority plural marker was approximately consistent across both conditions indicates that it is the number of types, rather than tokens, that forms the locus of children's generalisation behaviour. Meanwhile, adults did not tend to extend one form consistently to all their generalisations in either condition, but instead used the majority marker with the same frequency it occurred in the input, known as probability matching. This and a series of further experiments are also reported in Schuler (2017, unpublished); this work will be reviewed in detail in Chapters 4 and 5, in relation to the artificial language learning experiments and assessment of the TP carried out for the current thesis.

There is also evidence that infants generalise rules from inconsistent input in accordance with predictions of the Tolerance Principle. Koulaguina and Shi (2019) exposed 14-month-olds to an unfamiliar natural language (Russian) featuring an artificial word-order shift rule. The inconsistency in the training input took the form of non-applications rather than overt violations of this rule. When the rule applied to 50% of exemplars during training, with the remaining 50% being non-application cases where the shift did not take place, infants did not generalise the rule to new instances. If the non-application cases were considered by participants to be exceptions to the regular pattern, then the number of these exceptions exceeds the tolerance threshold according to Yang's algorithm (2016, p. 8-9). Given participants' failure to generalise in this task, the authors conclude that the non-application cases were indeed treated as exceptions and therefore did not support rule-learning, as predicted by the TP.

A second experiment manipulated the frequency of the inconsistent rule. To investigate whether it was the type or token frequency of non-application items that impeded rule-learning, the relative type frequency of rule cases was increased to 80% in Experiment 2 (and non-applications reduced to 20%), whilst maintaining the same *overall* frequency (i.e. type x token frequency) as Experiment 1. With this distribution, infants were now able to learn the word-order shift rule and generalise it to novel sentences. These results are also consistent with the TP: if infants considered the non-application cases to be exceptions (as they did in Experiment 1), then their number now falls below the tolerance threshold and therefore supports generalisation.

Furthermore, the results support the TP’s approach that rule productivity is calculated on the basis of type, rather than token, frequency. Overall, the authors suggest that their experiments provide evidence of abstract rule-learning on the basis of passive exposure in infancy, thereby demonstrating that this is an early and powerful process that can take place automatically. However, it should be noted that generalisation in these experiments was assessed according to looking times rather than productions (given the young age of the participants).

Together, these studies suggest that the generalisation behaviour of infants and young children in artificial or unfamiliar language learning experiments supports the categorical predictions made by the TP, as well as the underlying assumption that productivity operates on the basis of type frequency. More generally, they indicate that similar experimental paradigms could be fruitful methods with which to explore the TP in other domains.

2.9 Applying the TP to reading

In this thesis, I will investigate whether the mechanisms proposed by Yang (2016) for the generalisation of productive rules to novel items in spoken language could also apply to reading. Whilst Yang does not address generalisation in this domain, it is theoretically possible that these mechanisms could underlie the relationship between spelling and sound for readers within a quasi-regular alphabetic writing system. Specifically, I will examine whether the TP can predict which spelling-sound correspondences readers use productively when reading aloud. In doing so, I aim to address the wider issues surrounding the generalisation of orthography-phonology knowledge by skilled and developing readers raised in Chapter 1.

Using the TP as a new approach to word reading, I suggest that the productivity of a pronunciation rule (i.e. a spelling-sound correspondence) is dependent on whether the number of exceptions (i.e. irregular pronunciations) in a set of words falls below the tolerance threshold. This means that if a spelling-sound correspondence is sufficiently consistent to pass the tolerance test, the correspondence should be applied productively during reading to pronounce novel items. In this instance, a productive rule has been formed. However, if a spelling-sound correspondence is not consistent enough to pass the tolerance test, it should not be generalised to pronounce new items.

The TP is applicable to any size of orthography-phonology correspondence. However, in accordance with the recursive application of the TP, I suggest that when readers assign a pronunciation to a letter sequence, a more general rule (e.g. between a vowel grapheme and phoneme) will be sought before a more specific subset rule (e.g. between a word body and a rime). Therefore, productive rules which use smaller orthographic grain sizes will be prioritised over rules using larger orthographic grain sizes. Only when the more general rule fails the tolerance test due to its inconsistency will a more specific rule be sought within a subset of items, such as items with a shared word body. This results from the recursive application of the TP in the search for productive rules and the *Maximise Productivity* principle (Yang, 2016, p. 72).

In this way, the TP offers a parsimonious account of the conflicting results found in the skilled and developing nonword reading literature discussed in Chapter 1, in which pronunciations seem to use both smaller and larger orthographic grain sizes under different circumstances. It goes beyond previous accounts of word reading as it does not rely on a predetermined grain size, and can in fact accommodate multiple grain sizes. Further, it specifically predicts which grain size should be used productively, depending on the consistency of the spelling-sound correspondence in question. It also offers a novel and clearly defined line of consistency beyond which generalisation is not predicted. This line of consistency (the tolerance threshold) is determined by measures of computational efficiency, rather than a set of parameters specified by researchers.

Statistical models, including the Triangle and CDP+ models, also employ consistency in their accounts of word reading, but in a different way from the TP. In the Triangle model, consistency effects arise in the hidden layer of units between the orthographic input layer and the phonological output layer. In the CDP+ model, consistency effects arise in the sub-lexical phonological assembly route. In both cases, consistency is assumed to be a graded effect in word reading. In contrast, the TP's role for consistency is categorical; although it involves the statistical properties of the input, this information is entered into the tolerance algorithm to provide a categorical threshold. The statistical information involved in this process is the type frequency counts of alternative pronunciations of orthographic units in the words that the reader has encountered. There is reason to expect that a definition of consistency which uses type frequency will be successful, as there is evidence that nonword pronunciations are better predicted by type rather than token frequency measures – in particular by the proportion of irregular-body word neighbours (Andrews & Scarratt, 1998). Finally, the recursive application of the TP means that this measure of consistency is not restricted to a single level of representation, but allows consistency at multiple levels to form part of the generalisation process when required. Specifically, it can be used to predict an interaction between consistency of the vowel grapheme and of the word body in the context of word reading. However, it should be noted that the TP will be employed here only as a mathematical model, *not* as a mechanistic model of reading similar to the fully-developed DRC, Triangle or CDP+ models.

2.9.1 Acquiring and applying rules

According to the TP, the acquisition of a rule takes place as a learner encounters word types in their linguistic experience and builds evidence of a pattern across these word types. This process occurs as a result of the learning strategy to "pursue rules that maximise productivity" (Yang, 2016, p. 72), which is in turn driven by the need to process linguistic knowledge in the most efficient (i.e., quickest) way. For example, a learner may encounter five words that use the grapheme *"i"*, such as "*tin*", "*fill*", "*limp*", "*mint*" and "*bit*". As the grapheme "*i"* is pronounced /ɪ/ in all of these words, they each provide evidence for a "*i* -> /ɪ/" pronunciation rule. According to the TP, the pronunciation of the grapheme in these items need not be stored separately, but can instead be captured more efficiently by a general rule. As a learner encounters more items, the collation of evidence will continue. For instance, if the next item to be encountered is "*mind*", the pronunciation of this item would need to be stored separately as it does not follow the "*i* -> /ɪ/" pronunciation rule. This "exception" will be lexically listed but would not affect the status of the general rule. However, if the following items to be encountered also happened to be irregular, e.g. "*pint"*, "*find"* and "*climb",* the number of exceptions observed by this point would equal four out of a total of nine items. This number of exceptions exceeds the tolerance threshold where N = 9, and thus negates use of the productive rule "*i* -> /ɪ/" rule. According to the TP, it would be more efficient to store all nine items in a frequency ranked list than to keep the "*i* -> /ɪ/" and store only the four exceptions. To summarise: as knowledge of each word type is acquired, it contributes to the balance between regulars and exceptions; this tipping point will change as the total number of relevant items (*N*) increases. Whenever the number of exceptions falls below the tolerance threshold, use of a general rule is supported.

In terms of the use of a pronunciation rule by a skilled reader, I propose that the steps undertaken to pronounce the written form of a regular word (i.e., a word that can be pronounced using GPCs) such as "*bin*" would be as follows:

1) Start with the initial grapheme "*b*", which follows the exception-plus rule route as the consistency of the *b* -> */*b*/* rule in English words passes the tolerance test.

2) Search through a frequency-ranked list of exceptions in which "*b*" does not follow the productive rule *b* -> */*b*/,* such as *climb, tomb,* etc.

3) No match for the target word amongst the list of exceptions is found, so the productive rule *b* -> /b/ (which is stored in the reader's orthography-phonology rule system) can be applied.

4) Move next to the vowel grapheme "*i*", which follows the exception-plus rule route as the consistency of the *i* -> /ɪ/ rule in English words passes the tolerance test.

5) Search through a frequency-ranked list of exceptions in which "*i*" does not follow the productive rule *i* -> /ɪ/, such as *pint*, *rind,* etc.

6) No match for the target word amongst the list of exceptions is found, so the productive rule *i* -> /ɪ/ (which is stored in the reader's orthography-phonology rule system) can be applied.

7) Move to the final grapheme "*n*", which follows the exception-plus rule route as the consistency of the *n* -> /n/ rule in English words passes the tolerance test.

8) Search through a frequency-ranked list of exceptions in which "*n*" does not follow the productive rule *n* -> /n/, such as *hymn, damn,* etc.

9) No match for the target word amongst the list of exceptions is found, so the productive rule *n* -> /n/, (which is stored in the reader's orthography-phonology rule system) can be applied.

10) Finally, the pronunciation *bin* -> /b / + /ɪ/ + /n/ ->/bɪn/ can be assembled.

I propose that the steps undertaken by a skilled reader to pronounce the written form of an irregular word (i.e., a word that cannot be pronounced accurately using GPCs) such as "*soup*" would be as follows:

1) Start with the initial grapheme "*s*", which follows the exception-plus rule route as the consistency of the *s* -> */s/* rule in English words passes the tolerance test.

2) Search through a frequency-ranked list of exceptions in which "*s*" does not follow the productive rule *s* -> */s/*, such as *aisle, isle,* etc.

3) No match for the target word amongst the list of exceptions is found, so the productive rule *s* -> /s/ (which is stored in the reader's orthography-phonology rule system) can be applied.

4) Move next to the vowel grapheme "*ou*". This grapheme does not have a productive rule stored in the readers' orthography-phonology knowledge system, as no pronunciation of this grapheme in English words is consistent enough to pass the tolerance test. Therefore, the reader does not follow an exception-plus-rule route for this grapheme. However, the grapheme forms part of a number of more specific pronunciation rules in the reader's rule-system, triggering the reader to consider the consonantal context surrounding the vowel grapheme in the target item, in this instance the word body "*oup*".

5) The word body "*oup*" is associated with a productive pronunciation rule in the reader's orthography-phonology rule system, so the reader can follow the exception-plus-rule route for this body.

6) Search through a frequency-ranked list of exceptions that do not follow the "*oup*" -> /u:p/ rule, such as *coup.*

7) No match for the target is found amongst the list of exceptions, so the "*oup*" -> /u:p/ rule can be applied.

8) Finally, the pronunciation *soup* -> /s / + /u:p/ ->/su:p/ can be assembled.

I propose that the steps undertaken by a skilled reader to pronounce the written form of an unfamiliar nonword such as "*bip*" would be as follows:

1) Start with the initial grapheme "*b*", which follows the exception-plus rule route as the consistency of the *b* -> */*b*/* rule in English words passes the tolerance test.

2) Search through a frequency-ranked list of exceptions in which "*b*" does not follow the productive rule *b* -> */*b*/,* such as *climb, tomb,* etc.

3) No match for the target word amongst the list of exceptions is found, so the productive rule *b* -> /b/ (which is stored in the reader's orthography-phonology rule system) can be applied.

4) Move next to the vowel grapheme "*i*", which follows the exception-plus rule route as the consistency of the *i* -> */*ɪ/ rule in English words passes the tolerance test.

5) Search through a frequency-ranked list of exceptions in which "*i*" does not follow the productive rule *i* -> */*ɪ/, such as *pint, rind,* etc.

6) No match for the target word amongst the list of exceptions is found, so the productive rule *i* -> */*ɪ/ (which is stored in the reader's orthography-phonology rule system) can be applied.

7) Move to the final grapheme "*p*", which follows the exception-plus rule route as the consistency of the *p* -> */*p/ rule in English words passes the tolerance test.

8) Search through a frequency-ranked list of exceptions in which "*p*" does not follow the productive rule *p* -> */*p/, such as *coup, psalm,* etc.

9) No match for the target word amongst the list of exceptions is found, so the productive rule *p* -> /p/, (which is stored in the reader's orthography-phonology rule system) can be applied.

10) Finally, the pronunciation *bip* -> /b / + /ɪ/ + /p/ ->/bɪp/ can be assembled.

2.9.2 Similar approaches to reading

A previous exploration of print-to-sound mappings in English orthography bears some similarities to the Tolerance Principle's approach to reading introduced above. Specifically, Vousden et al. (2011) applied the Simplicity Principle (Chater & Vityani, 2003) to the English spelling system, with the aim of identifying which representational units specify the mappings from print to sound as simply as possible; information which could in turn be used to facilitate the reading acquisition process. As discussed above, the Simplicity Principle states that simpler explanations of data should be favoured over more complex ones. Like the TP, the underlying basis is one of computational efficiency, although here it is measured according to description length of the data rather than online processing time.

Vousden et al.'s application of the Simplicity Principle to English orthography balances the simplicity (or length) of a hypothesis that describes spelling-sound correspondences against how accurately this hypothesis can recreate the data (i.e., produce target pronunciations). Specifically, they compare the *total description length* of alternative hypotheses, which is a combination of the *description length of the current hypothesis* (the number of spelling-sound mappings used), and *the description length of the data under this hypothesis* (the accuracy with which it offers the target pronunciation as the most probable outcome). The authors tested a range of hypotheses which make use of different representational units, including graphemes, onsets and bodies, and whole words. According to the Simplicity Principle, the preferred hypothesis will be that with the shortest total description length.

Their initial analysis found that the shortest total description length was achieved by the grapheme-only hypothesis, followed by the head–coda and onset–body hypotheses, and finally the whole-word hypothesis. However, further analysis demonstrated that adding contextual information for some grapheme-phoneme mappings reduced ambiguity and the total description length associated with the grapheme-only hypothesis. Additionally, results varied according to the size of the vocabulary that was used; for example, the whole-word hypothesis offered the shortest total description length for very small vocabularies.

Overall, the authors conclude that the shortest total description of the data was provided by a hypothesis composed mostly of grapheme-phoneme mappings, but which also included a range of other unit types. However, beyond knowing that this is the preferred hypothesis according to the Simplicity Principle, it would be important to know in which particular orthographic instances larger unit types should be used to improve individual grapheme-phoneme mappings. Further, we cannot assume that the simplest way to capture the input data will necessarily be used as basis for learners' generalisations. The approach of the Tolerance Principle similarly involves a range of representational units, but further, can be used to explicitly specify when certain orthographic unit sizes (i.e., more specific rules) should be prioritised over others during generalisation. The success of the TP's predictions regarding English orthography will be examined experimentally in Chapter 3. Additionally, Vousden et al.'s finding that strategy choice will depend on the size of the specific vocabulary highlights the importance of taking into account the nature of expanding vocabularies, and associated changes in the required cognitive resources, in our understanding of reading development. The particular issue of acquiring knowledge of spelling-sound mappings during reading development will be also explored in more detail in Chapter 3.

2.10 Summary and thesis outline

These introductory chapters have offered an overview of research on word reading and linguistic productivity, with a specific focus on how knowledge is acquired and generalised within a quasi-regular system. Similar questions surrounding how the consistency and frequency of patterns affect their productivity have fuelled research in both these fields, but it is clear that for word reading, a number of issues remain unexplained. For instance, readers do not simply reproduce the distributions of their input (Treiman et al., 2003; Treiman & Kessler, 2019), but neither do they categorically use the most frequent pattern (Andrews and Scarratt, 1998; Pritchard et al., 2012). Their use of different orthographic grain sizes is particularly difficult to predict (Brown & Deavers, 1999). Consequently, computational models of word reading have been unable to satisfactorily capture human reading behaviour (Treiman et al., 2003). Meanwhile, Yang's (2016) Tolerance Principle has offered a novel account of linguistic productivity, in which categorical generalisation of a pattern is predicted according to whether the number of exceptions to the pattern crosses a critical threshold. The current thesis will apply this newly-proposed theory to word reading for the first time, in order to assess whether it can address outstanding issues in this field.

The following chapters present this investigation through four experiments. Experiment 1 offers an initial, exploratory application of the TP to reading, investigating whether the TP algorithm can be used to predict adults' and children's generalisation of familiar spelling-sound correspondences in a nonword reading aloud study. Experiment 2 examines whether the TP can predict adults' and children's generalisation of novel, inconsistent spelling-sound correspondences in an artificial orthography learning paradigm. Experiment 3 expands this investigation by manipulating the token frequency of regular and irregular items during training. Finally, Experiment 4 explores whether the recursive application of the tolerance threshold can predict adults' generalisation of novel, context-sensitive spelling-sound correspondences in an artificial orthography.

Chapter 3: Generalisation of orthography-phonology correspondences in nonword reading by adults and children

3.1 Introduction

Chapter 1 provided a comprehensive review of research on adult nonword reading aloud. This review highlighted a number of unresolved issues in the word reading literature including readers' use of different orthographic grain sizes; the characterisation of categorical, rule-based versus continuous, statistics-based orthographic knowledge of skilled readers; and the shortcomings of extant models in predicting nonword reading behaviour. Chapter 2 introduced Yang's (2016) Tolerance Principle as theory of linguistic productivity, and also laid out the potential application of the TP mechanism to word reading. In particular, I suggested that the TP's novel account of the way learners extract and extend regularities from quasi-regular input could be used to assess the consistency of spelling-sound correspondences and predict readers' productive use of these correspondences. The current chapter presents an exploratory study which uses the TP to predict orthography-phonology generalisation within a nonword reading paradigm for the first time. This experiment offers an initial opportunity to investigate whether Yang's linguistic theory can be successfully applied to another domain to predict readers' generalisation behaviour, and in so doing address some of the unresolved questions in reading research.

In addition to skilled reading research, nonword studies are also often used to understand more about the reading development process, for instance by comparing responses made by children of different ages and adults. This methodology can help reveal what pronunciation strategies readers use at different stages throughout development. For this reason, the current study will also employ the TP to investigate the nonword reading behaviour of developing readers, setting out to address outstanding issues in this area of research as described below.

3.1.1 The acquisition of spelling-sound knowledge in children

Results from studies of nonword reading by children suggest that whilst early readers are more likely to rely on GPCs than the body unit in nonword pronunciations (Marsh et al., 1981, Brown & Deavers, 1999), knowledge of the word body soon begins to emerge. Indeed, there is evidence that sensitivity to the word body is associated with reading ability: for example, Laxon et al. (1991) compared nonword pronunciations by children aged 6-13 who were either average or better-than-average readers for their age. The more skilled group demonstrated stronger consistency effects in their use of body-level pronunciations for nonwords, although both groups demonstrated some awareness of alternative pronunciations for inconsistent bodies. The authors suggest that knowledge of word bodies may develop as children become increasingly aware of patterns in high frequency words, and that sensitivity to regularity and consistency in general develops with increased exposure to printed words.

Additional research suggests that this trend towards an awareness of consistency and larger orthographic units does continue throughout development towards levels of skilled reading. For instance, Coltheart and Leahy (1992) found that for adults and children (grades 1-3), the majority of nonword pronunciations used GPCs. However, all readers also produced some rime-based analogies (classed as “irregular” pronunciations). Notably, this varied with age: children in grade 1 gave fewer irregular responses than children in grades 2 and 3, whilst adults demonstrated the highest degree of sensitivity to rime-level consistency. The authors suggest that knowledge of rime-level units is acquired after GPCs, possibly as a result of increased text experience. Similarly, Treiman et al. (1990) reported that even early readers (aged 6-7) make some use of body-level units rather than relying solely on GPCs, and that this trend continues (and perhaps strengthens) throughout reading experience and into adulthood.

Whilst this pattern of increasing sensitivity to the body unit through development is seen across the literature, it does not address under which circumstances readers use different levels of spelling-sound correspondence (i.e. smaller GPC units vs. larger body units). Indeed, Brown and Deavers (1999) suggested that the much-debated dichotomy between GPC or body strategies is a misplaced issue. Their results from a range of nonword reading tasks suggest that adults and children are flexible in their use of orthographic units, adapting their reading strategy as a function of task demands. Even the youngest readers demonstrated sensitivity to task demands in their use of orthographic units: they used a higher proportion of rime-based analogical pronunciations in a “clue word” task than they did in a task which involved reading isolated unfamiliar items. Consequently, the authors proposed the Flexible Unit Hypothesis, which allows readers to vary their pronunciation strategy depending on the type of generalisation involved.

In a later nonword study, Steacy et al. (2019) also considered when competing strategies are used, rather than whether one is favoured over another overall. Following a suggestion by Treiman et al. (2003), they explore the possibility that choice of context-dependent (i.e., rime-based) or context-independent (i.e., GPC) vowel pronunciation in nonwords is the result of a “trade-off between vowel GPC frequency and strength of context-dependent orthography-phonology relationships in the rime unit [of English words]” (2019, p.51). They find that consistency of a context-dependent rime pronunciation in English words is negatively correlated with use of the context-independent vowel pronunciation by participants, and positively correlated with use of the context-dependent pronunciation. Further, these relationships were moderated by reading skill, with rime consistency having a greater effect on pronunciation for more proficient readers. They suggest this is the result of increasing support for the alternative vowel pronunciations in written texts as reading experience develops. Similarly, in a discussion of spelling development, Kessler (2009) suggests that children may pay more attention to contextual information when there is no clear spelling candidate for a phoneme in isolation. He posits that there may be a pay-off involved in learning a conditional rule whereby this additional cost is justified only when the pronunciation of a vowel alone is inconsistent.

Neither Kessler (2009) nor Steacy et al. explore how the “trade-off” between the strength of competing pronunciations at different orthographic levels could be measured or characterised. In this thesis, I also suggest vowel pronunciations in nonwords are determined by a balance between competing grapheme- and body-level correspondences, and that this is based upon reading experience. However, my approach goes further by characterising the precise terms of this balance and the implicit process behind it: namely, that the balance is determined precisely by a consistency threshold provided by the Tolerance Principle, according to principles of computational efficiency.

3.1.2 Application of the TP to reading in Experiment 1

As described in detail in Chapter 2, Yang's account of generalisation in language acquisition takes a rule-based approach but also incorporates statistical information. The TP assesses the productivity of a rule according to a numerical balance between regulars and exceptions. It states that in order for a learner to form a productive rule, the number of exceptions to the rule must fall below a critical threshold. The threshold is generated by an algorithm (Yang, 2016, p. 8-9) which uses the total number of items a rule can apply to and the number of exceptions which do not follow the rule. In this way, the TP offers a categorical metric of consistency: a rule that is consistent enough to pass the threshold can be generalised to new items. Importantly, this account of consistency is based on type frequencies (i.e. how many different item types follow the rule), rather than token frequencies (i.e. the relative frequency of these items in the input). Further, Yang suggests that learners are driven by computational efficiency to "pursue rules that maximise productivity", (the *Maximise Productivity* principle (2016, p. 72)). This process leads to the recursive application of the TP: when the number of exceptions to a rule crosses the threshold, it triggers a search for more specific rules within subsets of the input. In this way, the TP can detect productivity within subclasses. This approach to generalisation is novel as it offers an account of consistency which can be applied to hierarchical (or nested) regularities.

In this chapter, I will use the TP's prediction for productive patterns, assessment of consistency in the input, and recursive search for sub-regularities, to address issues surrounding the generalisation of spelling-sound knowledge by skilled and developing readers. This investigation will assess the reading behaviour of adults and children using the English writing system. Specifically, I will explore whether the TP can predict which level (or grain size) of spelling-sound correspondence adult and child readers use when reading aloud. Using the TP algorithm, I will assess which orthography-phonology correspondences are sufficiently consistent across English words to pass the tolerance test; correspondences which pass the test should be used productively. More general pronunciation rules using smaller grain sizes (i.e. individual graphemes) should be prioritised over more specific pronunciation rules using larger grain sizes (i.e. the word body). Only when the more general rule does not pass the tolerance test should a recursive search for a more specific rule be triggered. This mechanism provides a precise balance between the productive use of graphemes and bodies based on a categorical measure of consistency (see *Section 2.9* for an introduction to this approach and *Section 3.2* for a detailed exposition). To assess the TP's predictions, Experiment 1 will examine which spelling-sound correspondences readers use when presented with novel items (i.e. nonwords) to pronounce in English orthography.

3.1.3 Comparing the TP with previous approaches to reading

A comparison between the TP's novel approach to reading and established computational models of word reading was introduced in Section *2.9*. This theme will be developed in the current chapter, through a comparison of nonword pronunciations produced by adult and child participants, three extant models of reading, and the TP. Therefore, it is worth highlighting again some important differences between these contrasting approaches to generalisation of spelling-sound correspondences. The DRC (Coltheart et al., 2001) is a rule-based model, according to which nonword pronunciations are generated using GPCs; there is no effect of consistency or longer letter sequences in these generalisations. The Triangle model (in this instance, the Chang et al., (2019) version of the Harm & Seidenberg (2004) model) is a distributed-connectionist model which is able to generalise more complex spelling-sound mappings such as body-rime correspondences. In this way it allows for graded consistency effects in nonword pronunciations. The hybrid CDP+ model (Perry et al., 2007) has an interactive activation lexical route and a distributed-connectionist non-lexical route which is sensitive to patterns of co-occurring graphemes. Therefore, its generalisation involves continuous consistency effects and multiple orthographic grain sizes. Both of the latter (statistical) accounts weight spelling-sound correspondences according to their token frequencies, whilst the DRC uses only type frequencies. The new TP account of orthography-phonology generalisation involves both smaller and larger grain sizes motivated by a recursive search for productive patterns. Although it is rule-based, the TP uses statistical information (specifically, type frequencies) to produce a tolerance threshold. This threshold provides a categorical metric of consistency which guides the generalisation of spelling-sound correspondences by predicting an interaction between the consistency of smaller and larger orthographic units.

This novel approach to orthography-phonology generalisation is a promising addition to research on word reading. By filling the theoretical gap between extant rule-based and statistical models, it is possible that the TP will also be able to redress the gap in readers' behaviour thus far unaccounted for by previous models. Namely, that readers make fewer regular GPC responses than rule-based models predict, but also fewer context-sensitive and lexicalisation responses than statistical models predict (e.g. Andrews & Scarratt, 1998; Pritchard et al., 2012). If successful, an account such as the TP, which predicts precisely when context-sensitive correspondences should and should not be used productively according to a consistency threshold, might offer a valuable development in our understanding of human reading behaviour.

3.1.4 Using the TP to inform our understanding of reading acquisition

There are three factors that I will consider when assessing the TP specifically within the context of reading development for the first time, particularly in the comparison of child and adult reading behaviour. These stem from the fact that spoken language acquisition and reading development are different challenges. The first involves reading instruction: children in UK primary schools learn to read through a systematic phonics instruction programme. This method emphasises the most frequent correspondences between graphemes and phonemes (GPCs) in written English. Currently, little is known about the relationship between instruction and the generation of productive rules using the TP, as it was developed as a theory of spoken language acquisition, which typically does not involve or require any explicit instruction. However, it is certainly possible that children who learnt to read using phonics are more likely to use GPCs when reading novel items than adults who may have learnt using other methods without this emphasis. Indeed, Thompson et al. (2009) found that adults who had learned to read using phonics demonstrated a "cognitive footprint" of their instruction in nonword pronunciations; they used more regular GPCs and fewer context-dependent pronunciations in their responses compared with adults who did not have childhood phonics instruction. This finding suggests that the type of reading instruction can have lasting influence over use of spelling-sound correspondences even after many years of reading experience.

The second factor involves reading experience. This is related to instruction, but is based more generally on the TP's central tenet that productive rules are built on an individual's direct experience. Yang's theory states that "the execution of the TP should be based on an individual learner's vocabulary" (2016, p. 70), as the balance between regular and irregular items for each learner is determined by the specific linguistic input they have received. Therefore, we may expect subtle differences between individuals' use of productive rules. In the current nonword study, this may be particularly likely for adult participants who may have received very varied reading input and instruction methods in schools across the country during their reading acquisition process, as well as years of reading experience involving varying text types and specialist knowledge. Child participants, meanwhile, may differ from each other less in their reading experience. They have had fewer years of reading in which to diverge, are likely to have encountered similar primary school texts, and have all undergone phonics instruction in the first years of primary school. As noted above, Thompson et al. (2009) report evidence of the long-lasting effect of reading instruction; adult readers who had learned to read using phonics produced significantly more regular nonword pronunciations than adults who had learned to read using a different method. For these reasons, we may expect nonword responses made by adult participants to be more variable than responses made by child participants.

The third factor involves previous findings about reading acquisition. As discussed in *Section 3.3.1*, studies have reported that early readers are likely to rely on GPC strategies, with use of the body unit increasing gradually throughout development (March et al., 1981; Treiman et al., 1990; Bruck & Treiman, 1992; Brown & Deavers, 1999), potentially as the result of increased text experience (Laxon et al., 1991). Similarly, Vousden et al. (2011) note that the most efficient pronunciation strategy according to the Simplicity Principle changes as vocabulary size increases. Therefore, in our study we may expect child participants to use the body unit less than adult participants if their use of the body unit has not yet reached adult levels, and this may interact with the TP's predicted use of the body pronunciation.

3.1.5 Similar approaches to reading development

Previous research has similarly sought to identify the most useful orthography-phonology correspondences for readers during development. For example, Vousden (2008) presents an exploration into how the statistical structure of English could inform teaching practice by measuring how well spelling-sound correspondences at different levels predict correct pronunciations of English words, thus determining their potential utility during explicit reading instruction. This investigation follows the theory of Rational Analysis (Anderson 1990), which seeks to determine what can be learned from the statistical properties of the environment and states that an optimal solution to a problem should be solved in a maximally efficient way (as discussed in *Section 2.4*). Assuming that adult readers are sensitive to the statistical structure of the reading system, this line of reasoning suggests that analysing this structure will reveal the most useful information for a learner seeking regularities in the input they receive.

Vousden (2008) highlights a notable feature of the statistical structure of the English language: that word frequency adheres to Zipf's Law. As discussed in *Section 2.5.3,* this means that the most frequent words occur very frequently compared to less frequent words. Vousden notes that this has important implications when considering reading instruction materials, because the most frequent words will account for a large proportion of words that a reader encounters. Therefore, she examined how well frequency data at different orthographic levels relates to Zipf's Law, in order to evaluate what proportion of text could be successfully read using knowledge of the most frequent sound-spelling mappings at these levels.

For the analysis, monosyllabic word frequencies were obtained from the CELEX database (Baayen et al., 1993), from which the frequencies of spelling-sound mappings at three different levels were calculated, namely for whole words, onsets and rimes, and graphemes. Importantly, only the most frequent (or "regular") spelling-sound mapping for inconsistent onsets/rimes and graphemes was selected for this analysis. This approach circumvented the issue of distinguishing between alternative pronunciations of orthographic units. Further, the 100 most frequent words were excluded from the onset/rime and grapheme analysis, as it was acknowledged that sublexical mappings may not predict the pronunciations for these words successfully (many of which have irregular pronunciations), and the aim of the investigation was to develop an optimal strategy for reading text beyond these words.

Overall, the analysis presented quantitative evidence that the frequency of orthography-phonology mappings at all three levels follows Zipf's Law. Additionally, there was a clear benefit of learning a small sight vocabulary of high frequency words, for which whole word learning is most appropriate. As vocabulary increases beyond these words, knowledge of a small number of the most frequent grapheme-phoneme mappings enabled a large proportion of the remaining text to be read successfully. Comparatively, knowledge of many more rime units must be acquired before these are more predictive than grapheme-phoneme mappings, suggesting that grapheme-phoneme mappings are more useful for early learners.

However, I suggest that the value of body-rime mappings may be more evident when the relative consistency of specific mappings at different levels is taken into account (which this study did not do, using only the most frequent of any alternative mappings at each individual orthographic level). Indeed, Vousden (2008) notes that issues around inconsistency at the grapheme-phoneme level may be alleviated by taking contextual information from the body-rime level into consideration. Overall, she suggests that this investigation provides a useful starting point for exploring the utility of spelling-sound mappings at different levels in a quantitative way; a similar approach to that taken up in this chapter.

3.2 Experiment 1

Experiment 1 offers an initial, exploratory investigation of the TP in the context of reading, aiming to evaluate whether skilled and developing readers generalise spelling-sound correspondences in English orthography according to the predictions of the TP. This was investigated using a nonword reading aloud task in which adult and child participants read aloud 198 nonword items written in English orthography. Using the tolerance algorithm provided by Yang (2016, p. 8), I calculated the tolerance thresholds of spelling-sound correspondences of vowel graphemes and word bodies in English monosyllabic words, using word frequencies from the CELEX database (Baayen et al., 1995). Identifying the productive spelling-sound rules in English according to this method produced the pronunciations for the nonword items that would be predicted by the TP. By comparing the TP predicted pronunciations with the nonword pronunciation responses of our participants, I could assess whether the TP is able to predict which spelling-sound correspondences adult and child participants use to pronounce novel items.

Specifically, the TP predicts that the pronunciation of a vowel grapheme which passes the tolerance test should be used by participants to pronounce a nonword item, as this offers a pronunciation rule at a more general level. Only when no pronunciation of the vowel grapheme passes the tolerance test should a more specific rule at the level of the word body be used by participants to pronounce a nonword item. This results from the *Maximise Productivity* principle (Yang, 2016, p. 72) and the TP's recursive mechanism, according to which failure to find a productive rule over a complete set of items will trigger the recursive application of the TP in order to seek more specific rules within subsets of items. Thus, use of a productive body rule should be modulated by the availability of a productive vowel rule. Meanwhile, the TP provides no predicted pronunciation for items that have no single pronunciation that passes the tolerance test outright at either the vowel or body level. For these nonword items, an exploratory analysis investigated the range of pronunciations participants used when reading them aloud. The aim of this investigation was to assess whether participants demonstrate sensitivity to a range of statistical properties of English orthography in their use of possible pronunciations for these items.

Following the analysis of nonword pronunciations in relation to the TP, I assessed the TP's predictive success for nonword reading in comparison to three extant models of word reading. These were the connectionist Triangle model (using the Chang et al., (2019) version of the Harm & Seidenberg (2004) model), the dual-route connectionist CDP+ model (Perry et al., 2007), and the rule-based DRC model (Coltheart et al., 2001). Finally, I explored whether the TP's novel role for consistency is key to its ability to predict human nonword reading behaviour. Specifically, I compared the TP's type-based, categorical metric of consistency which can apply to multiple orthographic grain sizes, with conventional, continuous measures of consistency based on type or token frequencies. Throughout these evaluations, I also considered how the TP may behave differently for adult and child participants, including the effect of systematic phonics instruction and reading experience on the use of certain spelling-sound correspondences, and the variability in participants' nonword responses.

3.3 Method

3.3.1 Participants

25 adult participants (age range: 18-40; 18 females and 7 males) were recruited from the student body of Royal Holloway, University of London. 29 child participants (age range: 8 years 3 months – 9 years 5 months; mean age: 8 years 10 months; 16 females and 13 males) were

recruited from two primary schools in the south of England. Participants were monolingual, native English speakers, with a Southern British English accent and no known language or learning difficulties. Participants had normal or corrected-to-normal vision. Adult participants received £5 for their involvement in the study; the schools received redeemable vouchers. The data from one adult participant were discarded due to difficulty completing the task. Therefore, data from 24 adult and 29 child participants were included in our analysis. The study received ethical approval from the Royal Holloway Ethics Committee.

3.3.2 Stimuli and design

The stimuli were 198 monosyllabic nonwords, constructed from existing English onsets and bodies, and using legal bigrams (see Appendix B for the complete stimuli set). Items were either selected from the ARC nonword database (Rastle et al., 2002), or constructed by the experimenter. The mean orthographic neighbourhood (N) size was 3.80. The mean item length was 4.83 letters (range 3-7 letters).

The consistency of each vowel grapheme and word body used in the nonword items was measured using the tolerance test. To do this, type frequencies of the occurrences of each vowel grapheme and word body in English words from the CELEX database (Bayyen et al., 1995) were entered into the tolerance algorithm (Yang 2016, p. 9, see discussion in *Section 2.5*):

$$e \leq \theta_N = \frac{N}{\ln N}$$

This process assessed whether the number of exceptions to any particular pronunciation of the relevant vowel grapheme and word body in English words fell below the tolerance threshold, meaning that the spelling-sound correspondence was consistent enough to pass the tolerance test and form a productive pronunciation rule.

Sixty-six vowel grapheme/word body pairs (e.g. EA/EAT) were each used in three nonword items in the stimuli set (e.g. SMEAT, PREAT, THEAT). In most items, the vowel grapheme was shorter than the word body (e.g. IE/IEND in the nonword TIEND); however, the vowel grapheme and word body were the same length in a small number of items (e.g. OW/OW in the nonword DOW). According to the outcome of the vowel grapheme/word body pair on the tolerance test, items were categorised into the following seven conditions:

Condition 1: Vowel winner, body winner, no conflict (60 items) e.g. YOOT
Condition 2: Vowel winner, body winner, conflict (30 items) e.g. SMEAD
Condition 3: Vowel all fail, body winner (51 items) e.g. CHOWL
Condition 4: Vowel all fail, body all fail (12 items) e.g. GLOWN
Condition 5: Vowel winner, body all fail (3 items) e.g. TROOD
Condition 6: Vowel all fail, body all pass (30 items) e.g. FOUTH
Condition 7: Vowel winner, body all pass (12 items) e.g. GLEAF

To exemplify, in the nonword YOOT, the vowel grapheme OO has a winning pronunciation /u:/ which passes the tolerance test. This is called the vowel winner pronunciation. The word body OOT also has a winning pronunciation /u:t/ which passes the tolerance test. This is called the body winner pronunciation. This body winner pronunciation does not conflict with the vowel winner pronunciation, so YOOT falls into condition 1 (vowel winner, body winner, no conflict). The nonword SMEAD has a vowel winner pronunciation (/i:/), and a body winner pronunciation (/ɛd/), but as these are different pronunciations, this item falls into condition 2 (vowel winner, body winner, conflict). The vowel grapheme OW in the nonword item CHOWL is not consistent enough to have a vowel winner pronunciation that passes the tolerance test, but OWL does have a body winner pronunciation. Therefore, this item falls into condition 3 (vowel all fail, body winner). Neither the vowel grapheme nor the word body of the nonword GLOWN is consistent enough to pass the tolerance test, so this item falls into condition 4 (vowel all fail, body all fail). The nonword TROOD has a consistent vowel grapheme with a vowel winner pronunciation but an inconsistent body so falls into condition 5 (vowel winner, body all fail). The nonword FOUTH has no vowel winner, and the number of English words which use its body OUTH is so low that all possible pronunciations pass the tolerance test (no exceptions to any pronunciation are high enough to exceed the tolerance threshold). This means there is no outright body winner pronunciation, so this item falls into condition 6 (vowel all fail, body all pass). The nonword item GLEAF falls into condition 7 (vowel winner, body all pass) because the vowel grapheme EA has a vowel winner pronunciation, but all pronunciations of the body EAF pass the tolerance test. The size of each condition was restricted by the number of possible vowel grapheme/word body pairs found in English monosyllabic words that fell into each category. Thus, the number of items in each category varied accordingly.

3.3.3 Procedure

Adult participants were tested individually in a language lab in the Department of Psychology at Royal Holloway, University of London. Participants were informed that they would be presented with 198 nonwords to read aloud. The nonword stimuli were presented one at a time in uppercase in white font on a black background, in the centre of a computer screen, using DMDX software (Forster & Forster, 2003). The stimuli were presented in a randomised order for each participant. Participants were asked to read each nonword aloud into a microphone, and their responses were recorded using the audio-capture capacity of DMDX (Forster & Forster, 2003). Participants were instructed to read each item as quickly and accurately as possible. The duration of each recording simultaneously with the presentation of each nonword on the screen was 2500 ms. In between the presentation of each nonword, a focus screen displaying < > was presented for 2000 ms. Participants were provided with a set of 4 practice items before beginning the test phase. Midway through the experiment, after 99 items, participants were given the opportunity to rest, and could choose when to resume the experiment by pressing the spacebar on the keyboard. The duration of the experiment was approximately 15 minutes.

Child participants were tested individually on primary school sites. Children carried out the same testing procedure as adults except that participants were asked to read made-up words as carefully as possible; items were presented on the screen until the participant pressed the spacebar and for a minimum of 2500 ms; and participants were given the opportunity to rest and resume after every 33 items, with a total of six 33 item blocks presented in a randomised order.

Additionally, child participants carried out the Test of Word Reading Efficiency – Second Edition (TOWRE-2; Torgesen, Wagner, & Rashotte, 2012), to assess individual's ability to pronounce printed words and nonwords accurately. Participants carried out both the Sight Word Efficiency (SWE) and Phonemic Decoding Efficiency (PDE) subtests, in which they read aloud as many words (SWE) and nonwords (PDE) from a list as they could within 45 seconds.

3.3.4 Transcription

Recordings were transcribed into symbols representing the DRC's phonemic vocabulary (Coltheart & Rastle, 1999: Appendix A). On occasion (0.01% of responses), multiple answers were recorded by a participant in response to a single stimulus, for example because the participant had attempted to repeat or correct themselves mid-way through or after their initial response. In these cases, the first complete pronunciation of the nonword was selected and transcribed by the experimenter. In 0.5% of cases either no response was recorded, or the recording of the participant's pronunciation was cut short at the end of the recording duration and thus these responses were excluded.

3.4 Results

The analysis addressed three broad questions. Firstly, I investigated whether the TP can predict adult and child participants' use of vowel grapheme and word body orthographic units in their nonword pronunciations, and also considered what pronunciations participants use for items which have no pronunciation predicted by the TP. Secondly, I compared the ability of the TP to predict adult and child nonword pronunciations with that of three computational models of reading, finding that the TP offers the most successful account. Thirdly, I investigated whether the TP's categorical metric of consistency offers value beyond a continuous metric of consistency in predicting adult and child nonword reading behaviour. Together, results suggest that the TP offers a relatively successful account of adult and child nonword reading behaviour and provides a novel contribution to our understanding of both developing and skilled reading.

Additionally, child participants were assessed on SWE and PDE background measures using the TOWRE-2 sub-tests. The mean raw score for SWE was 62.0 (*SE* = 1.1); the mean raw score for PDE was 32.9 (*SE* = 1.7). The mean age-scaled score for SWE was 99.2 (*SE* = 1.9); the mean age-scaled score for PDE was 101.5 (*SE* = 2.2).

3.4.1 Can the TP predict skilled and developing readers' nonword pronunciations at the vowel grapheme and word body level?

The analyses in this section addressed two predictions of the TP. The first was that participants should use the available vowel winner in their pronunciations irrespective of the status of the word body; use of the vowel winner should not be modulated by the body status. The second was that participants' use of a body winner should be modulated by the presence of a vowel winner. Each of these predictions is addressed in turn using mixed-effects logistic regression models. An exploratory analysis then investigated the range of pronunciations participants use when reading aloud nonword items without vowel or body pronunciations predicted by the TP.

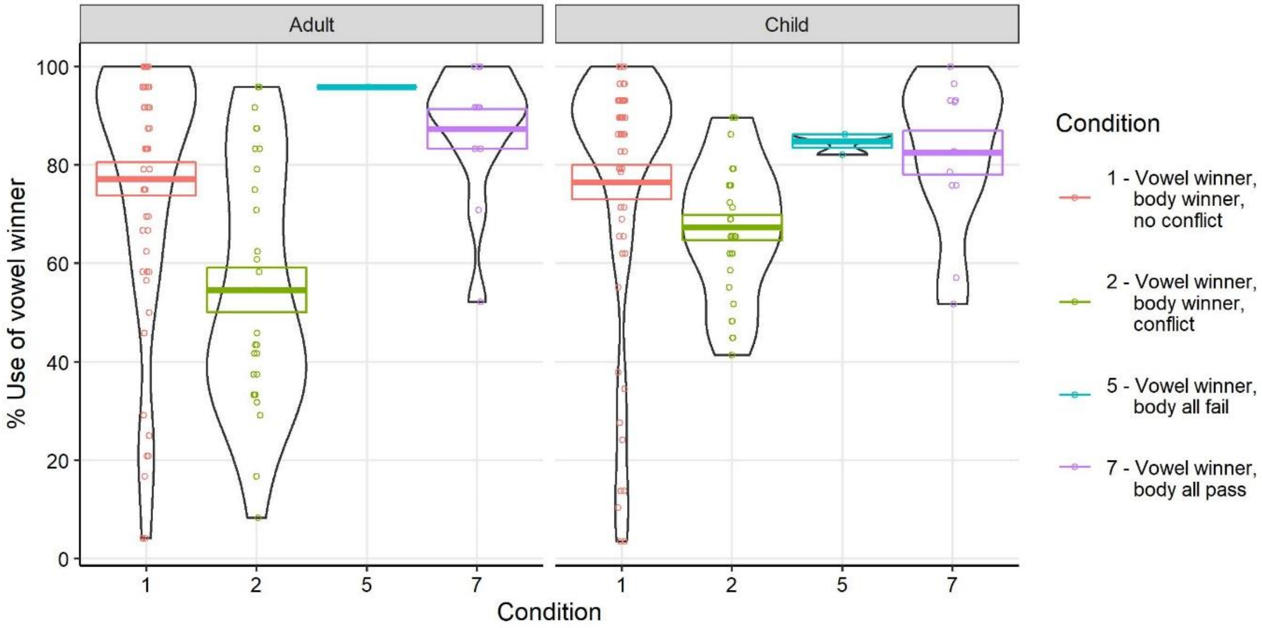
3.4.1.1 Is the vowel winner pronunciation used regardless of body status?

Pronunciations which pass the tolerance test outright at the level of the vowel grapheme and offer a productive rule are referred to as “vowel winners”. The TP predicts that when a nonword has a vowel winner pronunciation available, participants will use this pronunciation regardless of the nonword's body status. Evaluating this prediction involved 105 items from the four conditions with vowel winner pronunciations, or “vowel winner conditions”: vowel winner, body winner, no conflict (1); vowel winner, body winner, conflict (2); vowel winner, body all fail (5); and vowel winner, body all pass (7). The vowel winner pronunciation was used in 72.22% of responses by adult participants and 74.84% responses by child participants, suggesting that it is the favoured pronunciation for items with a vowel winner, as predicted by the TP.

To evaluate this prediction statistically, I examined participants' use of the vowel winner in each condition. Recall that the TP predicts the vowel winner will be the pronunciation used in all four vowel winner conditions, regardless of whether the body also passes the tolerance test. This is because the vowel offers a more general pronunciation rule which should be prioritised over a more specific body-level correspondence. Figure 3.1 shows use of the vowel winner pronunciation in vowel winner conditions by adults and children. Note that condition 5 contains only three items.

Figure 3.1

Use of the Vowel Winner Pronunciation (%) by Adult and Child Participants in Vowel Winner Conditions



Note. In this figure and subsequent similar figures (unless specified otherwise), the horizontal line represents the mean, the box around the mean represents standard error, data points represent individual nonword items, and the borders around data points are smoothed density curves.

A logistic mixed-effects analysis was used to assess whether use of the vowel winner varied as a function of condition. For this and further analyses below, I used R (version 3.6.0; R Development Core Team, 2019) and the *lme4* package (version 1.1-21; Bates et al., 2015). This approach is able to include predictors as fixed effects and participant and item as random effects simultaneously in the same models. The *p*-values reported are based on the Wald Z statistic for each effect (Jaeger 2008). A maximal random effects structure was sought in each model (following Barr et al., 2013). When a model failed to converge, the random effects structure was simplified until the model converged.

For the vowel winner analysis, adults and children were treated as separate participant age groups, as a mixed-effects model using a Condition x Age Group interaction explained significantly more data variance than a reduced model using Condition and Age Group as the fixed effects (($\chi^2(3) = 33.849$, $p < .001$). As a maximal mixed-effects logistic regression model failed to converge, I ran an intercepts-only model with the Condition x Age Group interaction as a fixed effect (rotating each condition and participant group as the reference levels), and item and subject random intercepts, with use of the vowel winner as the outcome measure. Table 3.1 presents the output for the model using adult as the age group reference level and conditions 1, 2 and 7 as the condition reference levels. The output of the model using condition 5 as the reference level is omitted as this failed to converge due to too few items. Results suggest that adults' use of the vowel winner pronunciation was lower in the vowel winner, body winner, conflict condition (2) than in the three other conditions. There were no significant differences between adults' use of this pronunciation in the other vowel winner conditions. Table 3.2 presents the output for the models using child as the age group reference level and conditions 1, 2 and 7 as the condition reference levels. Results suggest that children's use of the vowel winner pronunciation was lower in the vowel winner, body winner, conflict condition (2) than in conditions 1 and 7. There were no significant differences between children's use of this pronunciation in the other vowel winner conditions. These results are inconsistent with the TP prediction that body status should be irrelevant for items with a vowel winner. Instead, data from adults and children showed that responses were less likely to use the vowel winner pronunciation when there was a conflicting body winner.

Table 3.1

Output from mixed-effects model comparing use of the vowel winner pronunciation in vowel winner conditions, using Adult as the age group reference level
*glmer(Vowel Winner Score ~ Condition*Age + (1|Participant) + (1|Item))*

Fixed Effect	Est.	St. Error	*z* value	*p* value	Inverse Logit Probability[8]
(1) Vowel winner, body winner, no conflict vs. (2) Vowel winner, body winner, conflict (Adult)	-1.527	0.333	-4.588	<.001	0.178
(1) Vowel winner, body winner, no conflict vs. (5) Vowel winner, body all fail (Adult)	1.580	1.010	1.564	0.118	0.829
(1) Vowel winner, body winner, no conflict vs. (7) Vowel winner, body all pass (Adult)	0.602	0.490	1.229	0.219	0.646
(2) Vowel winner, body winner, conflict vs. (5) Vowel winner, body all fail (Adult)	3.107	1.024	3.033	0.002	0.957
(2) Vowel winner, body winner, conflict vs. (7) Vowel winner, body all pass (Adult)	2.129	0.521	4.083	<.001	0.894
(7) Vowel winner, body all pass vs. (5) Vowel winner, body all fail (Adult)	0.978	1.086	0.900	0.368	0.727

[8] Logarithm of the odds back-transformed to a probability value, indicating probability of the modelled event. For example, the likelihood of adults using the vowel winner pronunciation more often in condition 2 than condition 1 is 17.8% (i.e., it is quite unlikely that the vowel winner pronunciation will be used more often in condition 2 than in condition 1).

Table 3.2

Output from mixed-effects model comparing use of the vowel winner pronunciation in vowel winner conditions, using Child as the age group reference level
*glmer(Vowel Winner Score ~ Condition*Age + (1|Participant) + (1|Item))*

Fixed Effect	Est.	St. Error	*z* value	*p* value	Inverse Logit Probability
(1) Vowel winner, body winner, no conflict vs. (2) Vowel winner, body winner, conflict (Child)	-0.857	0.330	-2.596	0.009	0.298
(1) Vowel winner, body winner, no conflict vs. (5) Vowel winner, body all fail (Child)	0.135	0.880	0.154	0.878	0.534
(1) Vowel winner, body winner, no conflict vs. (7) Vowel winner, body all pass (Child)	0.202	0.476	0.426	0.670	0.550
(2) Vowel winner, body winner, conflict vs. (5) Vowel winner, body all fail (Child)	0.992	0.897	1.106	0.269	0.729
(2) Vowel winner, body winner, conflict vs. (7) Vowel winner, body all pass (Child)	1.060	0.508	2.085	0.037	0.743
(7) Vowel winner, body all pass vs. (5) Vowel winner, body all fail (Child)	-0.067	0.962	-0.070	0.944	0.483

Using the Condition x Age Group interaction as a fixed effect also allowed a comparison between adult and child participants' use of the vowel winner in each condition. It was hypothesised that children may use the vowel winner pronunciation more often than adults as a result of their systematic phonics instruction. Table 3.3 presents output of the model comparing use of the vowel winner by adults and children in these conditions (condition 5 is omitted as above). Results showed that there was no significant difference between adults' and children's use of the vowel winner in conditions 1 and 7, but adults' use of the vowel winner was significantly lower than children's in condition 2 (vowel winner, body winner, conflict). Possible reasons for this pattern of results will be discussed in *Section 3.5.1*.

Table 3.3

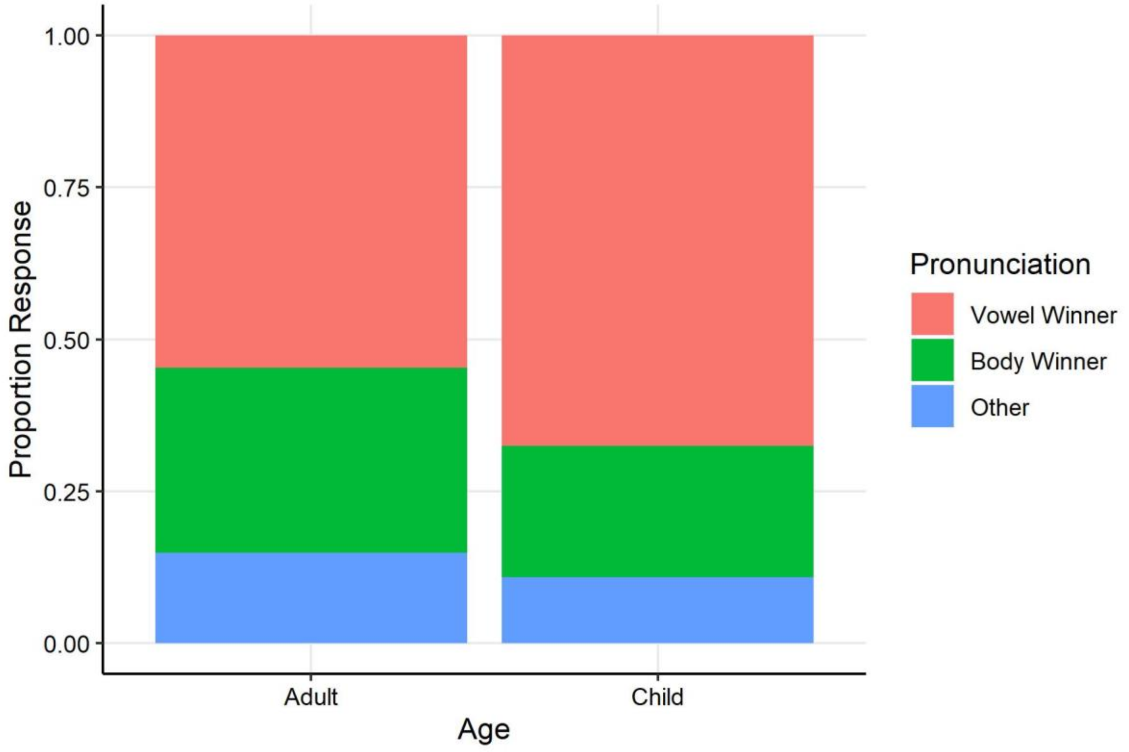
Output from mixed-effects model comparing adult and child participants' vowel winner use in conditions 1, 2 and 7
*glmer(Vowel Winner Score ~ Condition*Age + (1|Participant) + (1|Item))*

Fixed Effect	Estimate	St. Error	*z* value	*p* value	Inverse Logit (Probability)
(1) Vowel winner, body winner, no conflict (Adult vs. Child)	-0.017	0.180	-0.092	0.927	0.496
(2) Vowel winner, body winner, conflict (Adult vs. Child)	0.653	0.185	3.528	<.001	0.658
(7) Vowel winner, body all fail (Adult vs. Child)	-0.416	0.281	-1.476	0.140	0.397

Previous results suggested that - contrary to the TP prediction – both adult and child participants used the vowel winner less often when nonwords contained a conflicting body winner (condition 2). Crucially, this condition provides the only opportunity to assess which pronunciation participants use for items that have conflicting vowel winner and body winner pronunciations. To analyse this behaviour, pronunciation responses in this condition were categorised according to whether participants used the vowel winner pronunciation, the body winner pronunciation, or any other pronunciation. Figure 3.2 displays the percentage of adults' and children's responses using each of these pronunciations.

Figure 3.2

Proportion of Responses Using the Vowel Winner Pronunciation, Body Winner Pronunciation and Other Pronunciations by Adults and Children in the Vowel Winner, Body Winner, Conflict Condition (2).



The TP predicts that participants' choice of pronunciation ("vowel winner", "body winner" or "other") should not be evenly distributed across the three categories, as the vowel winner should be used more frequently than the body winner or other pronunciations. The results of chi-square goodness-of-fit tests were significant both for adults (χ^2(2, *n* = 720) = 173.73), *p* < .001) and for children (χ^2(2, *n* = 865) = 475.54), *p* < .001), suggesting an uneven distribution across the pronunciation categories. Paired-sample t-tests with a Bonferroni adjusted alpha level of .025(.05/2) confirmed that adults used the vowel winner more often than the body winner (*t*(29) = 3.219, *p* = .006) and other pronunciations (*t*(29) = 5.792, *p* < .001). Similarly, children used the vowel winner more often than the body winner (*t*(29) = 9.101, *p* < .001) and other pronunciations (*t*(29) = 18.075, *p* < .001). Whilst a substantial percentage of pronunciations in this condition do use the conflicting body winner, this analysis shows that the vowel winner is still the most frequent pronunciation for both adults and children.

3.4.1.2 Is use of the body winner modulated by the vowel winner?

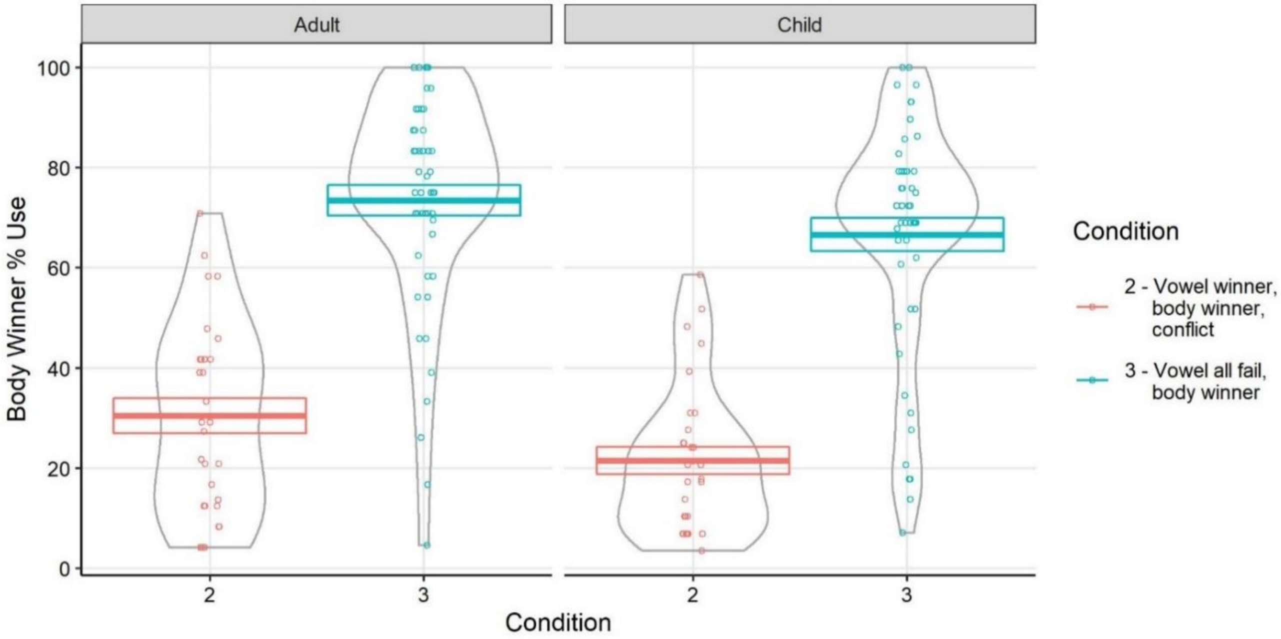
When applied as an account of word reading, the TP can incorporate information from the orthographic level of the word body as well as the vowel grapheme, and predicts when readers will use this information in nonword pronunciations. Specifically, the TP predicts that use of the "body winner" (a single pronunciation of a word body that passes the tolerance test) should be modulated by the presence of a vowel winner: the body winner should only be used when a nonword has no vowel winner pronunciation available, as more specific rules should only be sought if a more general rule does not pass the tolerance test.

Evaluating the TP's predicted use of the body winner pronunciation involved the vowel winner, body winner, conflict condition (2) and the vowel all fail, body winner condition (3). The TP predicts use of the body winner to vary across these conditions, such that participants should use the body winner to pronounce items in condition 3, where all vowel pronunciations fail the tolerance test, and should *not* use the body winner to pronounce items in condition 2, where there is a competing vowel winner. Figure 3.3 displays adult and child participants' use of the body winner for items in these conditions.

A mixed-effects model using a Condition x Age Group interaction did not explain significantly more data variance than a reduced model using Condition and Age Group as fixed effects (($\chi^2(1) = 0.195$, $p = .659$), suggesting that adults' use of the body winner did not differ from children's as a function of condition. A maximal mixed-effects logistic regression model failed to converge, but Table 3.4 presents the output of the reduced model using Condition and Age Group as fixed effects, random slopes and intercepts for participant, and random intercepts for item. Results suggest that adult and child participants' use of the body winner pronunciation was significantly higher in the vowel all fail, body winner condition (3) than in the vowel winner, body winner, conflict condition (2). This accords with the TP prediction, as use of the body winner seems to be modulated by the presence of a vowel winner: if an item has no vowel winner then the body winner will be used, but if it does have a vowel winner then the body winner will not be used. A significant effect of participant group suggests that children's use of the body winner across these conditions was significantly lower than adults'.

Figure 3.3

Adult and Child Participants' Use (%) of the Body Winner Pronunciation in the Vowel Winner, Body Winner, Conflict Condition (2) and the Vowel All fail, Body Winner Condition (3), by Item.



Note. Data points represent individual nonword items.

Table 3.4

Output from mixed-effects model comparing use of the body winner pronunciation in the vowel winner, body winner, conflict condition (2) and the vowel all fail, body winner condition (3) by adults and children
glmer(Body Winner Score ~ Condition + Age + (1+Condition|Participant) + (1|Item))

Fixed Effect	Est.	St. Error	*z* value	*p* value	Inverse Logit (Probability)
(2) Vowel winner, body winner, conflict vs. (3) Vowel all fail, body winner	2.558	0.290	8.828	<.001	0.928
Participant Group (Adult vs. Child)	-0.443	0.208	-2.129	0.033	0.391

3.4.2 What do participants say when the TP provides no predicted pronunciation?

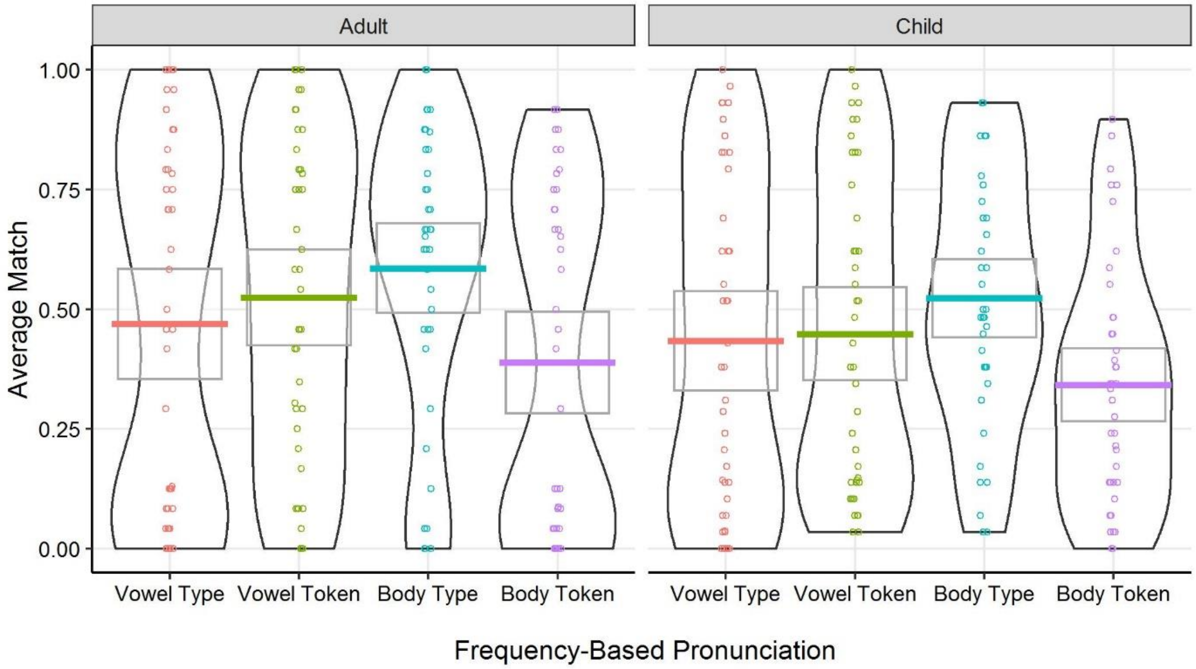
42 nonword items from conditions 4 (vowel all fail, body all fail) and 6 (vowel all fail, body all pass) have no single predicted pronunciation provided by the TP. An exploratory analysis investigated the possible pronunciations participants could use to read these items aloud, based on different statistical properties of English words.

Participants' pronunciation responses were assessed by their match to possible pronunciations based on four different spelling-sound frequencies in English words: the most common pronunciation of the vowel grapheme by type; the most common pronunciation of the vowel grapheme by token; the most common pronunciation of the word body by type; and the most common pronunciation of the word body by token. The type and token frequencies of vowel graphemes and word bodies in English monosyllabic words were taken from the CELEX corpus (Baayen et al., 1995). It should be noted that it is possible for multiple frequency measures to offer the same pronunciation for a letter sequence. For example, the most common pronunciation of the letter sequence OW is the same when measured by vowel type frequency, vowel token frequency, and body type frequency.

Figure 3.4 displays the average match between adult and child participant responses and the pronunciations provided by the four corpus frequency-based measures for items in conditions 4 and 6. This shows that participants do make use of information from all four frequency counts in their pronunciations, and that the most frequent pronunciation of the body by type seems to be the pronunciation most often used by participants. However, this should be interpreted with caution, as this frequency count can offer more than one pronunciation: for some letter sequences there are different pronunciations of the body that occur in the same number of word types in the corpus, thus inflating the likelihood that a participant's response will match the pronunciation offered by this frequency measure.

Figure 3.4

Adult and Child Participants' Average Match between Pronunciation Response and Corpus Frequency-Based Pronunciation for 42 Nonword Items in Conditions 4 and 6



Note. Data points represent individual nonword items. Box around mean represents 95% confidence interval.

This descriptive analysis shows that participants used pronunciations based on all four corpus frequency-based measures for items without a TP prediction. However, this group-level analysis may mask the behaviour of individual participants. For example, an individual participant may use a combination of all frequency measures in their pronunciations, or they may favour some over others. To investigate this further, Figures 3.5 and 3.6 present the data according to the behaviour of individual adult and child participants. Here can be observed the average match between the response and pronunciation according to frequency measure by each participant, revealing that all participants used a combination of information from the four frequency measures in their pronunciations, following a relatively similar pattern of behaviour.

Figure 3.5

Individual Adult Participants' Average Match Between Pronunciation Response and Corpus Frequency-based Pronunciation for 42 Nonword Items in Conditions 4 (Vowel All Fail, Body All Fail) and 6 (Vowel All Fail, Body All Pass)

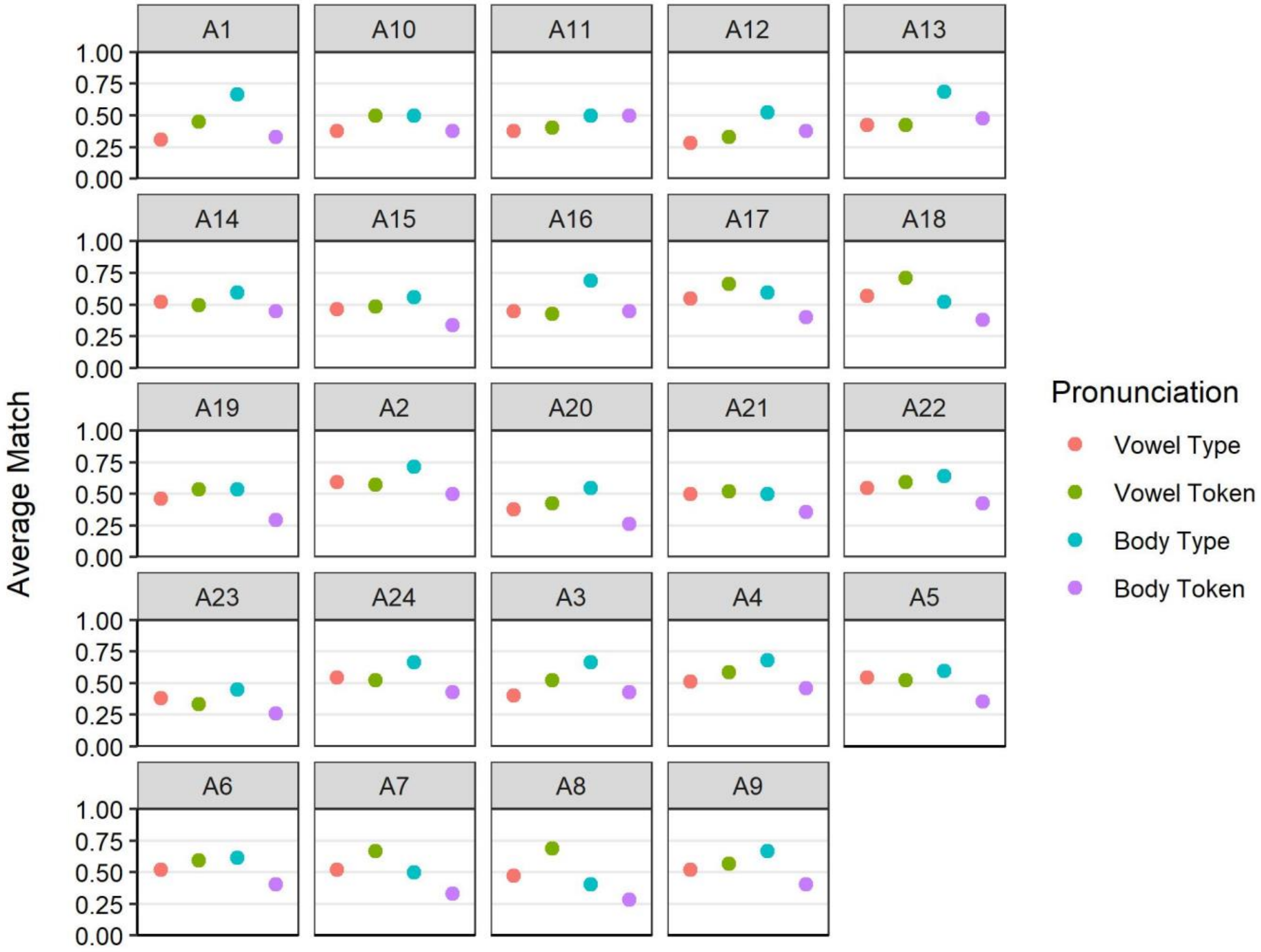
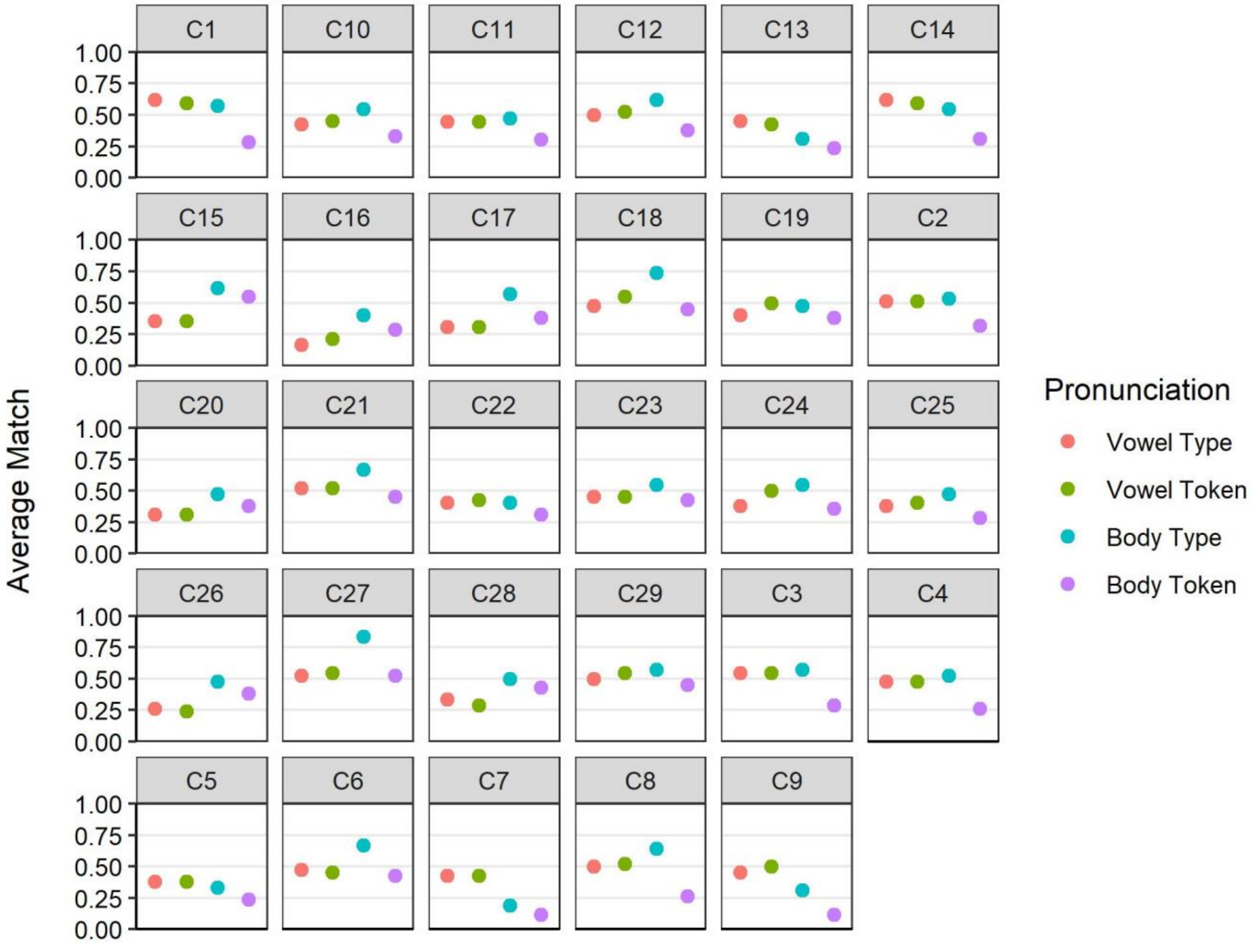


Figure 3.6

Individual Child Participants' Average Match Between Pronunciation Response and Corpus Frequency-based Pronunciation for 42 Nonword Items in Conditions 4 (Vowel All Fail, Body All Fail) and 6 (Vowel All Fail, Body All Pass)



3.4.3 Can the TP predict nonword pronunciations more successfully than three computational models of reading?

In addition to the TP, I analysed the predicted pronunciations of three computational models of reading: the DRC (Coltheart et al., 2001), the CDP+ (Perry et al., 2007) and the Triangle model (Chang et al., 2019). Mixed-effects logistic regression models were used to compare the match between the model prediction and participant response firstly across all items, secondly for items with a vowel winner pronunciation (from conditions 1, 2, 5 and 7), and thirdly for items whose vowel grapheme fails the tolerance test (from condition 3).

3.4.3.1 Model comparisons

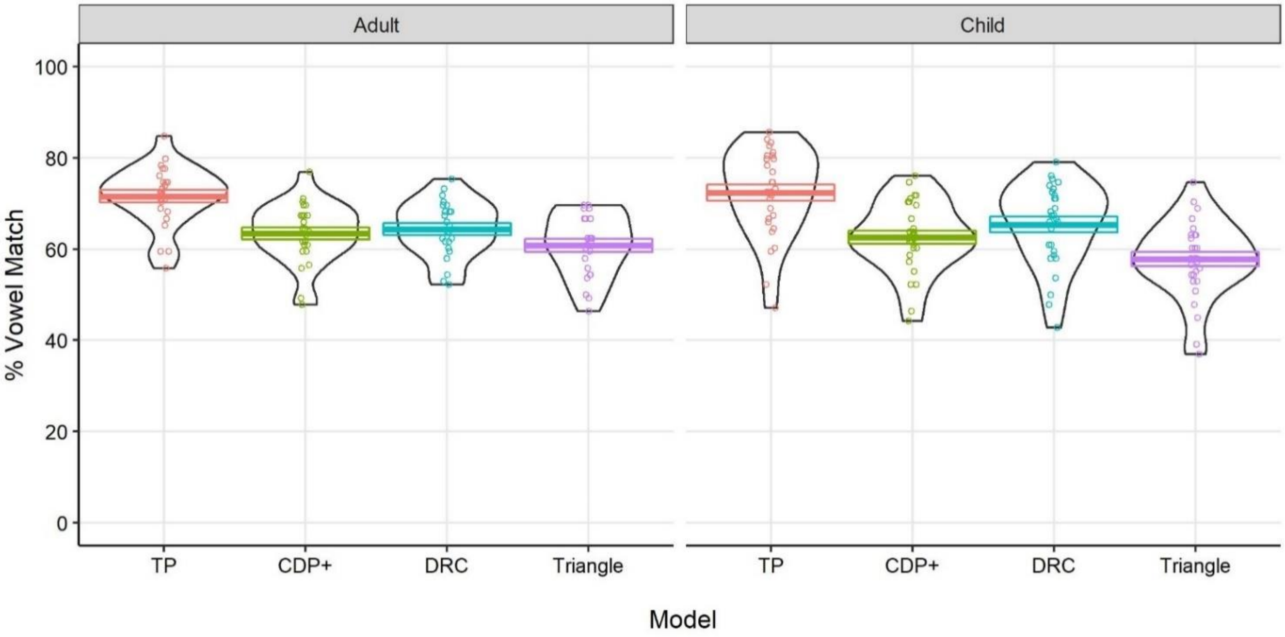
Taking a participant-level approach, the model comparison began by investigating the similarity between participants' responses and pronunciations predicted by the TP. This analysis involved 156 items with a TP prediction; the 42 items from conditions 4 and 6 were removed as they do not have a single pronunciation passing the tolerance test and therefore have no predicted pronunciation. In terms of the percentage of adult participants' pronunciations that matched the TP predicted pronunciation, the lowest level of similarity between the TP and an individual participant was 56.41% and the highest was 87.01%; the median was 74.19% and the mean was 72.85% (SE = 1.41%). For child participants, the lowest level of similarity was 47.01% and the highest was 85.51%; the median was 73.19% and the mean was 72.39% (SE = 1.76%).

Next, I examined the predictions made by the three established computational models of reading cited above, to place these results in the context of existing literature and assess the TP's relative performance. The analysis used 138 nonword items. In addition to the 42 items removed from the original data set for which TP could provide no prediction (from conditions 4 and 6), 18 items were removed because the Triangle model predicted use of a rhotic vowel in the pronunciation. This is because the version of the Triangle model used (Chang et al., 2019) uses US pronunciations in the training set, meaning that these rhotic predictions are incompatible with our participant responses and corpus frequencies, which use Southern British English pronunciations.

Throughout the model comparison, the match between predicted pronunciation and participant response was scored according to the pronunciation of the vowel grapheme. Figure 3.7 displays for adult and child participants the percentage of vowel pronunciations that matched the predictions made by the TP and the three reading models. The match between participants' responses and the DRC prediction can also be interpreted as participants' rate of vowel regularisation (i.e., use of the most common pronunciation of the vowel grapheme by type), as the DRC prediction for nonword pronunciations is always a regular response. By observation, it seems that the behaviour of participants is closer to the predictions of the TP than to the DRC, CDP+ or Triangle models. This indication is supported by the results of a mixed-effects logistic regression model using a Model x Age Group interaction. This model was a significantly better fit to the data than a reduced model using Model and Age Group as fixed effects (($\chi^2(3)$ = 12.233, *p* = .007); thus, adult and child participant groups were treated separately. As a maximal mixed-effects logistic regression model failed to converge, I ran an intercepts-only model with the Model x Age Group interaction as the fixed effect (rotating models of reading and age groups as the reference levels), and item and subject random intercepts. Table 3.5 presents the output from the model using adult as the age group reference level. Results indicate that the TP is a significantly better match for adult participants' responses than the three reading models, and the Triangle model performs significantly worse than all other models. There is no significant difference between performance of the CDP+ and DRC. The results of the mixed-effects model using child as the age group reference level is presented in Table 3.6. This analysis suggests that the TP is a significantly better match for child participants' responses than the three reading models, the Triangle model is a worse match than the other models, and the DRC is a better match than the CDP+ model.

Figure 3.7

The Percentage of Vowel Pronunciations Produced by 24 Adult Participants and 29 Child Participants that Matched the Predictions Made by Each Model of Reading



Note. Data points represent individual participants.

Table 3.5

Output from mixed-effects model comparing reading model performance, using adult as age group reference level

*glmer(Score ~ Model*Age + (1|Participant) + (1|Item), family =binomial)*

Comparison	Estimate	Std. Error	*z* value	*p* value	Inverse Logit (Probability)
TP vs. CDP+	-0.516	0.062	-8.328	<.001	0.374
TP vs. DRC	-0.453	0.062	-7.291	<.001	0.389
TP vs. Triangle	-0.671	0.062	-10.873	<.001	0.338
CDP+ vs. DRC	0.064	0.060	1.094	0.292	0.516
CDP+ vs. Triangle	-0.155	0.060	-2.589	0.010	0.461
Triangle vs. DRC	0.219	0.060	3.642	<.001	0.555

Table 3.6

Output from mixed-effects logistic regression comparing reading model performance, using child as age group reference level
*glmer(Score ~ Model*Age + (1|Participant) + (1|Item), family =binomial*

Comparison	Estimate	Std. Error	*z* value	*p* value	Inverse Logit (Probability)
TP vs. CDP+	-0.635	0.057	-11.130	<.001	0.346
TP vs. DRC	-0.461	0.057	-8.035	<.001	0.387
TP vs. Triangle	-0.920	0.058	-16.182	<.001	0.285
CDP+ vs. DRC	0.174	0.055	3.145	0.002	0.543
CDP+ vs. Triangle	-0.284	0.055	-5.206	<.001	0.429
Triangle vs. DRC	0.459	0.055	8.327	<.001	0.613

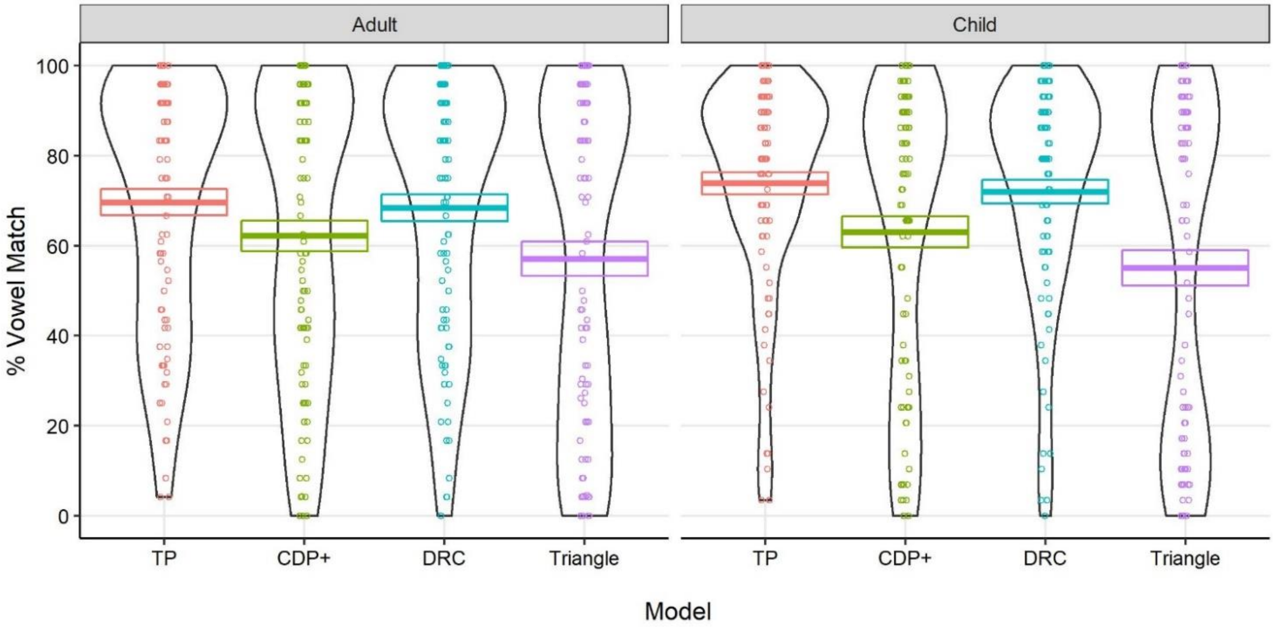
3.4.3.2 Where do the models differ in their performance?

To understand more about why the TP outperforms the other models in predicting nonword reading behaviour, the following analysis separates “vowel winner” items in conditions with a vowel winner (1, 2, 5 and 7) from “vowel fail” items in condition 3 where the vowel fails the tolerance test. Analysing the models’ performance for items with consistent and inconsistent vowels separately offers an insight into each model’s strengths and weaknesses in respect of the consistency of spelling-sound correspondences.

Figure 3.8 displays each reading model's average match between the predicted pronunciation and participants' responses for the 90 vowel winner items. To compare the models' performance on these items, I ran a mixed-effects logistic regression model with a Model of Reading x Age Group interaction, which explained significantly more variance in the data than a reduced model using Model of Reading and Age Group as fixed effects (($\chi^2(3) = 20.298$, $p < .001$). Therefore adult and child age groups are treated separately. As a maximal mixed-effects structure failed to converge, I ran the model using the Model of Reading x Age Group interaction as a fixed effect (with TP as the model reference level and rotating the age group reference levels) with random intercepts for participant, and random slopes and intercepts for item. Tables 3.7 and 3.8 present the output of the model for adult and child age groups. For both adults and children, results indicated no significant difference between the performance of the TP and the DRC, which is expected as both models predict use of the most common vowel pronunciation type for these items. The CDP+ and Triangle models both perform significantly worse than the TP.

Figure 3.8

Match (%) Between Adult and Child Participant Response and Model Prediction for Pronunciation of the Vowel Grapheme in 90 Vowel Winner Items



Note. Data points represent individual nonword items.

Table 3.7

*Output from mixed-effects model comparing reading model performance for vowel winner items, using adult as the age group reference level and TP as the model reference level glmer(Score ~ Model*Age + (1|Participant) + (1+Age|Item), family =binomial)*

Comparison	Estimate	St. Error	*z* value	*p* value	Inverse Logit (probability)
TP vs. CDP+ (Adult)	-0.489	0.078	-6.251	<.001	0.380
TP vs. DRC (Adult)	-0.077	0.079	-0.973	.330	0.481
TP vs. Triangle (Adult)	-0.807	0.078	-10.309	<.001	0.308

Table 3.8

Output from mixed-effects model comparing reading model performance for vowel winner items, using child as the age group reference level and TP as the model reference level
*glmer(Score ~ Model*Age + (1|Participant) + (1+Age|Item), family =binomial)*

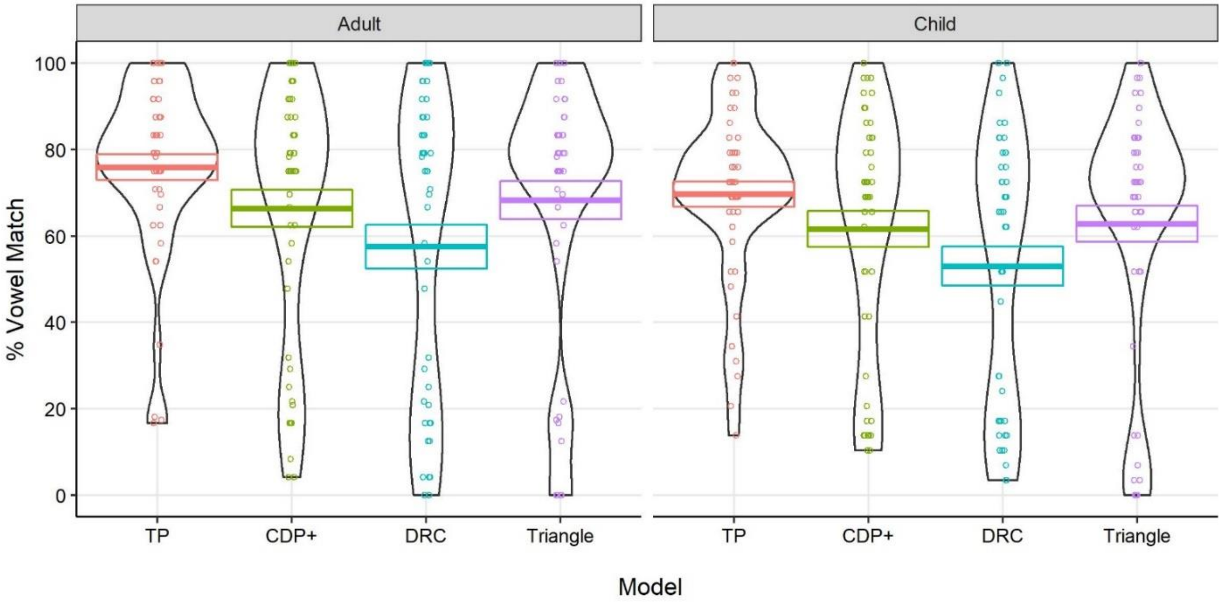
Comparison	Estimate	St. Error	*z* value	*p* value	Inverse Logit (probability)
TP vs. CDP+ (Child)	-0.738	0.073	-10.117	<.001	0.323
TP vs. DRC (Child)	-0.137	0.075	-1.842	0.066	0.466
TP vs. Triangle (Child)	-1.233	0.073	-16.869	<.001	0.226

Figure 3.9 displays each model's average match between the predicted pronunciation and the participants' responses for the 48 vowel fail items. A mixed-effects model using a Model of Reading x Age Group interaction did not explain significantly more variance in the data than a reduced model using Model of Reading and Age Group as fixed effects (($\chi^2(3) = 0.937$, $p = .817$). The maximal mixed-effects logistic regression model failed to converge; thus, Table 9 presents the output of the reduced model using Model of Reading (with TP as the reference level) and Age Group as fixed effects, with random intercepts for participant, and random slopes and intercepts for item. Results suggest that for adult and child participants, the TP performs significantly better than all other models for vowel fail items. There was no significant main effect of age group. Together, this vowel winner vs. vowel fail analysis reveals that whilst the TP is no better than the DRC at predicting pronunciations for items with consistent vowels (i.e., vowel winner items), it is more successful than all other models at predicting pronunciations for

items with inconsistent vowels (i.e., vowel fail items), thus offering the strongest account overall. Possible reasons for the TP's particular success in predicting inconsistent vowel pronunciations will be explored in *Section 3.5.3*.

Figure 3.9

Match (%) Between Adult and Child Participant Response and Model Prediction for Pronunciation of the Vowel Grapheme in 48 Vowel Fail Items.



Note. Data points represent individual nonword items.

Table 3.9

Output from mixed-effects model comparing reading model performance for vowel fail items by adults and children, using TP as the model reference level
glmer(Score ~ Model + Age + (1|Participant) + (1+Age|Item), family =binomial)

Comparison	Estimate	St. Error	*z* value	*p* value	Inverse Logit (probability)
Intercept	1.660	0.283	5.857	<.001	0.840
TP vs. CDP+	-0.577	0.072	-7.973	<.001	0.360
TP vs. DRC	-1.098	0.072	-15.246	<.001	0.250
TP vs. Triangle	-0.480	0.073	-6.623	<.001	0.382
Adult vs. Child	-0.364	0.225	1.615	0.106	0.410

To complete the picture of each model's performance, I identified items for which the models performed particularly poorly, revealing how the models deviate most dramatically from participants' behaviour. Tables 3.10 and 3.11 present, for each model, the nonword items for which predicted pronunciations were used in less than 5% of adult and child participant responses. For the DRC, the majority of items with predicted pronunciations used rarely by participants are from the vowel fail, body winner condition (3). These items have an inconsistent vowel pronunciation (the vowel fails the tolerance test), yet the DRC predicts use of the most common grapheme-phoneme correspondence for the vowel (as for all nonword items). However, participants seem to avoid using this correspondence for these condition 3 items.

The TP makes one predicted pronunciation which is not used by any adult participants and two that are not used by any child participants. These are use of the most frequent vowel grapheme-phoneme correspondence in the items THEIL, CREIL, and CHEIL from the vowel winner, body winner, no conflict condition (1).

The CDP+ model makes a number of unsuccessful GPC predictions for items from the vowel fail, body winner condition (3); like the DRC, it predicts use of this vowel pronunciation despite the inconsistency of the vowel grapheme. Both the CDP+ and the Triangle model predict body analogies for some items in vowel winner conditions (in which the vowel GPC is relatively consistent) which are not used by participants. Both connectionist models also predict lexicalisations of some nonwords which participants did not make. In addition, the Triangle model makes a number of erroneous predictions produced by no participants.

Table 3.10

Pronunciations predicted by models which are used less than 5% of the time by adult participants. Predictions are transcribed using symbols representing the DRC's phonemic vocabulary (see Appendix C).

Model	Nonword	Condition	Prediction	Use by Participants	Description
CDP+	THEIL	1	DEl	0.000	n/a
CDP+	KAID	1	k{d	0.000	Lexicalisation
CDP+	SMEAD	1	smEd	0.042	Body analogy
CDP+	PRIELD	3	pr2ld	0.042	GPC
CDP+	BOUP	3	b6p	0.042	GPC
CDP+	FRIMB	7	frQm	0.000	Lexicalisation
DRC	YOOT	1	wt	0.000	n/a
DRC	THEIL	1	T1l	0.042	GPC
DRC	MIEF	3	mif	0.000	GPC
DRC	PRIEF	3	pr2f	0.000	GPC
DRC	ZIELD	3	z2ld	0.000	GPC
DRC	PRIELD	3	pr2ld	0.042	GPC
DRC	BOUP	3	b6p	0.042	GPC
DRC	FROUP	3	fr6p	0.042	GPC
TP	THEIL	1	T1l	0.042	GPC
Triangle	THEIL	1	T1l	0.042	GPC
Triangle	PLINT	1	pl2nt	0.000	Body analogy
Triangle	DRAUNCH	1	dr{nJ	0.000	n/a
Triangle	VAID	1	vEd	0.000	Body analogy
Triangle	THAID	1	TQd	0.000	n/a
Triangle	VEIGHT	1	vEt	0.000	Lexicalisation
Triangle	DREIGHT	1	drEt	0.000	n/a
Triangle	GEANT	2	_Int	0.042	n/a
Triangle	HIEF	3	hQf	0.000	n/a

Triangle	JIEK	3	_Qk	0.000	Lexicalisation
Triangle	CHOUP	3	J{p	0.000	Lexicalisation
Triangle	FROUP	3	frIp	0.000	n/a
Triangle	SILD	7	s2ld	0.042	Body analogy

Table 3.11

Pronunciations predicted by models which are used less than 5% of the time by child participants. Predictions are transcribed using symbols representing the DRC's phonemic vocabulary (see Appendix C)

Model	Nonword	Condition	Prediction	Use by participants	Description
CDP+	FRIMB	7	frQm	0.000	Lexicalisation
CDP+	KAID	1	k{d	0.000	Lexicalisation
CDP+	NOVE	1	nVv	0.000	Body analogy
CDP+	CHEIL	1	J1l	0.034	GPC
CDP+	CREIL	1	kr1l	0.034	GPC
CDP+	THEIL	1	DEl	0.034	n/a
DRC	YOOT	1	wt	0.000	n/a
DRC	BIELD	3	b2ld	0.034	GPC
DRC	HIEF	3	h2f	0.034	GPC
DRC	ZIELD	3	z2ld	0.034	GPC
DRC	CHEIL	1	J1l	0.034	GPC
DRC	CREIL	1	kr1l	0.034	GPC
TP	CHEIL	1	J1l	0.034	GPC
TP	CREIL	1	kr1l	0.034	GPC
Triangle	DREIGHT	1	drEt	0.000	n/a
Triangle	NOVE	1	nVv	0.000	Body analogy
Triangle	THAID	1	TQd	0.000	n/a
Triangle	VAID	1	vEd	0.000	Body analogy
Triangle	VEIGHT	1	vEt	0.000	Lexicalisation
Triangle	CHOUP	3	J{p	0.000	Lexicalisation
Triangle	HIEF	3	hQf	0.000	n/a
Triangle	JIEK	3	_Qk	0.000	Lexicalisation
Triangle	CHEIL	1	J1l	0.034	GPC
Triangle	CREIL	1	kr1l	0.034	GPC

Triangle	GEANT	2	_Int	0.034	n/a
Triangle	BOUP	3	bUp	0.034	n/a
Triangle	FROUP	3	frIp	0.034	n/a

3.4.4 How is the TP's performance related to corpus consistency and participant variability?

These analyses suggest that the TP offers a more successful account of adult nonword reading aloud than three computational models, including those which use continuous, token frequency-weighted measures of spelling-sound consistency (i.e. the connectionist CDP+ and Triangle models). In contrast, the TP employs a type-based, categorical account of spelling-sound consistency which can apply recursively to involve multiple orthographic levels, and which I suggest may lie behind its success. I explored this approach further by investigating the relationship between the TP, item-based consistency (i.e. the spelling-sound consistency of vowel graphemes in the CELEX corpus), and participant-based consistency (i.e. the variability of pronunciation responses between participants). The aim was to determine whether the TP's type-based, categorical metric of consistency offers an improved account of reading behaviour beyond continuous measures of consistency based on either type or token corpus frequencies. If the TP metric can predict participants' pronunciations beyond these continuous measures, this result would indicate that the TP's particular approach to consistency lies behind its ability to capture nonword reading behaviour more successfully than the models using continuous measures, examined above.

Corpus type- and token-based spelling-sound consistency of the vowel grapheme and variability of participants' responses was quantified using a measure of entropy known as the H statistic (Shannon, 1949). An H value of 0 denotes no variability across items or participants, with a single pronunciation used for each item, whereas a higher H value represents more variability in the pronunciation across items or participants. This measure of variability is modulated by two factors, as described by Andrews and Scarratt (1998, p. 1061). The first factor is the number of different pronunciations an item or letter pattern has; H is higher when an item has many different pronunciations. The second factor is the number of exemplars of each pronunciation; H is higher when the number of exemplars of each pronunciation is equal, meaning that one pronunciation is not weighted more than another. The H value is calculated using the formula $\Sigma[-pi \times log2(pi)]$, where pi is the proportion of items or participants using a certain pronunciation (see also Andrews & Scarratt, 1998; Zevin & Seidenberg, 2006; Mousikou et al., 2017).

Three sets of analyses were used to investigate the relationships between item-based consistency, participant-based consistency and the TP in explaining nonword reading behaviour. The first set of analyses investigated the role of item-based consistency in participants' use of a particular pronunciation. It used mixed-effects logistic regression models to assess whether participants use the regular (most common) pronunciation of a vowel grapheme more often when the pronunciation of this grapheme is consistent in the corpus, and whether this relationship is modulated by the TP's own prediction for regularisation (using the recursive application of the type-based, categorical metric of consistency). Here, the TP would predict that there is a categorical, rather than continuous, effect of item-based consistency on participants' regularisation.

The second set of analyses investigated the role of item-based consistency on the variability of responses produced by participants. It used linear regression models to assess whether participants' responses are more variable if a nonword's vowel grapheme is inconsistent, and whether this relationship is modulated by the TP's categorical metric of consistency. The TP would predict that there is a categorical, rather than continuous, effect of item-based consistency on participants' variability.

The third set of analyses investigated the effect of age on the variability of responses produced by participants. It assessed whether the responses of child participants are less variable than those of adult participants as hypothesised according to the TP's role for individual experience in generalisation behaviour.

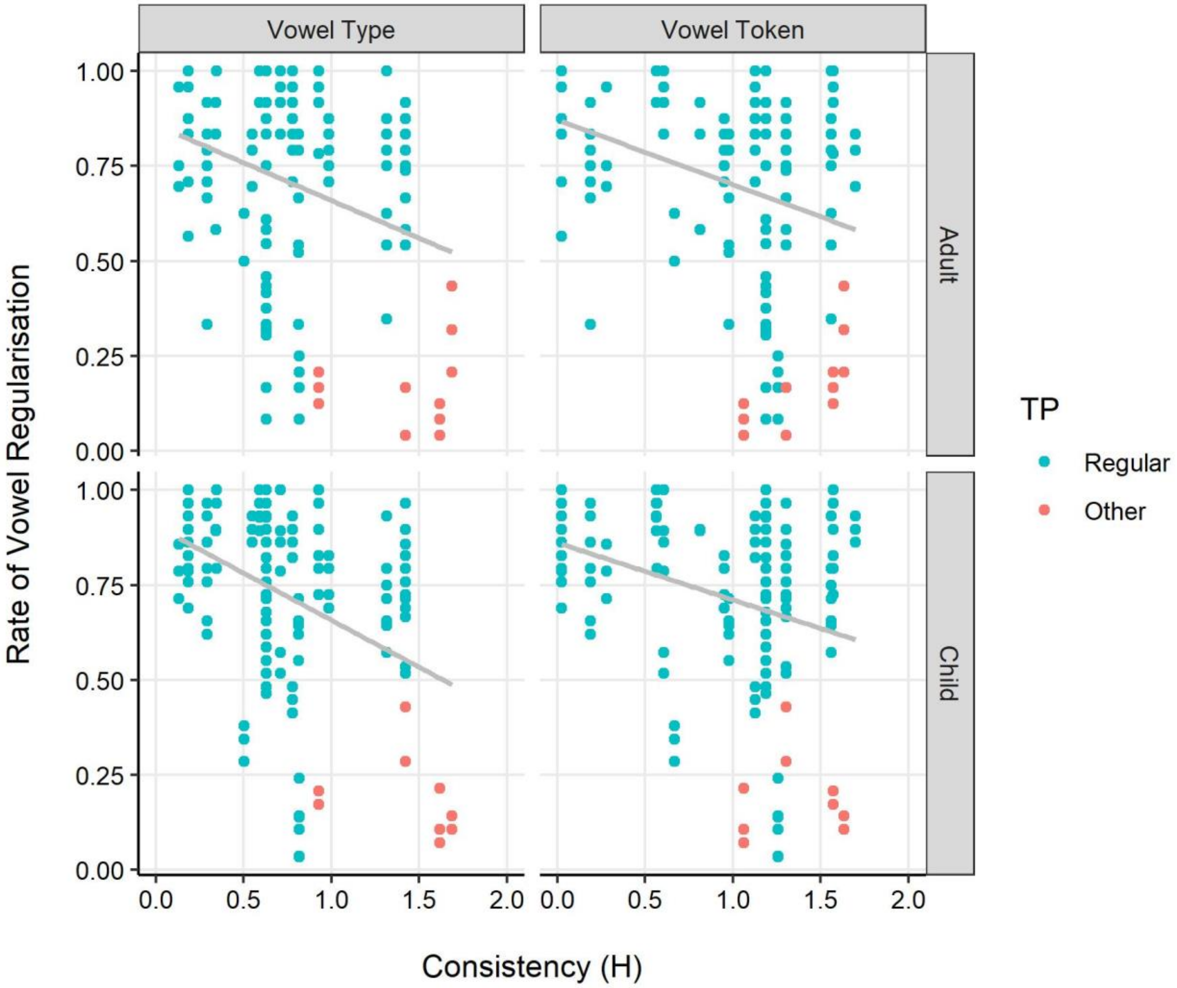
3.4.4.1 Is the relationship between vowel regularisation and corpus consistency of the vowel modulated by the TP?

Two continuous measures of consistency were considered in this analysis: the H value of the vowel grapheme calculated by type and by token. Type and token frequency values from the CELEX corpus (Baayen, et al. 1995) were used to produce H values representing the pronunciation consistency of each vowel grapheme and word body in the 156 items with a pronunciation predicted by the TP.

If participants pronounce nonwords according to a simple linear relationship between regularisation and consistency, then participants' regularisation of a vowel grapheme (or use of the most common pronunciation type, i.e. the vowel GPC) should increase as corpus consistency of the vowel grapheme increases. Adults' mean vowel regularisation for these items was 70.50% (*SE* = 2.26); children's mean vowel regularisation was 71.43% (*SE* = 1.99). Figure 3.10 plots participants' vowel regularisation by vowel type and token consistency (measured by H value) for 156 items. Note that a higher H value represents lower consistency. The figure shows that for both type and token consistency measures, there are low consistency items (high H values) that have high rates of regularisation, suggesting that there is not a simple linear relationship between consistency and regularisation. Additionally, each item is colour coded according to whether the TP predicts the regular vowel pronunciation (blue) or not (red), according to its own categorical measure of consistency (i.e., whether this particular pronunciation of the vowel is consistent enough to pass the tolerance test at either the vowel grapheme or body level). The coding shows that when participants regularised vowels with high corpus inconsistency, this corresponded with the TP's predicted regularisation for these items: the TP may modulate the relationship between regularisation and consistency.

Figure 3.10

Adult and Child Participants' Use of the Regular Vowel Type Pronunciation by Corpus Vowel Type and Token Consistency (Measured by H Value) for 156 Items



Note. Data points are colour coded by TP prediction for regularisation of the vowel and represent individual items.

The suggestion that the TP offers an improved account of vowel regularisation compared to continuous measures of consistency alone was supported by a series of mixed-effects logistic regression models. These investigated the effect of vowel consistency calculated by type and token on participants' vowel regularisation (use of the most common vowel pronunciation by type). A maximal mixed-effects model failed to converge, but Table 3.12 presents the output of a reduced model using vowel type consistency, vowel token consistency, and participant age (adult or child) as fixed effects, with random intercepts for participant. The results suggest that vowel type consistency but not vowel token consistency has a significant effect on participants' vowel regularisation. There was no significant difference between adults' and children's vowel regularisation.[9]

Table 3.12

Output from mixed-effects model comparing the effects of vowel type consistency, vowel token consistency and age on participants' vowel regularisation
glmer(Vowel Regularisation ~ Vowel Type Consistency + Vowel Token Consistency + Age + (1|Participant), family =binomial)

	Est.	St. error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	2.538	0.305	8.321	<.001	0.927
Vowel Type Consistency	-1.185	0.430	-2.756	.006	0.234
Vowel Token Consistency	-0.386	0.364	-1.062	.288	0.405
Age Group (Adult vs. Child)	0.079	0.152	0.515	.606	0.520

[9] An interaction between the consistency measures and participant age was not included in the model as VIF values above 5 suggested that there was multicollinearity between effects.

A second model added the binary variable of the TP's prediction for vowel regularisation as a fixed effect. Table 3.13 presents the model output, showing that the TP has a significant effect on vowel regularisation: participants are more likely to regularise the vowel in items which are predicted to be regularised by the TP. Meanwhile, vowel token consistency, but not vowel type consistency, also had a significant effect on participants' vowel regularisation.

Table 3.13

Output from mixed-effects model comparing the effects of vowel type consistency, vowel token consistency, TP regularisation prediction and age on participants' vowel regularisation glmer(Vowel Regularisation ~ Vowel Type Consistency + Vowel Token Consistency + Age + TP + (1|Participant), family =binomial)

	Estimate	St. error	*z* value	*p* value	Inverse Logit Probability
Intercept	0.730	0.385	1.896	.058	0.675
Vowel Type Consistency	-0.079	0.411	-0.192	.848	0.480
Vowel Token Consistency	-0.667	0.321	-2.078	.038	0.339
Age Group (Adult vs. Child)	0.079	0.153	0.518	.604	0.520
TP Vowel Regularisation	2.843	0.441	6.446	<.001	0.945

A chi-square test comparing these two models found that adding the TP significantly improved the model's fit to the data ($\chi^2(1) = 36.281$, $p < .001$). This indicates that the TP's type-based, categorical account of consistency based on multiple orthographic grain sizes is able to explain variance in participants' vowel regularisation behaviour that vowel consistency measured continuously by type and token frequency cannot.

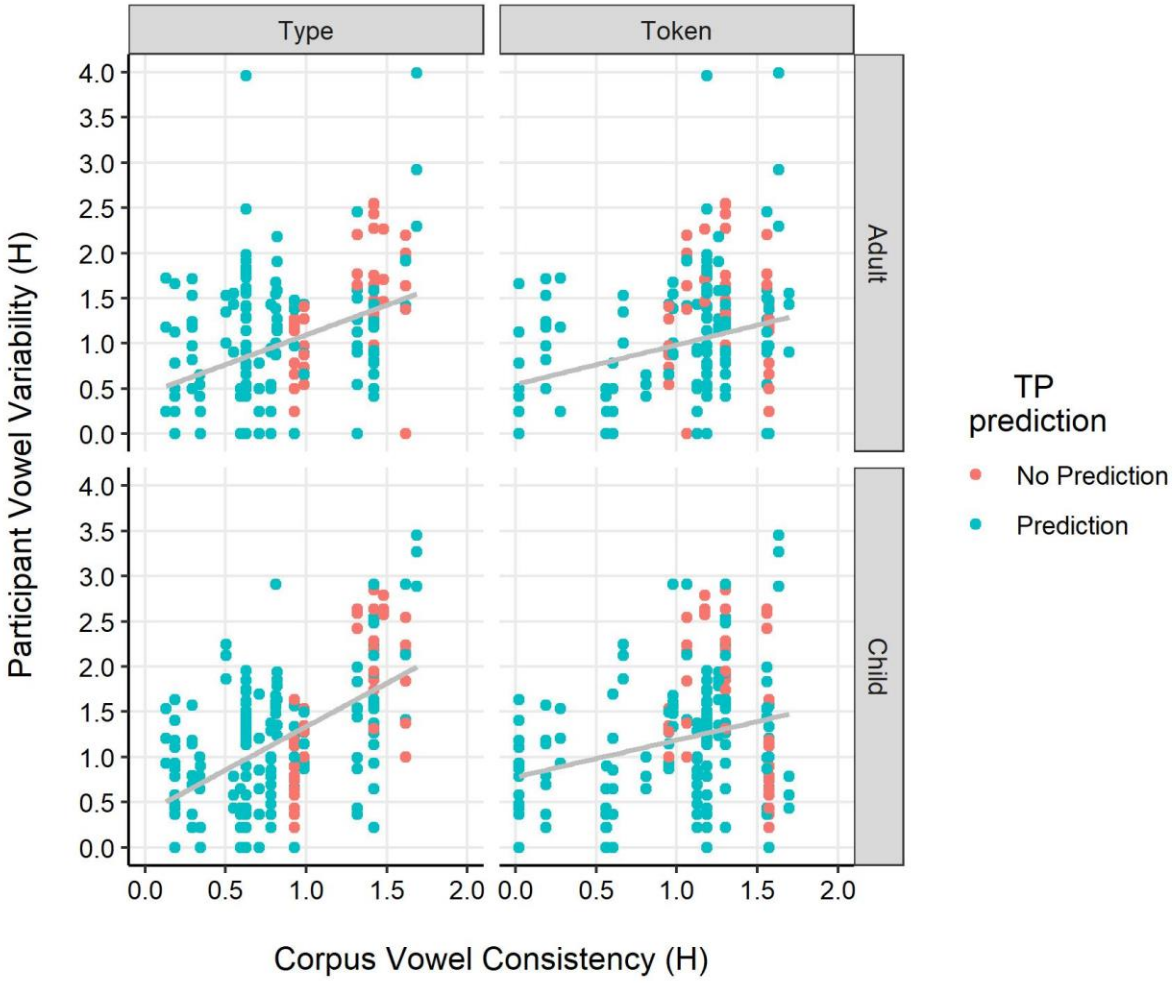
3.4.4.2 Is the relationship between variability in participants' vowel responses and corpus consistency of the vowel modulated by the TP?

The second set of consistency analyses investigated whether participants' vowel responses became more variable as the corpus type consistency of the vowel decreased, and whether this was modulated by the TP. The binary TP predictor used was whether or not the TP predicts a pronunciation for that nonword item (i.e., whether the nonword has a spelling-sound correspondence at the vowel grapheme or body level which is consistent enough to pass the tolerance test and should therefore be used productively by participants).

Figure 3.11 plots corpus vowel type and token consistency (measured by H value) by variability of adult and child participants' vowel pronunciations in nonword items (measured by H value). Note that a higher H value denotes lower consistency and higher variability. Each item is colour coded according to whether the nonword has a TP prediction (blue) or not (red). Table 3.14 presents the output of a linear regression model predicting variability of adult and child participants' vowel pronunciations based on corpus type and token consistency of the vowel, and participant age. Results suggest that corpus vowel type consistency predicts variability in participants' vowel pronunciations, but vowel token consistency does not. The effect of age on participants' variability was also significant, such that children's pronunciations of the vowel grapheme were more variable than adults'.

Figure 3.11

Variability in Adult and Child Participants' Vowel Pronunciations by Corpus Vowel Type and Token Consistency (Measured by H Value) for 198 Items.



Note. Items are colour coded according to whether or not the TP predicts a pronunciation for the nonword item

Table 3.14

Output from linear regression model comparing the effects of vowel type corpus consistency, vowel token corpus consistency and age on the variability of participants' pronunciation of the vowel grapheme

lm(Participant Vowel Variability ~ Vowel Type Consistency + Vowel Token Consistency + Age)

	Estimate	Standard error	*t* value	*p* value
Intercept	0.360	0.091	3.959	<.001
Vowel Type Consistency	0.903	0.106	8.524	<.001
Vowel Token Consistency	-0.126	0.095	-1.322	0.187
Age Group (Adult vs. Child)	0.201	0.066	3.033	0.003

I then considered whether an item having a pronunciation predicted by the TP (i.e. a pronunciation consistent enough to pass the tolerance test at either the vowel level, body level, or both) was also a predictor of participants' variability beyond continuous measures of vowel consistency calculated by type and token frequency. Table 3.15 presents the output of a model using vowel type consistency, vowel token consistency, participant age and the TP as predictors[10]: results indicate that vowel type consistency is a significant predictor of variability in participants' vowel pronunciations, but vowel token consistency and having a TP predicted pronunciation are not. The result of a chi-square test comparing the models confirmed that

[10] An interaction between the consistency measures and participant age was not included as VIF values above 5 suggested that there was multicollinearity between effects.

adding the TP as a predictor did not significantly improve the model's fit to the data (($\chi^2(1)$ = 0.218, *p* = .640). Together, these results indicated that as the corpus type consistency of the vowel increases, participants become less variable in their pronunciations of the vowel grapheme, and this relationship is not modulated by having a pronunciation predicted by the TP at the vowel or body level.

Table 3.15

Output from linear regression model comparing the effects of vowel type corpus consistency, vowel token corpus consistency, age, and having a TP predicted pronunciation on the variability of participants' pronunciation of the vowel grapheme lm(Participant Vowel Variability ~ Vowel Type Consistency + Vowel Token Consistency + Age + TP)

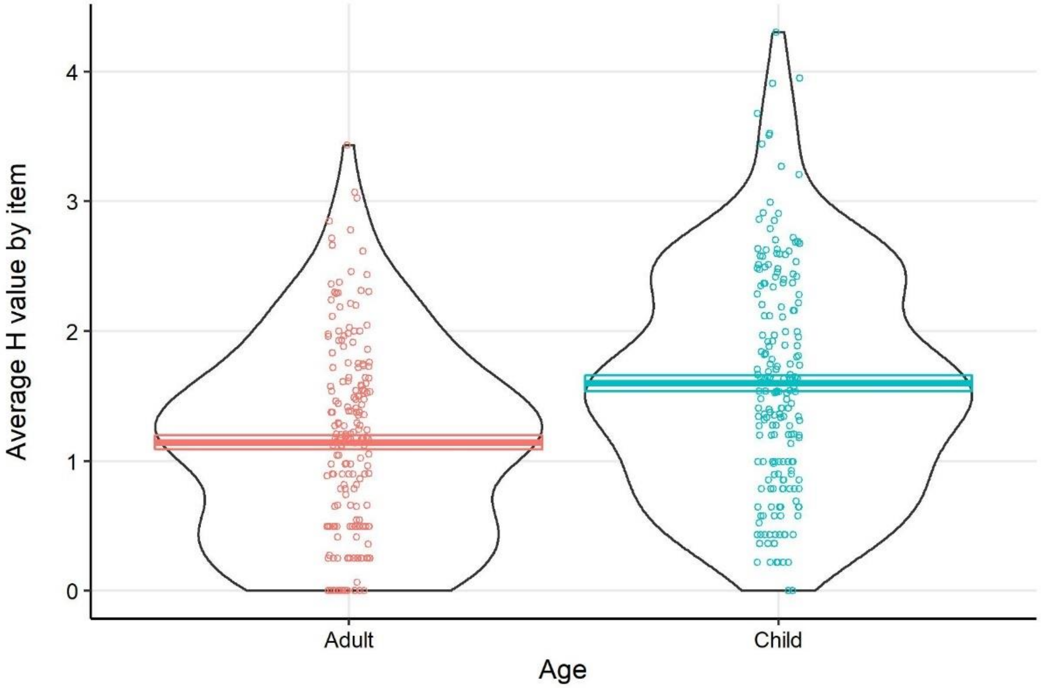
	Estimate	Standard error	*z* value	*p* value
Intercept	0.408	0.137	2.986	.003
Vowel Type Consistency	0.887	0.112	7.955	<.001
Vowel Token Consistency	-0.127	0.096	-1.329	.185
Age Group (Adult vs. Child)	0.201	0.066	3.030	.003
TP Prediction	-0.042	0.089	-0.468	.640

3.4.4.3 Are responses produced by children less variable than responses produced by adults?

According to the TP, an individual's productive rule system is based on the language input they have specifically received. It is likely that adult participants have received much more varied reading input and instruction than child participants, who have received only a few years' reading experience in a primary school context. Therefore, I hypothesised that there may be a greater difference between adults' individual rule systems than between children's, giving rise to more varied responses to nonword items by adults than by children. Figure 3.12 displays the average H value of whole-word responses to 198 items by adults and children, with a higher H value indicating more varied responses by participants. A paired-sample t-test found that responses by children had a significantly higher H value than those by adults (t(197) = 9.506, p < .001), suggesting that, contrary to expectation, children provided more varied responses to the stimuli than adults. Possible reasons for this will be explored in *Section 3.5.4.3*.

Figure 3.12

Average H Value of Responses for 198 Items by Adult and Child Participants



Note. Error bands represent standard error.

3.5 Discussion

For decades, debate has questioned whether readers use categorical rules involving individual grapheme-phoneme correspondences, or graded information about the consistency of larger orthographic units, when reading words aloud. In particular, research has focused on how skilled and developing readers generalise this orthography-phonology knowledge to read new words they have not encountered before (e.g. Glushko, 1979; Marsh et al., 1981; Ryder and Pearson, 1980; Treiman et al., 1990; Coltheart & Leahy, 1992; Andrews & Scarratt, 1998; Brown & Deavers, 1999; Treiman et al., 2003; Steacy et al. 2019), but no firm conclusions have been reached. In Experiment 1, I investigated whether a new account of productivity could answer these longstanding questions and bridge the gap between previous, opposing approaches.

First developed as a theory of linguistic generalisation, here I applied the Tolerance Principle (Yang, 2016) to the field of reading. Twenty-four adults and twenty-nine children aged 8-9 years read aloud 198 monosyllabic nonword items written in English orthography. I used the tolerance algorithm (2016, p. 9) to assess which spelling-sound correspondences were sufficiently consistent to pass the tolerance test, and thus predict which correspondences readers should use productively to pronounce the nonwords. By analysing participants' responses, I was able to assess the TP's ability to predict nonword pronunciations, as well as that of three models of reading: the DRC (Coltheart et al., 2001), the Triangle model (using the Chang et al., (2019) version of the Harm & Seidenberg (2004) model) and the CDP+ (Perry et al., 2007). Further, I investigated the role of consistency in the TP's account of nonword reading and its relative success in comparison to the computational models.

3.5.1 Can the TP predict skilled and developing readers' nonword pronunciations at the vowel grapheme and word body level?

The TP predicts that if the pronunciation of a vowel grapheme is sufficiently consistent to pass the tolerance test (a "vowel winner" pronunciation), participants should use this productively to read aloud a nonword item. This should be the case regardless of the consistency of the word body; a productive pronunciation for a vowel grapheme is a more general rule, so the search for a more specific pronunciation rule is redundant. Adult and child participants did use this pronunciation at a high rate for items with a vowel winner (72.22% and 74.84% respectively). However, contrary to the TP's prediction, use of the vowel winner in these items was modulated by condition: vowel winner use was significantly lower in condition 2 (vowel winner, body winner, conflict) than in the other conditions. In this condition, items have both a consistent vowel pronunciation and a conflicting consistent body pronunciation that pass the tolerance test. Use of the body winner pronunciation by adults in 30.5% of these items and by children in 21.5% suggests that there is some interference from orthographic information at the level of the word body, and that generalisation behaviour is not as categorical as the TP predicts. Nevertheless, the vowel winner remained the most commonly used pronunciation for items in this condition by both children and adults, as predicted by the TP.

Children’s use of the vowel winner was significantly higher than that of adults in these items. One explanation for this is that the explicit instruction of vowel GPCs that children receive within the synthetic phonics programme helps to override interference from a conflicting signal from the word body unit. Indeed, Thompson et al. (2009) found long-lasting effects of phonics instruction in adults who used more GPCs in nonword pronunciations than adults who had not been taught to read using this method. It is also possible that the body unit gives less interference for developing learners regardless of instruction, given earlier findings that early readers rely primarily on GPCs (Treiman et al., 1990) and that use of the body increases throughout development (Laxon et al., 1991). It is notable that children’s use of the vowel winner was not significantly different from adults’ in other vowel winner conditions, despite their rigorous instruction in phonics.

Beyond the vowel grapheme, the TP is also able to predict when the spelling-sound correspondences at the word body level should inform nonword pronunciation. The TP theory states that a more specific rule should only be sought if a more general rule does not pass the tolerance test, employing the TP’s recursive mechanism that seeks to find a productive rule amongst a subset of items when the initial test is unsuccessful. Therefore, it was expected that if an item has an inconsistent vowel grapheme which does not pass the test, this should trigger the search for a more specific pronunciation rule at the level of the body (a “body winner”). In this way, use of the body winner should be modulated by the presence of a vowel winner; a body winner pronunciation should be used only when a nonword has no vowel winner pronunciation. Results supported this prediction, with adults’ and children’s use of the body winner pronunciation significantly higher in condition 3 (vowel all fail, body winner) than in condition 2 (vowel winner, body winner, conflict).

This evidence that adult and child readers demonstrate knowledge of body-rime correspondences in nonword pronunciations is consistent with similar findings from a range of earlier research from Glushko (1979), Ryder and Pearson (1980), Treiman et al. (1990), Andrews and Scarratt (1998), and Treiman et al. (2003). More specifically, the finding that information from the body level is used more often when the vowel grapheme is inconsistent lends weight to the suggestion by Steacy et al. (2019) that nonword pronunciations are determined by a trade-off between the strength of competing pronunciations at the vowel and body orthographic levels.

However, the current results go further, by supporting the TP's use of a categorical threshold of consistency and its recursive application to predict a specific interaction between vowel and body consistency. Importantly, the use of information from multiple orthographic grain sizes is not pre-determined arbitrarily by a modeller. Instead, it is based on principles of computational efficiency (Yang, 2016, p. 49), according to which productive rules are postulated if they offer the most efficient way to process the data, beginning at the most general level and becoming more specific when required.

Additionally, results suggested that children's use of the body winner across these two conditions was lower than adults', which is in line with previous literature suggesting that children make less use of the body unit than adults when reading aloud (March et al., 1981, Treiman et al., 1990, Bruck & Treiman, 1992, Coltheart & Leahy, 1992, Brown & Deavers, 1999). This is likely because use of more complex contextual information in orthography-phonology mappings is more difficult than use of simpler grapheme-phoneme mappings. Increased use of these context-sensitive mappings may be supported through extensive experience (Treiman and Kessler, 2019), in which readers gain cumulative exposure to these patterns in print (Laxon et al., 1991).

3.5.2 What do participants say when the TP provides no predicted pronunciation?

Items in conditions 4 (vowel all fail, body all fail) and 6 (vowel all fail, body all pass) do not have a single winning pronunciation predicted by the TP at either the vowel or body level, as the pronunciations of these letter patterns are highly inconsistent in English words. Nevertheless, participants do, of course, provide responses for these items. An exploratory analysis found that both adult and child participants make use of a variety of sources of statistical information from their input in their responses, including the type and token frequency of vowel and body pronunciations in English monosyllabic words according to the CELEX corpus (Baayen et al., 1995). Although individual participants sometimes differed in their relative use of these frequencies in their pronunciations, all participants demonstrated use of all four frequency-based measures, with the most frequent pronunciation of the body by type being the most common match, and the most frequent pronunciation of the body by token the least frequent match.

However, confounds within these frequency counts mean that these findings should be interpreted with caution.

3.5.3 Comparing the TP with three computational models of reading

In addition to assessing the TP's success at predicting nonword reading behaviour, I compared it to three computational models of reading. To do so, I analysed the pronunciations of the nonword items generated by a rule-based model (the DRC (Coltheart et al., 2001)), and two statistical models (the Triangle model (Chang et al., 2019) and the CDP+ (Perry et al., 2007)). This approach allowed me to place the TP's rule-based, statistical account of nonword reading within the context of models from opposing theoretical standpoints. For the 138 items included in the analysis, the TP predicted adult and child vowel pronunciations more successfully than the three other models; the Triangle model was the least successful model for both age groups. For adults, there was no difference between the performance of the DRC and CDP+. For children, the DRC was a significantly better match than the CDP+, perhaps because of younger readers' increased reliance on GPCs (Marsh et al., 1981; Brown & Deavers, 1999), which are the pronunciations predicted by the DRC.

In order to understand why the TP is most successful overall, I first considered where the TP matches participant behaviour particularly well in comparison to the computational models. For items with a vowel winner pronunciation (in which pronunciation of the vowel grapheme is consistent enough to pass the tolerance test), the two statistical models perform less well than the TP. However, there is no difference between performance of the TP and DRC, as both predict use of the most common vowel pronunciation by type for these items, which seems to capture adults' and children's behaviour relatively well. Meanwhile, the statistical models may allow more interference from the level of the word body, which does not necessarily reflect human reading behaviour for these items which have consistent vowel pronunciations in English words. For vowel fail items (in which pronunciation of the vowel grapheme is not consistent enough to pass the tolerance test), the TP is a better predictor of adults' and children's vowel pronunciations than all three models. For these items, the TP deviates from the DRC predictions (which are always based on the most common pronunciation of the vowel grapheme); instead, it uses contextual information about the word body to predict the vowel pronunciation. This is achieved through recursive application of the TP: if the pronunciation of the vowel grapheme is too inconsistent to pass the tolerance test, then a more specific productive pattern is sought using the word body. In this way, the categorical threshold of consistency and its recursive application to multiple orthographic levels predicts the precise way in which pronunciation of the vowel grapheme should be informed by adjacent letters. Even though the statistical models are also able to take information about consistency and multiple orthographic grain sizes into account, their predictions are not as successful as those of the TP for these items. A potential reason for this highlighted by Pritchard et al. (2012) is their oversensitivity to token-based statistics. Further, Treiman et al.'s (2003) assessment of a wide range of both rule-based and statistical models found that none successfully accounted for readers' pronunciations of contextually-conditioned vowels.

Looking at predictions from all four accounts that were matched by very few participants provides a fine-grained analysis of how the models deviate dramatically from readers' behaviour, which in turn offers valuable insight into skilled and developing reading. For the rule-based DRC, the majority of pronunciations that were not used by participants were from condition 3, where the vowel grapheme fails the tolerance test. As noted above, the DRC predicts the most common pronunciation of the vowel grapheme for these items, despite the inconsistency of these grapheme-phoneme correspondences in English words. Indeed, participants do not use this level of correspondence to pronounce these condition 3 items, perhaps because it does not offer a pronunciation rule that is strong enough to be generalised in nonword pronunciation. This suggests that a model predicting categorical use of the most common pronunciation for inconsistent vowel graphemes may diverge from human reading behaviour in certain instances.

Turning to the statistical models, both the CDP+ and Triangle models predict some vowel pronunciations according to consonantal context (i.e. body analogies) for items in vowel winner conditions. However, participants do not use these pronunciations for some items in these conditions, perhaps because they have a sufficiently consistent vowel pronunciation (passing the tolerance test), meaning that contextual information from the level of the word body is not required. Therefore, a successful computational model of reading should avoid making body analogies in instances where pronunciation of the vowel grapheme in isolation is relatively consistent, as this can deviate from human reading behaviour. Further, both statistical models predict lexicalisations for some items that are not made by any participants. Whilst individual participants may produce occasional lexicalisations, this type of response does not accurately capture nonword reading behaviour across adult and child participants.

All models fare badly in their predicted pronunciations of the EI/EIL letter sequences. Despite the fact that these items are in condition 1, meaning that one pronunciation of both the vowel grapheme and the body pass the tolerance test without conflict, participants are notably low in their use of this pronunciation. It seems that this letter sequence in particular generates unpredicted pronunciations by both child and adult readers. With the exception of these items, the TP does not predict any pronunciations which are used by no participants, suggesting that its use of a categorical, type-based threshold of consistency and the recursive application to multiple orthographic levels avoids unnatural generalisations of spelling-sound correspondences. Indeed, the TP's approach seems to avoid the pitfalls of other models which result in some pronunciations rarely produced by readers, namely predicting use of a vowel GPC when this pronunciation is inconsistent; allowing interference from the body level when the vowel GPC is relatively consistent; and predicting lexicalisations.

3.5.4 Understanding the TP's performance: the importance of corpus consistency and participant variability

As detailed above, the TP's ability to predict nonword reading behaviour compared favourably with that of three computational models of word reading, including two statistical models that use continuous, token frequency-weighted measures of consistency. Considering that results discussed above indicated the key to the TP's relative success lay in its use of a type-based, categorical metric of consistency that can be applied recursively, a more detailed investigation of the role of consistency in the TP's account of nonword reading was warranted. The aim of this investigation was to assess whether the TP's novel role for consistency captured reading behaviour more successfully than the standard, continuous measure of consistency (the H value, Shannon, 1949), using either type or token frequency.

3.5.4.1 The relationship between corpus consistency and participant regularisation

The investigation began by considering whether the relationship between corpus consistency of a vowel grapheme and participants' vowel regularisation (use of the most common vowel pronunciation by type) was modulated by the TP's own prediction for regularisation. This prediction is based on the pronunciation which passes the tolerance test at either the vowel or body level according to the categorical, type-based threshold which can be applied recursively. Results found that the TP could explain variance in participants' vowel regularisation beyond the effects of continuous consistency measured by type and token. Firstly, this suggests that it is not simply the use of type-based consistency which underlies the TP's ability to predict regularisation, as the TP was an improvement on the simple type-based measure. Instead, I propose that this ability is driven by use of a categorical threshold which provides the trigger to use information from the next orthographic level. In this way, the TP not only captures information from multiple grain sizes, but also integrates this information in a specific way, by predicting an interaction between the effects of vowel and body consistency on reading behaviour.

3.5.4.2 The relationship between corpus consistency and participant variability

Participants reading aloud nonword items do not all respond with the same pronunciations. Therefore, an important part of understanding reading behaviour involves capturing when participants are more or less varied in their responses. Previous research has found that greater inconsistency of spelling-sound mappings gives rise to greater pronunciation variability by participants (Mousikou et al., 2017). Therefore, I investigated the relationship between corpus vowel consistency and variability in participants' vowel responses, and whether it is modulated by the TP. Specifically, I considered whether the participant variability increases with corpus consistency in a linear manner, or whether it is captured more successfully by a binary predictor, namely whether or not there is a spelling-sound correspondence (at either the vowel or body level) which passes the tolerance test. The TP would predict this pronunciation to be used productively by readers, resulting in less variability in participant responses.

Results from this analysis suggested that as corpus type consistency of the vowel increases, participants become less variable in their pronunciations of the vowel grapheme, which is consistent with findings by Mousikou et al. (2017). However, the analysis also indicted that this relationship between type-based consistency and participant variability is not modulated by having a pronunciation which passes the tolerance test. This does not support the suggestion that the TP's categorical metric of consistency (and its recursive application) is able to capture pronunciation variability beyond the effect of a continuous measure of consistency. Whilst analyses above suggested that participants' use of a particular pronunciation is predicted more successfully by the TP than by continuous measures of consistency, these results suggest that this success does not extend to capturing variability across participants.

3.5.4.3 The relationship between participant variability and age

An important aspect of the TP theory is that an individuals' productive rule system is determined by the specific language input to which they have been exposed (Yang, 2016, p. 69). Research has found that text exposure predicts unique variance in orthographic knowledge during reading acquisition (Cunningham and Stanovich, 1992). The 8-9 year-old child participants would have had just a few years' reading experience and exposure to a relatively narrow range of text types compared to adult participants. Moreover, they would all have learned to read through systematic training in synthetic phonics, meaning that their instruction would be very similar across schools, and would have highlighted the most common mappings between phonemes and graphemes. This instruction may have long-lasting effects on participants' use of alternative pronunciations of nonwords, as found in adult readers by Thompson et al., (2009). Meanwhile, adult participants had learned to read before the introduction of compulsory phonics, and would also have had an increased opportunity over many years to become much more varied in their reading experience. For these reasons, I hypothesised that child participants may have formed more similar pronunciation rule systems to each other than adult participants may have, with the result that responses across child participants would be less variable than across adult participants.

However, results indicated that children's responses were in fact more variable than adults', contrary to this prediction. I suggest that this could be due to the fact that at this age, children are still undergoing development in their reading ability which can progress at different rates for different children for a variety of reasons (Powell et al., 2014), despite their similar instruction. Their productive rule systems may reflect these differences, and give rise to a greater variability in responses. Adult readers, meanwhile, may have reached a point at which the rule systems they have developed are actually more stable and similar to each other's, as individual differences in text exposure (at least in terms of the frequency of exposure to certain spelling-sound correspondences) become less pronounced over years of broad, cumulative reading experience.

3.5.5 Summary: what can the TP capture about nonword reading behaviour?

This initial investigation of the TP within the field of reading suggests that Yang's (2016) theory is able to offer a novel stance within the existing literature on nonword reading behaviour. Namely, it is a rule-based approach which uses statistical information about consistency, thus providing a bridge between previous rule-based models of word reading (such as the DRC) and statistical models (such as the Triangle and CDP+ models). Results from the current study suggest that this approach can indeed account for adults' and children's nonword reading aloud more effectively than these models, which have previously been found wanting (Treiman et al., 2003).

Overall, the TP's role for consistency seems to underlie its success in predicting nonword pronunciations. This approach to consistency is novel in a number of ways: firstly, it is based on type frequency counts alone; the number of alternative pronunciations is not weighted by token frequency within the algorithm. According to Yang, "the empirical frequencies of words are ignored entirely" (2016, p. 76) when determining productivity, which is instead calculated on the basis of the number of regular and irregular items. Secondly, it provides a categorical metric of consistency; the threshold produced by the algorithm determines the specific level of inconsistency (i.e., the number of exceptions) that can be tolerated by a productive pattern. The particular consistency of a pattern is immaterial until it reaches this threshold. Thirdly, the TP applies recursively: when a productive pattern is not found across a total set of items (i.e. the pattern "fails the tolerance test"), the TP applies to smaller subsets in order to find more specific, productive patterns. For reading, not only does this mean that the TP can capture information at different orthographic levels (such as the grapheme and the word body) but that it can predict a precise interaction between them. For example, it predicts that information from the word body is only used when the inconsistency of a vowel grapheme goes beyond the threshold. Whilst earlier work has suggested that there may be an interaction between the consistency of context-dependent and context-sensitive pronunciations (Treiman et al., 2003; Kessler, 2009; Steacy et al., 2019), the TP is the first account to predict explicitly when this interaction should occur. Furthermore, this prediction is borne out by the findings that it can capture nonword reading behaviour more successfully than the three models of reading considered here.

Together, these facets of the TP's account of consistency enable a precise prediction about which spelling-sound mappings will be used productively by readers, in a way that is more successful than other models of word reading or type- and token-based measures of consistency. However, it was not able to account for variability in responses across participants, which perhaps hints at the importance of taking into account the role of an individual's reading experience in order to understand human reading behaviour.

This study has also informed our understanding of skilled and developing nonword reading behaviour beyond the TP. Whilst the TP's type-based consistency metric captures reading behaviour most successfully, findings have also suggested that readers may also pay attention to other statistical properties of text. In particular, an analysis of items without pronunciations predicted by the TP provided preliminary evidence that token frequency information from the grapheme and body level can inform nonword pronunciations by adults and children. The analysis of the relationship between corpus consistency and regularisation for items with a TP prediction also found that a token-frequency based measure of consistency had an effect on regularisation alongside the effect of the TP. Further investigation into the role of token frequency may shed light on the finding that in general, nonword reading behaviour is not as categorical as TP predicts.

Building on earlier findings, this study has provided additional evidence that both adult and child readers use contextual information from larger orthographic grain sizes than the grapheme in nonword reading. Further, children used the word body to inform their vowel grapheme pronunciations less often than adults, in line with previous research (March et al., 1981; Treiman et al., 1990; Bruck & Treiman, 1992; Brown & Deavers, 1999). Despite their more similar reading instruction and experience, children were more varied in their responses than adults, suggesting that their productive systems of orthography-phonology mappings are still undergoing development. Together, these results suggest that younger readers continue to gather information from their reading experience over many years, including knowledge of more complex body-rime pronunciation patterns.

Chapter 4: Testing the Tolerance Principle in adults and children learning an artificial orthography

4.1 Introduction

The aim of Experiment 2 was to examine whether the Tolerance Principle (Yang, 2016) could predict adult and child learners' generalisation of novel spelling-sound correspondences. Using an artificial orthography paradigm, participants were first trained to read aloud nonword items using novel vowel symbols, and then asked to pronounce untrained items in order to capture their generalisation of the novel spelling-sound correspondences. This experiment manipulated whether the consistency of the spelling-sound correspondence for each novel vowel symbol passed the tolerance test or not according to the tolerance algorithm (Yang, 2016, p. 8-9).

This chapter begins by reviewing existing research that explores the TP experimentally; this work is then placed in the context of findings from other artificial language learning studies and the wider statistical learning literature. I then review existing research using artificial orthography learning paradigms, before demonstrating how this methodology can be used to explore the TP, address wider issues in statistical learning, and inform the debates surrounding models of word reading.

4.1.1 Previous experimental work on the TP

There have been no previous studies exploring the TP using artificial orthography paradigms. However, a series of studies (Schuler et al., 2021; Schuler, 2017) used an artificial language paradigm to investigate whether the TP could predict adult and child participants' generalisation of morphological patterns. Artificial language learning paradigms are a valuable tool for research on language learning and generalisation, as they allow precise control over both the design of the input language and exposure to the language in the learning environment (Taylor et al., 2011; Taylor et al., 2017).

Schuler's (2017) artificial language was formed of 9 nonsense nouns paired with novel plural morphemes. Twenty adults and fifteen children (aged 6-8 years) were exposed to one of two language conditions: one in which 5 of the 9 nouns used a regular plural maker, and the other in which 3 of the 9 nouns used the regular plural marker. In the 5 Regular/4 Exceptions (5R4E) condition, the number of exceptions did not cross the tolerance threshold for a set of 9 items (4 exceptions), so the TP predicts that a productive rule should be formed. In the 3 Regular/6 Exceptions (3R6E) condition, the number of exceptions exceeded the tolerance threshold, so the TP predicts that a productive rule should not be formed.

In Experiment 1, noun frequency during exposure to the language varied along a Zipfian distribution (Zipf, 1949), with nouns that used the regular marker appearing most frequently in both conditions. Following the exposure, participants' use of a productive rule was assessed by producing plural markers for untrained novel nouns. It was hypothesised that participants who formed a productive rule as predicted by the TP would use the regular marker 100% of the time during generalisation; participants who did not form a rule should use this marker significantly less often than 100% of the time. Indeed, child participants' use of the regular marker did not differ significantly from 100% in the 5R4E condition, whilst regularisation was much lower in the 3R6E condition at 16.9%. It thus appears that children behave categorically as predicted by the TP: they used a productive rule in the 5R4E condition but not the 3R6E condition. Meanwhile, adults displayed a different pattern of results. Their use of the regular marker in the 5R4E condition was at 65.0%, significantly lower than the predicted 100%. Furthermore, this rate of regularisation was not significantly different from the rate of 51.7% in the 3R6E condition. Therefore, adults did not obey the TP as children do in their generalisation behaviour. Instead, Schuler found that adults' use of the regular marker matched the token frequency with which it occurred in the linguistic input, a strategy known as probability matching (Hudson Kam & Newport 2004, 2005).

The author notes that the children's generalisation behaviour here could be explained by a simpler evaluation metric than the TP: it is possible that learners only required a majority of forms to follow the regular pattern in order to support productivity. The TP employs a stricter threshold for productivity: a substantial majority is required, with the number of tolerated exceptions decreasing proportionally as the overall number of items increases. To investigate this further, Experiment 3 involved a more rigorous test of the TP, using a new artificial language that consisted of 10 regular nouns and 6 irregular nouns. Here, the number of exceptions exceeded the tolerance threshold (5), so if learners are using the TP then a productive rule should not be formed. However, if learners are using a simple Majority of Forms metric, then a productive rule should be formed. Results suggest that children aged 5-7 followed the TP, as their use of the regular marker (39.9%) was significantly lower than 100% and not significantly different from chance. Meanwhile, adults' use of the regular marker was not significantly lower than 100%, as predicted by a Majority of Forms approach. However, this rate of regularisation also did not differ from the high token frequency of the regular form in the input, meaning that their behaviour could also be explained by probability matching. Moreover, the low number of participants in this experiment (ten children and seven adults) means that it is difficult to form any firm conclusions.

Schuler's (2017) finding that children regularised inconsistent grammatical markers in accordance with the TP's categorical predictions, whilst adults used an alternative probability matching strategy, opens the possibility that the TP will also have differential effects on adults' and children's generalisation behaviour in the current artificial orthography learning study. There is a theoretical basis for predicting that, in general, children rather than adults will follow the TP. Yang (2016, p. 66-67) suggests that children are superior language learners to adults because their vocabularies are smaller, and as the TP tolerates a larger proportion of exceptions for smaller groups of items, this allows children (with smaller vocabularies) to form productive rules more easily than adults (with larger vocabularies). However, this is unlikely to be a relevant motivation in an artificial language or orthography experiment where all participants are exposed to the same stimuli in the same way.

Instead, Schuler (2017, p. 79-91) proposes that the TP applies uniquely to children as a result of cognitive differences between children and adults which result in an early maturational state optimally suited to the TP. Building on Newport's (1990) "Less is More" hypothesis, she proposes that these differences involve memory and cognitive constraints during development which allow the TP to operate as a low-level competition, rule-forming mechanism for children (2017, p. 90). Adults, with fully-developed cognitive capacity, can override this low-level mechanism and use more complicated strategies such as probability matching, thereby failing to acquire productive rules in the way that children do. Indeed, it is effectively children's cognitive limitations which enable them to acquire these rules so successfully. The current study allows an expanded investigation of this hypothesis to compare children and adults' rule-learning behaviour in the field of reading, specifically of spelling-sound correspondences.

Additional experiments from Schuler (2017) will also be reviewed in Chapter 5, as part of a more detailed discussion about the importance of type and token frequency input statistics for learning. In summary, the results described above suggest that children, but not adults, follow the TP predictions of generalisation behaviour in a strikingly categorical way. It must be noted that the small number of participants in these experiments (together with use of a between-subjects design in Experiment 1) means that the results should be interpreted with caution. Nevertheless, this research demonstrates that an artificial language training-generalisation procedure can be used as an effective method for testing the TP experimentally.

4.1.2 Statistical learning

The results obtained by Schuler (2017) have interesting implications in the context of previous literature on statistical learning, both regarding the extraction of statistical information to form regularities, and the differences between adults and children. Statistical learning is an approach to language acquisition which maintains that learners use general learning mechanisms to extract information about statistical distributions in the linguistic input. These mechanisms are shaped by memory and processing constraints, with the result that underlying similarities between languages are not necessarily the result of innate linguistic knowledge, but rather the nature of the learning process (Saffran, 2003).

Early statistical learning studies demonstrated that learners are able to track basic statistical properties of the input. Saffran et al. (1996) found that adults presented with an artificial language were able to segment a speech stream into syllables, and acquire syllable order, according to the distributional characteristics of the continuous input. Saffran et al. (1997) demonstrated that children aged 5-7 were also able to do so, noting that this seemed to be an implicit process without any direction of participants' attention at the speech stream.

Exploring possible constraints on statistical learning, Saffran et al. (1996) found that 8-month-old infants could discriminate between "words" and "nonwords" after a two minute exposure to a simplified corpus of trisyllabic nonsense words, by using information about transitional probabilities in the input. Aslin et al. (1998) found that 8-month-olds could go further by discriminating words from part-words, and demonstrated that this was indeed due to syllabic conditional probabilities rather than syllable frequencies. The authors concluded that infants of this age are performing an analysis of the statistical distributions in the input involving a large number of different conditional probabilities rapidly and simultaneously, and without explicit direction of attention. Saffran et al. (1999) conducted a similar study with infants the same age but using a non-linguistic tone stream rather than a speech stream as input. Again, the infants were able to differentiate between “tone words” and “tone part-words”, suggesting that the statistical mechanism participants had used in the linguistic studies was capable of performing similar computations on non-linguistic input. Together, these early statistical learning studies demonstrate that infants, children and adults are able to extract distributional information from the input, and that this capability may not solely be based on language-specific learning mechanisms.

According to Thiessen et al.’s (2013) Two-Process Account, this kind of statistical learning process should be categorised as *extraction*, whereby discrete units are extracted from a continuous input using conditional statistical information such as transitional or conditional probabilities (as described in the studies above). In contrast, the subsequent process of *integration* involves combining information across these extracted units to reveal the central tendencies of a set of items. According to Thiessen et al., this process involves distributional statistical information such as frequency and variability (Maye et al., 2002; Thiessen & Pavlik 2012). It is this second process of statistical learning which is particularly relevant to our investigations, and which has been the focus of a line of research preceding Schuler’s work, asking: how do learners use statistical distributions in the input to identify patterns, and is the outcome of this process probabilistic or all-or-none?

Hudson Kam and Newport (2005) explored these questions, investigating what patterns learners are able to acquire from inconsistent linguistic input using an artificial language learning paradigm. Specifically, they asked whether participants learn and reproduce the statistical distributions they are exposed to, or whether they make the language more consistent through regularisation, i.e. reducing variability by maximising one pattern from the input or imposing a new pattern, and using this consistently. Experiment 1 manipulated the consistency of determiners during exposure by varying their frequency of occurrence with nouns across different input conditions. During a production test, adult learners generally reproduced the level of variation in the input by using the determiners at the rate they had heard them during training. This pattern of results suggests that participants had learned veridically rather than by adopting a general rule to characterise the language (i.e., *regularising* the inconsistencies). In Experiment 2, they compared the behaviour of adults and children (aged six) using a simpler version of the language, but again manipulated frequency of the determiner across input conditions. Examination of individual participant productions found that children used the determiners systematically even in the inconsistent conditions, which suggests they impose regular structure on inconsistent input. Meanwhile, adults varied their use of the determiner when the input was inconsistent, and demonstrated systematic use of determiners only when their input was consistent. The authors suggest this reveals dramatic differences between the outcomes of learning at different ages, with children showing a strong tendency to regularise the language whilst adults match the variation they are exposed to, known as probability matching. However, it is worth noting that despite these contrasting patterns of behaviour, Experiment 2 did not reveal significant differences between children and adults, perhaps as a result of the small sample sizes.

In another series of artificial language learning experiments, Hudson Kam and Newport (2009) again investigated whether learners reproduce or regularise variation from the linguistic input. In Experiment 1, adults were presented with different levels of “scattered inconsistency” during training: one “main” and multiple “noise” determiners occurred with nouns at varying frequency across input conditions. In a production test, participants who had received more “scattered” or complex input used more main determiner forms than participants with less scattered input. Therefore, it seems that increasingly complex and inconsistent input induces higher rates of regularisation in adults. Results from Experiment 2 suggested that when determiners had the same complex range of variation (or “scatter”) and low level of frequency, but occurred in consistent rather than inconsistent noun contexts, adults no longer regularised their use of determiners. Instead, they learned the patterns veridically and reproduced the variation present in the input. Using a simplified version of Experiment 1, Experiment 3 compared adult and child learners aged 5-7 years. They found that during production, children almost always regularised their determiner use, regardless of their input condition. In contrast, adults only displayed systematic use of the determiner in the most consistent condition; otherwise, they used determiners in a variable way, as in their input. This suggests that adults, unlike children, reflected the inconsistency of the determiners in the input. As the scattered inconsistency here was not as complex as that found in Experiment 1, adults did not regularise their determiner use as they did in the earlier experiment.

In a similar study, Wonnacott et al. (2017) investigated adults' and six-year-old children's ability to track distributional statistics in a semi-artificial language learning paradigm, specifically looking at the learning of co-occurrence relationships between particles and nouns. Learners were exposed to either a "skewed" language in which five nouns used the same particle and one noun used a second particle, or an "unskewed" language, in which three nouns used one particle, and three nouns used a second particle. Both children and adults demonstrated better accuracy when producing particles for trained input nouns from the skewed than unskewed language. Importantly, this was true for both majority and minority noun-particle pairings in the skewed language. Children also demonstrated better accuracy for "minimal exposure" nouns (introduced after initial training) in the skewed language; results from adults were inconclusive for these items. The authors suggest these results reveal stronger lexical learning and less generalisation in the skewed than unskewed language. However, this study did not explicitly assess generalisation to novel items; the only opportunity to assess (over-)generalisation was in incorrect particle productions for trained or minimal exposure items that had not been successfully learned. Due to this design, it is impossible to tease apart productive generalisation behaviour from learning success or failure. Moreover, it is also possible that learners in the skewed condition are actually demonstrating *more refined* generalisation than those in the unskewed condition; that learning a majority pattern (perhaps as a productive rule) could make it easier for a small number of exceptions to be learned individually, stored separately, and thus not over-generalised. Indeed, this is consistent with a TP approach to generalisation and rule-learning.

Together, results from these artificial language learning studies suggest that learners are able to track statistical or probabilistic information in the input, as maintained by a statistical learning account of language acquisition. The outcome of the process can be veridical (i.e., reflecting distributions from the input), or it can result in regularisation of the language (i.e., imposing additional structure that goes beyond the input statistics by removing inconsistent or probabilistic patterns). Children in particular appear to regularise systematically, ironing out inconsistencies in the input, whereas adults may only regularise in instances where the inconsistencies or complexities of the input are too difficult to track. Otherwise, the outcome of their learning seems to reproduce the statistical patterns and inconsistencies they were exposed to. These results are largely in line with Schuler's recent work, which also found that children are likely to regularise their input whilst adults reproduce the variation from the input. However, this work also revealed limits on children's regularisation, namely a level of inconsistency in the input that crosses the tolerance threshold (in the 3R6E condition). These findings will be explored further in the current experiment, using an artificial orthography paradigm.

4.1.3 Artificial orthography studies

Early artificial orthography studies focused on whether participants could learn individual sound to symbol mappings for novel letters, without using whole-word items. For instance, Byrne (1984) and Byrne and Carroll (1989) found that adults could learn associations between sounds and novel symbols, but could not acquire or generalise knowledge of the mappings between phonetic features and orthographic elements that were embedded in the novel symbols. More recently, artificial orthography studies have begun to examine extraction of sub-word spelling-sound mappings from word-level items.

For example, Bitan and Karni (2003) investigated adults' learning of spelling-sound patterns using novel words written in an artificial script, in which each phoneme was represented by a sequence of two or three symbols. Using three different sets of stimuli, participants completed explicit, implicit and arbitrary (pictographic) training conditions, with the order of training counterbalanced across participants. The authors found that only after explicit training involving letter decoding instruction could participants segment word items into specific symbol strings and transfer this letter knowledge from the trained to novel items. This suggests that adults require explicit instruction in order to extract individual spelling-sound correspondences from whole words and to generalise this knowledge (see also Rastle et al. (2021) for evidence of the dramatic impact of direct instruction on learners' generalisation of novel spelling-sound regularities). However, certain aspects of the methodology mean that these results should be interpreted with caution. Firstly, (as noted by Taylor et al., 2011) use of a sequence of two or three symbols to represent each phoneme – a feature designed to reduce interference from alphabetic knowledge – may increase the difficulty of extracting individual spelling-sound correspondences. Secondly, within each stimulus set, combinations of only three different symbols were used to create the “letter” sequences, which may compound this extraction difficulty by impeding differentiation between the symbol sequences. Indeed, Bitan and Karni note that the complexity of the segmentation rules may have hindered extraction of the letter sequences. Thirdly, there were only six words in each training set, which would likely reduce learners’ ability to extract patterns without instruction due to an insufficient number of exemplars, given that variability is necessary for generalisation (Tamminen et al., 2015). Finally, only a small number of participants (nine) took part in the study.

Subsequent studies have, in contrast, found that adult learners *are* able to extract spelling-sound correspondences from exposure to word-level items written in a novel orthography, without explicit instruction. The first to do so was by Taylor et al. (2011), who also explored whether adults’ learning and generalisation of spelling-sound correspondences was influenced by their consistency and frequency during training. Four novel vowel characters were used in the artificial orthography: two were consistent with a one-to-one grapheme-phoneme mapping, and two were inconsistent, pronounced in one way when preceded by a particular consonant character and in a different way when preceded by any other consonant character. Type frequency of the characters during training was also manipulated, with certain spelling-sound correspondences either “high” or “low” frequency. Results showed that learners were sensitive to the frequency, consistency and consonantal context of spelling-sound mappings during learning and generalisation, indicating that they were sensitive to the statistics of the learning environment. Specifically, spelling-sound mappings were learned and generalised more accurately if they were highly frequent or highly consistent. In learning, there was an interaction between frequency and consistency, such that consistency only affected items with low-frequency vowels, and the advantage of items with high over low frequency vowels was found only when vowels were inconsistent. In generalisation, this interaction did not reach significance but the data followed a similar trend to that seen during the learning phase.

Using an artificial orthography based on Chinese phonograms, Zhao et al. (2018) also found an effect of consistency on learning. For both orthography-phonology and orthography-semantics mappings, higher consistency during training gave rise to more successful learning of both types of correspondences by adults. Furthermore, they demonstrated that learners could develop knowledge of sub-lexical regularities after exposure to the language, without any explicit training of these regularities. Taylor et al. (2017) also used an artificial orthography that included phonological and semantic mappings, and found that training which emphasised orthography-phonology correspondences rather than orthography-semantics correspondences was more effective for adult learners. The benefits included faster and more accurate reading aloud of trained items, faster generalisation of novel items, and higher accuracy in comprehension of written words earlier in learning. Furthermore, their fMRI neuroimaging results found that there is large overlap between neural activity when reading aloud English words, pseudowords, and the items written in the artificial orthography. This supports use of artificial orthographies as an effective way to reveal insights about the systems underlying reading. Additional neuroimaging evidence that these paradigms are not general problem-solving tasks but instead tap specific reading mechanisms is presented by Taylor et al. (2019): two weeks' training on an artificial orthography yielded a hierarchy of neural representations for orthographic, phonological, and semantic information along the ventral stream in adult learners.

In contrast to this recent research with adult participants, very few artificial orthography studies have investigated learning by children; those that did involve children have tended to examine the effects of dyslexia on learning (e.g. Law et al., 2018; Tong et al., 2020). Some work has been carried out with typically developing children investigating whether statistical learning processes – similar to those observed in spoken language acquisition – underlie the development of knowledge between spelling and sound. For example, Samara et al. (2019) explored whether learners could generalise novel graphotactic rules (i.e. permissible letter combinations) without explicit instruction in two different linguistic contexts (Turkish and English). Turkish and English children aged 6-7 were exposed to CVC nonwords in their respective orthographies which contained contingencies between the medial vowel and either word-initial or word-final consonants. Following exposure, children were able to discriminate between permissible and impermissible novel items, which suggests they could learn and generalise graphotactic constraints using information about statistical distributions in the input rather than through instruction. Beyond these findings that children seem to make use of statistical information as they acquire knowledge of spelling-sound patterns, the current study is one of the first to explore what input characteristics support children's learning and generalisation of unfamiliar spelling-sound patterns using an artificial orthography.

In summary, the body of research described above has demonstrated that adult learners are able to extract spelling-sound correspondences from whole-word items written in artificial orthographies, and that their learning and generalisation of this knowledge is sensitive to the consistency, frequency and training of the stimuli. I therefore consider an artificial orthography paradigm to be an appropriate methodology to investigate a number of broad research questions. Firstly, it can assess whether learners acquire and generalise spelling-sound correspondences according to the numerical predictions of the TP by manipulating type frequency, token frequency and consistency of spelling-sound mappings during training in a precise way. This allows greater control over the input statistics than is possible in natural language research. Secondly, it can be used to address wider issues in statistical learning such as which statistical properties of the input are most important for learning and generalisation, and whether participants' learning reflects distribution of the input or imposes increased regularity. Thirdly, the results will be informative in the debates surrounding models of word reading. Different computational models (including the DRC, CDP+ and Triangle models) place varying weight on input variables including type frequency, token frequency and consistency. Using a methodology that allows complete control over these variables in order to examine their effects on learning and generalisation can therefore be useful to assess the merits of extant reading models. Finally, by training adults and children in the same paradigm, we are able to investigate potential age-related differences in the generalisation of spelling-sound knowledge. These may be parallel to the differences between adults' and children's generalisation of grammatical knowledge found in the artificial language learning studies above.

4.2 Experiment 2

Experiment 2 investigated whether the Tolerance Principle (Yang, 2016) could predict adult and child learners' generalisation of novel spelling-sound correspondences. The TP states that there is there is a categorical distinction between productive rules that can be generalised (applied to unseen items), and rules which are lexically-specific and unproductive as the number of exceptions to the rule exceeds the tolerance threshold (Yang, 2016, p. 9; p. 34; see *Section 2.5* for detailed discussion). Yang provides a tolerance algorithm (Yang 2016, p. 8-9) which numerically predicts the tolerance threshold for a set of N items. The TP can therefore be used to predict which spelling-sound correspondences or "pronunciation rules" pass the tolerance test and should be used productively by learners.

Using an artificial orthography paradigm, participants were trained to read aloud nonword items which used three novel vowel symbols. The novel vowel symbols varied in the consistency of their relationship to spoken forms, such that they either passed the tolerance test or did not. Participants were then tested on trained items to assess the orthography-phonology knowledge they had learned from this artificial orthography, and critically were also tested on untrained items to assess their generalisation of this knowledge. Additionally, an old-new test of trained and untrained items was used to assess learners' recognition memory of trained items.

The consistency of the three novel spelling-sound correspondences during training was manipulated by condition. Each condition had 10 nonword items using one novel vowel symbol. The TP states that for a set of 10 items, the number of exceptions a productive rule can tolerate is 4. In the 8 Regular/ 2 Irregular condition (8R2I), the vowel symbol had a regular pronunciation in 8 items and irregular pronunciations in 2 items. In the 6 Regular/ 4 Irregular condition (6R4I), the vowel symbol had a regular pronunciation in 6 items and irregular pronunciations in 4 items. Following the TP, participants should form a productive pronunciation rule in both the 8R2I and the 6R4I conditions as the number of irregular pronunciations does not exceed the tolerance threshold of 4. However, in the 4R6I condition, the vowel symbol had a regular pronunciation in 4 items and irregular pronunciations in 6 items. In this condition, the TP predicts that participants should not form a productive pronunciation rule for the vowel symbol.

During the Generalisation task in the testing phase, participants read aloud a set of 30 untrained items using the three novel vowel symbols learned during the training phase. Analysis of these pronunciations allowed us to assess whether participants had formed a productive rule for each novel vowel symbol on the basis of exposure to these vowels during training. The TP predicts that learners will *regularise* their pronunciation of the vowel (i.e., productively use the most common type of vowel pronunciation from the training phase) for 100% of generalisation items from the 8R2I and 6R4I conditions, where a productive rule is predicted. Critically, the TP predicts there should be no difference in participants' regularisation in these two conditions. In contrast, the TP predicts learners will use the regular pronunciation at no more than chance level (25%) for items from the 4R6I condition, where a productive rule is not predicted. Although the TP predicts the same pattern of regularisation for both adults and children, Schuler et al. (2021) found that the TP successfully predicted children's generalisation but not adults' in an artificial grammar paradigm. Therefore, we may expect the TP to have a greater effect on children's vowel regularisation than adults' in the Generalisation task.

The TP prediction for regularisation in each condition differs from that of an established rule-based model of reading: the DRC (Coltheart et al., 2001) would expect learners to use the regular pronunciation for generalisation items in all conditions, as it stipulates that the most frequent pronunciation of a grapheme (by type) should be used productively.

In contrast to the rule-based approaches of the TP and DRC, statistical learning frameworks predict that generalisation behaviour is sometimes (particularly for adults) based on the statistical distributions from the learning environment in a continuous way. This could involve matching the type or token frequency of forms in the input (e.g. probability matching (Hudson Kam and Newport 2004, 2005)). Therefore, this approach would predict type and token frequency of the regular vowel during training to have an effect on vowel regularisation. Indeed, Schuler (2017) found the token frequency of regular plural-markers during training to be the basis of adults' regularisation behaviour in an artificial grammar paradigm. In the current study, the token frequency of regular items varied across all participants, as items from the artificial language were assigned a rank on the Zipfian distribution at random for each participant. Meanwhile, the type frequency of the regular pronunciation remains at a constant 80% in the 8R2I condition, 60% in the 6R4I condition, and 40% in the 4R6I conditions, for all participants.

Finally, I considered the relationship between successful learning of trained regular items assessed in the Final Reading Aloud test and the rates of vowel regularisation for untrained items in the Generalisation Task, in order to explore the relationship between learning and generalisation of the regular pattern.

4.3 Method

4.3.1 Participants

Twenty-seven adult participants (mean age 21; 20 females and 7 males) were recruited from the student body of Royal Holloway, University of London. Twenty-six child participants (mean age: 10 years 4 months, range: 9 years 10 months; 10 years 9 months; 11 females and 15 males) were recruited from a primary school in Somerset, UK. Participants were monolingual, native English speakers, with a Southern British English accent and no known language or learning difficulties. Participants had normal or corrected-to-normal vision. Each adult participant received £10 for their involvement in the study. Each child participant received a certificate and stickers for their involvement in the study; the school received redeemable vouchers. Three adult participants were excluded, due to technical error (*n* = 1) and not fulfilling eligibility requirements (*n* = 2). Two child participants were excluded due to poor performance, by scoring below 20% accuracy in the final Reading Aloud test of the 30 exposure items. Therefore, data from twenty-four adult participants and twenty-four child participants were included in our analysis. The study received approval from the procedures of the Ethics Committee at Royal Holloway, University of London.

4.3.2 Stimuli and design

To assess whether participants who learn to read aloud an artificial orthography follow the predictions of the Tolerance Principle (TP), I designed an artificial language consisting of 30 nonword items. Each item had a consonant-vowel-consonant (CVC) structure. In the orthography of this artificial language, the consonant graphemes were 11 familiar letters from the English alphabet (d, t, p, b, k, g, m, n, v, f, l) which corresponded to their most frequent pronunciations (/d/, /t/, /p/, /b/, /k/, /g/, /m/, /n/, /v/, /f/, /l/ respectively). The vowel graphemes were three “novel” letters: “ʕ” “ϱ” “Ϸ”, the forms of which were borrowed from the Semitic, Greek and Bactrian alphabets respectively. In this artificial orthography, these three graphemes had inconsistent vowel pronunciations, with a one-to-many grapheme-phoneme mapping. The consistency of the pronunciations of the three vowel graphemes were manipulated to form three conditions: two conditions in which the TP predicts that learners should form a productive rule for pronunciation of the vowel grapheme, and one condition in which the TP predicts learners should *not* form a productive rule for pronunciation of the vowel grapheme.

Each of the three conditions consisted of 10 nonword items. Each vowel grapheme was used in only one condition, appearing in the medial position of all 10 items for that condition. Yang's TP algorithm (2016, p. 8-9) was used to calculate the number of exceptions a productive rule can tolerate for a set of 10 items; the predicted threshold of tolerated exceptions is 4 items. This threshold was used to form two conditions in which, according to the TP, the pronunciation of the vowel grapheme is sufficiently consistent to form a productive rule (passing the tolerance test), and one condition in which the pronunciation of the vowel grapheme is not sufficiently consistent to form a productive rule (failing the tolerance test). The first condition had 8 regular items following a pronunciation rule and 2 irregular pronunciations (8R2I). In this condition, the TP would predict learners to form a productive rule as the number of irregular items (2) falls below the tolerance threshold (4). The second condition had 6 items following a regular pronunciation and 4 irregular pronunciations (6R4I). In this condition, the TP would again predict learners to form a productive pronunciation rule for the vowel grapheme, as the number of irregular items does not exceed the tolerance threshold of 4. However, this condition provides a more rigorous test of the TP as the number of irregular items reaches, but does not breach, the tolerance threshold for a set of 10 items. Critically, the TP predicts no difference in regularisation behaviour for these two conditions that pass the tolerance test. The third condition had 4 regular items following a pronunciation rule and 6 irregular items (4R6I). In this condition, the TP would predict that learners do not form a productive rule for the pronunciation of the vowel grapheme as the number of exceptions (6) exceeds the tolerance threshold of 4 irregular items.

The regular pronunciation of the vowel grapheme corresponded to the phoneme /ɪ/ in the 8R2I condition, the phoneme /ɒ/ in the 6R4I condition, and the phoneme /iː/ in the 4R6I condition. The phonemes /e/, /uː/ and /æ/ were used as irregular pronunciations for the vowel graphemes, each appearing in a total of four nonword items each across all three conditions. The use of three vowel graphemes in each condition was rotated so that three mappings (A, B and C) were counterbalanced across participants, as seen in Table 4.1.

Table 4.1

Rotated vowel grapheme mappings in three conditions

	Mapping A	Mapping B	Mapping C
8R2I vowel	ƍ	Þ	ʕ
6R4I vowel	Þ	ʕ	ƍ
4R6I vowel	ʕ	ƍ	Þ

The full artificial language consisting of three conditions and 30 nonword items, with pronunciations and orthographic representations using Mapping A, is shown in Table 4.2.

Table 4.2

The condition, orthographic form, pronunciation and regularity of 30 nonword training items

Condition	Orthographic Representation	Pronunciation	Mapping Regularity
8R2I	**pǫb**	**/pɪb/**	**Regular**
8R2I	**bǫp**	**/bɪp/**	**Regular**
8R2I	**kǫg**	**/kɪg/**	**Regular**
8R2I	**gǫn**	**/gɪn/**	**Regular**
8R2I	**tǫv**	**/tɪv/**	**Regular**
8R2I	**lǫf**	**/lɪf/**	**Regular**
8R2I	**fǫd**	**/fɪd/**	**Regular**
8R2I	**vǫk**	**/vɪk/**	**Regular**
8R2I	mǫl	/mel/	Irregular
8R2I	nǫm	/nuːm/	Irregular
6R4I	**lÞn**	**/lɒn/**	**Regular**
6R4I	**nÞp**	**/nɒp/**	**Regular**
6R4I	**vÞk**	**/vɒk/**	**Regular**
6R4I	**fÞd**	**/fɒd/**	**Regular**
6R4I	**mÞt**	**/mɒt/**	**Regular**
6R4I	**dÞm**	**/dɒm/**	**Regular**
6R4I	tÞb	/teb/	Irregular
6R4I	gÞv	/gæv/	Irregular
6R4I	kÞf	/kæf/	Irregular
6R4I	pÞg	/puːg/	Irregular
4R6I	**pʕb**	**/pi:b/**	**Regular**
4R6I	**mʕg**	**/mi:g/**	**Regular**
4R6I	**gʕl**	**/gi:l/**	**Regular**
4R6I	**tʕf**	**/ti:f/**	**Regular**
4R6I	kʕk	/kek/	Irregular

4R6I	vʕd	/ved/	Irregular
4R6I	lʕt	/læt/	Irregular
4R6I	dʕv	/dæv/	Irregular
4R6I	fʕn	/fuːn/	Irregular
4R6I	bʕp	/buːp/	Irregular

Frequency of the consonant graphemes, as well as word position and co-occurrence with the vowel graphemes, was counterbalanced both within conditions and across the entire exposure set. This ensured that pronunciation of any vowel grapheme could not be associated with or predicted by use of the consonant graphemes.

Participants were exposed to this set of exposure items from the artificial language throughout the training phase (discussed in further detail under *Procedure* below*)*. During training, the nonword items varied approximately along a Zipfian frequency distribution (Zipf 1949), according to which the frequency of a word is inversely proportional to its rank. This means that the most frequent word occurs about twice as often as the second most frequent word, and three times as often as the third most frequent word. This design was chosen as word frequency in natural language follows an approximately Zipfian distribution, and the derivation of Yang's TP also assumes this is the case (Yang, 2016). The training phase consisted of 131 total presentations of the 30 unique nonword items; the most frequent item appearing 24 times and the least frequent items appearing 3 times each. Items were randomly assigned a rank on the Zipfian distribution for each participant, meaning that items were encountered a different number of times by each participant. This ensured that the token frequency of regular and irregular items to which individual participants were exposed could be examined, whilst avoiding the application of a consistent, arbitrary assignment of items to frequencies for the entire language across all participants. The token frequencies of exposure to items for all participants during the training phase are available here: https://osf.io/wsc24/?view_only=10e04739627b4a27861154c5edafdaf6.

4.3.3 Procedure

Participants were briefed on the nature of the task by being informed that they would be trained to read items from an artificial language using an artificial script. They were informed that some letters from the script would be familiar, English letters and others would be novel symbols for them to learn. They were informed that they would be trained to read the artificial script by carrying out reading aloud and spelling activities on the computer. The procedure was run using E-Prime software.

The training phase began with an exposure to the set of 30 trained items. Participants were presented with the written form of each item one at a time on the computer screen for a duration of 6 seconds, and also heard a pre-recorded pronunciation of the item commencing after 2 seconds of the visual presentation. Participants were asked to try to remember the pronunciation of the items. Each item was presented once and items were presented in a randomised order.

After completing the exposure to each of the 30 items, participants carried out a Reading Aloud task. During this task, the written form of each item was presented to the participant on the screen and the participant was asked to say aloud the pronunciation of each item. Each item was presented one at a time on the screen for a maximum of 9 seconds, or until the participant had made their response and pressed the spacebar. Following the participant's response, which was audio recorded by the E-Prime software, the participant heard a pre-recorded correct pronunciation of the item. The 30 items were presented in a random order and their frequency followed a Zipfian distribution (as described in *Section 4.3.2*).

Following the Reading Aloud task, participants carried out a Spelling task. During this task, each participant was presented with the auditory form of each exposure item and was asked to spell the written form of the item by using mouse clicks to select letters from a matrix on the screen. The matrix contained all letters from the artificial language (11 consonants and 3 vowels). Selected letters appeared at the top of the screen, and after three letters had been selected for the spelling of each item, the correct written form of the item was presented. Feedback on whether the participant's selected letters were correct or not was also presented on the screen. The 30 exposure items were presented in a random order and their frequency followed a Zipfian distribution. This task concluded the training phase of the experiment.

The testing phase of the experiment immediately followed the training. This phase began with a Generalisation reading aloud task. During this task, the participant was presented with the written form of 30 untrained items and asked to say aloud the pronunciation of each item. Each item was presented one at a time on the screen for a maximum of 9 seconds, or until the participant had made their response and pressed the spacebar in order to proceed to the next item. Participants failed to respond within 9 seconds in 0.3% of trials overall. The 30 new items were also three-letter nonwords with a CVC structure, using the 11 consonants and 3 vowels of the artificial language. Each vowel letter appeared in 10 items. The 30 items were presented one at a time in a random order. No feedback was given to participants during the task. This task tested the participants' ability to generalise their newly-acquired knowledge of the novel vowel letter pronunciations to untrained items, offering an opportunity to assess their rule-learning ability.

Following the Generalisation task, adult participants carried out an Old/New task. In this task, participants were presented with the written form of the 30 original exposure items and 30 novel items that they had not previously encountered during the experiment, but which were composed from the same set of consonants and vowel symbols. Each item was presented one at a time on the screen in a randomised order. Participants were asked to indicate using a keyboard response whether each item had been encountered previously in training (Old) or was a novel item (New). Participants were asked to press the Z key on the keyboard when they considered the item to be Old and the M key when they considered the item to be New. Each item was presented one at a time on the screen for a maximum of 9 seconds, or until the participant had made their response. In no experimental trials throughout this task did a participant fail to respond in time. No feedback was given to participants during the task. This task assessed participants' recognition memory of the 30 original trained items. Child participants did not complete the Old/New task, progressing directly from the Generalisation task to the final Reading Aloud task.

The testing phase concluded with a final Reading Aloud test of the 30 original trained items. Participants were presented with the written form of all 30 items and asked to read aloud the correct pronunciation of each item. Items were presented on the screen one at a time for a maximum of 9 seconds, or until the participant had made their response and pressed the spacebar to proceed to the next item. In only one experimental trial throughout this task did a participant fail to respond in time. No feedback was given to participants during the task. This task assessed how accurately each participant had learned the pronunciation of each of the 30 original exposure items.

4.4 Results

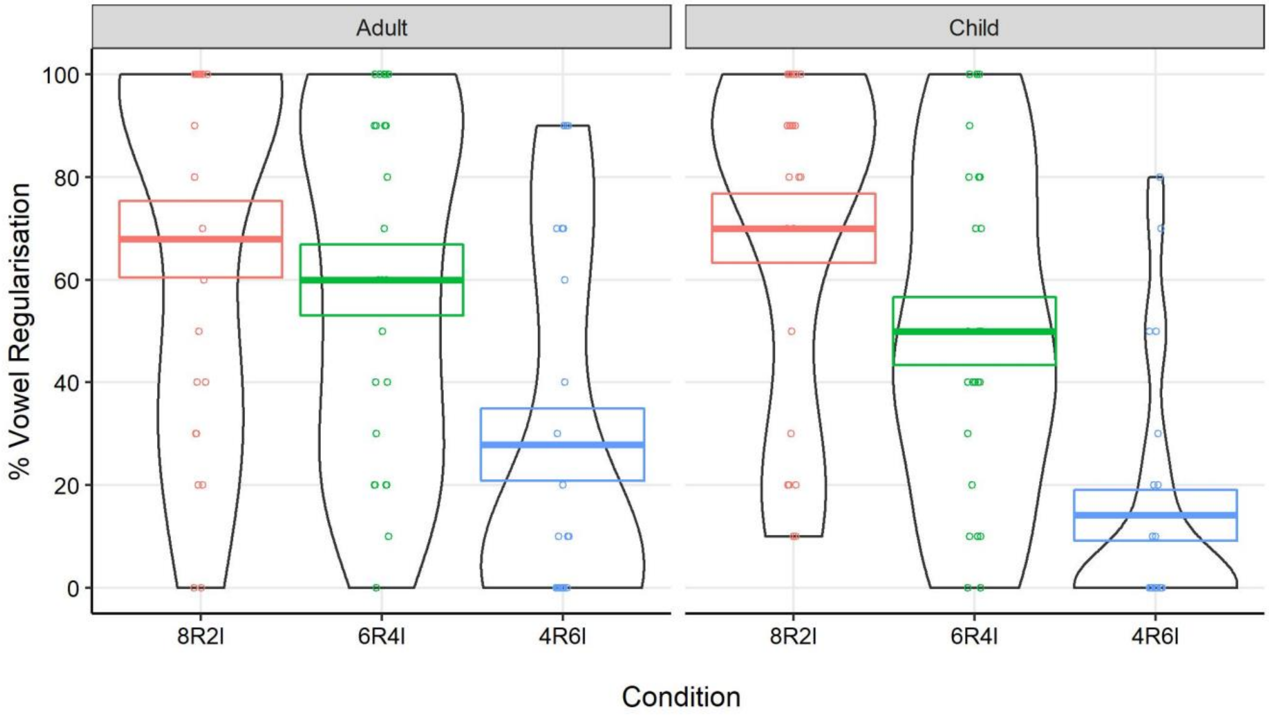
Results from adult and child participants are analysed separately, as adults and children demonstrated significantly different regularisation behaviour in the Generalisation task (see *Section 4.4.6* below) and significantly different levels of accuracy in the Reading Aloud task (see *Section 4.4.7* below). Results from adult participants are presented first, followed by results from child participants, including a comparison of adults' and children's behaviour in the Generalisation task and final Reading Aloud task.

4.4.1 Vowel regularisation by adults in the Generalisation task

In the Generalisation task, each participant read aloud 30 untrained items. Figure 4.1 presents participants' regularisation of the vowel symbol (use of the most common vowel phoneme by type) in pronunciations of these untrained items, by condition. Adults' mean vowel regularisation for untrained items across all conditions was 51.98% (*SE* = 4.34). Analysis of participants' vowel regularisation involved a series of mixed-effects logistic regression models. For this and further analyses below, I used R (version 3.6.0; R Development Core Team, 2019) and the *lme4* package (version 1.1-21; Bates et al., 2015). This approach allows predictors as fixed effects and participant as a random effect to be included simultaneously in the same models. The *p*-values reported are based on the Wald Z statistic for each effect (Jaeger 2008). A maximal random effects structure was sought in each model (following Barr et al., 2013). When a model failed to converge, the random effects structure was simplified until the model converged.

Figure 4.1

Adult and Child Participants' Vowel Regularisation (%) in Untrained Items by Condition in the Generalisation Task



Note. In this figure and subsequent figures, the horizontal line represents the mean, the box around the mean represents standard error, data points represent individual participants, and the borders around data points are smoothed density curves.

For this analysis, regularisation of the vowel in the pronunciation of each untrained nonword item was treated as the binary outcome variable, coded as 1 if the vowel was regularised and 0 for any other response. To compare regularisation by condition, I used a model with condition as the fixed effect (rotating each condition as the reference level) and participant as a random effect. Item was not added as a random effect as it resulted in a model with singular fit, due to an overly complex random effects structure. Table 4.3 presents the results from this model. Adults' regularisation was lower in the 6R4I condition than the 8R2I condition, although the related *p* value approached the threshold of significance set at .05. Therefore, it is difficult to conclude whether adults' behaviour contradicts the TP prediction that regularisation in these conditions (which both pass the tolerance test) should be similar. Meanwhile, regularisation in the 4R6I condition was significantly lower than both the 8R2I and 6R4I conditions. These results are in line with the TP prediction, according to which regularisation should be lower in the condition that does not pass the tolerance test than in the two conditions which do.

Results from one-sample t-tests suggest adults' percentage vowel regularisation does not reach 100% in the 8R2I condition (*t*(23) = -4.31, *p* < .001) or the 6R4I condition (*t*(23) = -5.87, *p* < .001). This is contrary to the TP prediction that regularisation will be categorical with participants using a productive rule 100% of the time for items in these conditions, and suggests that regularisation behaviour is not directly based on the TP alone. As the TP predicts, regularisation in the 4R6I condition was significantly lower than 100% (*t*(23) = -10.27, *p* < .001). These results do not provide evidence in support of the rule-based DRC model of word reading, which would predict that the most frequent, regular pronunciation type should be used at the same rate of 100% in all conditions.

Table 4.3

Output from mixed-effects model comparing the effect of condition on adults' vowel regularisation in the Generalisation task

glmer(Regularisation ~ Condition + (1|Participant), family = binomial)

	Estimate	Standard error	*z* value	*p* value	Inverse logit (probability)
8R2I vs 6R4I	-0.432	0.211	-2.042	.041	0.394
8R2I vs 4R6I	-2.087	0.230	-9.086	<.001	0.110
6R4I vs 4R6I	-1.656	0.221	-7.477	<.001	0.160

4.4.1.1 The effect of token frequency and the TP on vowel regularisation

Although the results above do not display the categorical generalisation behaviour predicted by the TP, it is difficult to conclude whether the TP may play an underlying role in participants' regularisation during generalisation. Additionally, this initial analysis did not involve the distribution of token frequencies in the input. The following analysis explored the effect of token frequency of regular pronunciations during training, and addressed whether the TP is able to explain regularisation behaviour beyond this input frequency variable; specifically, whether a regularised response is more likely when a spelling-sound correspondence passes the tolerance test.[11]

I used a series of mixed effects models to examine the effect of token frequency and the TP on vowel regularisation. Table 4.4 presents results from a model using token frequency of the regular vowel pronunciation during training as a fixed effect, and participant as random effect. This analysis shows that increased token frequency is associated with a greater likelihood of a regularised response.

Table 4.4

Output from mixed-effects model investigating the effect of token frequency or regular vowel pronunciations on adult participants' vowel regularisation glmer(Regularisation ~ Token + (1|Participant), family = binomial)

	Estimate	Standard error	*z* value	*p* value	Inverse logit (probability)
Intercept	-2.390	0.360	-6.641	<.001	0.084
Token	4.141	0.480	8.628	<.001	0.984

[11] Type frequency of the regular pronunciation during training was not included as an independent variable in this analysis, as a VIF value above 5 indicated that this variable has high collinearity with passing the TP. Therefore, a direct comparison between the categorical TP measure and a conventional type frequency measure as predictors of regularisation was not possible as part of this analysis.

A second model examined whether passing the tolerance test had an effect on regularisation above the effect of token frequency. This model used passing the tolerance test (coded as 1 for pass and 0 for fail) as a fixed effect. Item was not added as a random effect as it resulted in a model with singular fit, due to a too complex random effects structure. Table 4.5 presents the output from the model. This analysis shows that the TP has a significant effect on regularisation, meaning that items using vowel symbols from conditions which pass the tolerance test are significantly more likely to be regularised by participants. Token frequency also had a significant effect on regularisation.

Table 4.5

Output from mixed-effects model investigating the effect of token frequency of regular vowel pronunciations and passing the TP on adult participants' vowel regularisation glmer(Regularisation ~ Token + TP + (1|Participant), family = binomial)

	Estimate	Standard error	*z* value	*p* value	Inverse logit (probability)
Intercept	-1.097	0.469	-2.338	0.019	0.250
Token	1.589	0.748	2.123	0.034	0.830
TP	1.342	0.310	4.331	<.001	0.793

A chi-square test compared the two models to assess whether the model including the TP offers a better fit to the data than the model using only token frequency as a predictor. The result suggests that including the TP as a predictor significantly improved the model (χ^2 (1) = 19.705, p < .001), thereby indicating that the TP is able to explain variance in participants' regularisation behaviour that token frequency alone cannot.

4.4.2 Adults' recognition memory in the Old/New task

In the Old/New task, participants' mean recognition accuracy for trained ("old") items was 72.1% (*SE* = 2.87) and for untrained ("new") items, 56.4% (*SE* = 2.72). Following Merkx et al. (2011), Tamminen et al. (2012) and Tamminen et al. (2015), I also analysed this data by calculating signal detection measures (*d'*) to allow for response bias. Recognition accuracy was measured by calculating the difference between the *z*-transformed proportion of correct "old" responses to trained items (hits) and incorrect "old" responses to untrained items (false alarms). The mean *d'* value for adult participants' recognition accuracy across all items was 0.783 (*SD* = 0.073). There were no significant differences between participants' *d'* values for items in the 8R2I condition (*M* = 0.880, *SD* = 0.661) and the 6R4I condition (*M* = 0.815, *SD* = 0.118), (*t*(23) = 0.374, *p* = .712), the 6R4I condition and the 4R6I condition (*M* = 0.655, *SD* = 0.127), (*t*(23) = 1.003, *p* = .326), or the 8R2I condition and the 4R6I condition (*t*(23) = 1.234, *p* = .230).

4.4.3 Adults' performance in final Reading Aloud task

Figure 4.2 presents adult participants' percentage accuracy of trained items by condition in the final reading aloud test. Mean accuracy for trained items across all conditions was 49.58% (*SE* = 3.03). I used a mixed-effects logistic regression model to explore the effect of condition on participants' accuracy. Accuracy of the vowel pronunciation in each trained item was treated as the binary outcome variable, coded as 1 if the vowel was pronounced correctly and 0 if the vowel was pronounced incorrectly. Condition was used as a fixed effect (rotating each condition as the reference level) with random intercepts for item, and random slopes and intercepts for participant. This analysis suggests that participants' accuracy was not significantly different in the 8R2I and 6R4I conditions, or in the 6R4I and 4R6I conditions. However, accuracy in the 8R2I condition was significantly higher than the 4R6I condition.

A further mixed-effects model was used to examine the effect of vowel regularity on accuracy. Table 4.6 presents the results of a model with accuracy as the binary outcome variable, regularity as the fixed effect (using irregular as the reference level), random intercepts for item, and random slopes and intercepts for participant. The results suggest that across conditions, adult participants' accuracy was significantly higher for regular items than for irregular items.

Figure 4.2

Adult Participants’ Accuracy (%) for Trained Items by Condition in the Final Reading Aloud Task

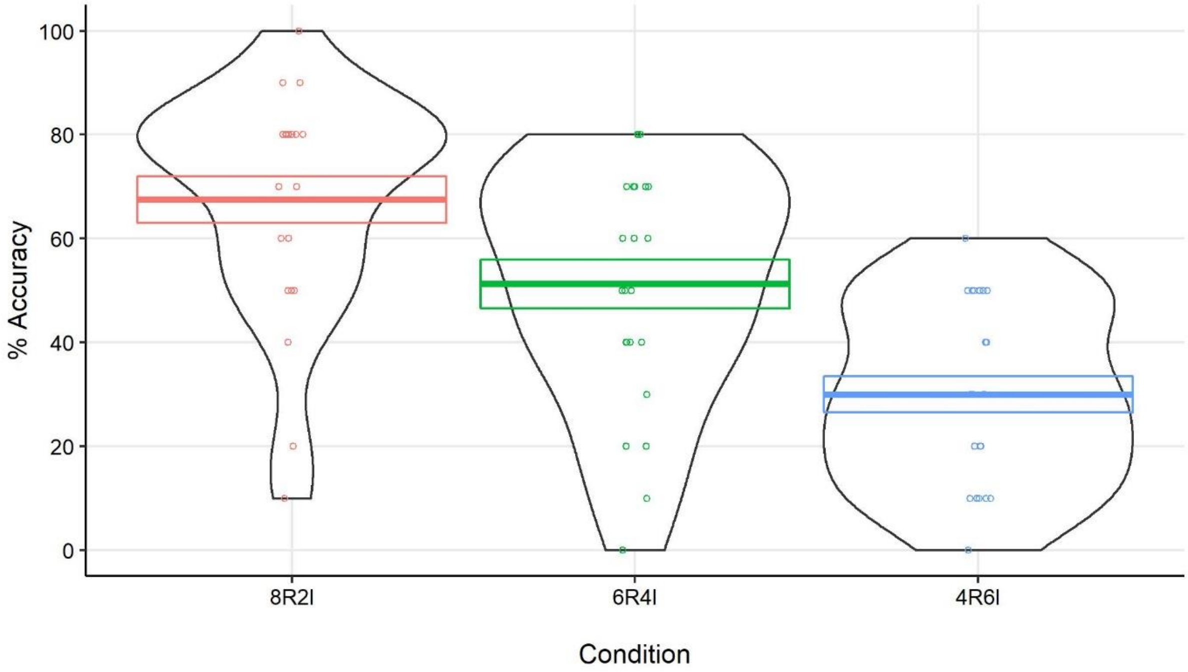


Table 4.6

Output from mixed-effects model comparing the effect of condition on adults' accuracy in the Reading Aloud task

glmer(Accuracy ~ Condition + (1+Condition|Participant) + (1|Item), family = binomial)

	Estimate	Standard error	*z* value	*p* value	Inverse logit (probability)
8R2I vs 6R4I	-0.915	0.528	-1.733	0.083	0.286
8R2I vs 4R6I	-1.932	0.519	-3.724	<.001	0.127
6R4I vs 4R6I	-1.017	0.526	-1.934	0.053	0.266

Table 4.7

Output from mixed-effects model comparing the effect of regularity on adults' accuracy in the Reading Aloud task

glmer(Accuracy ~ Regularity + (1+Condition|Participant) + (1|Item), family = binomial)

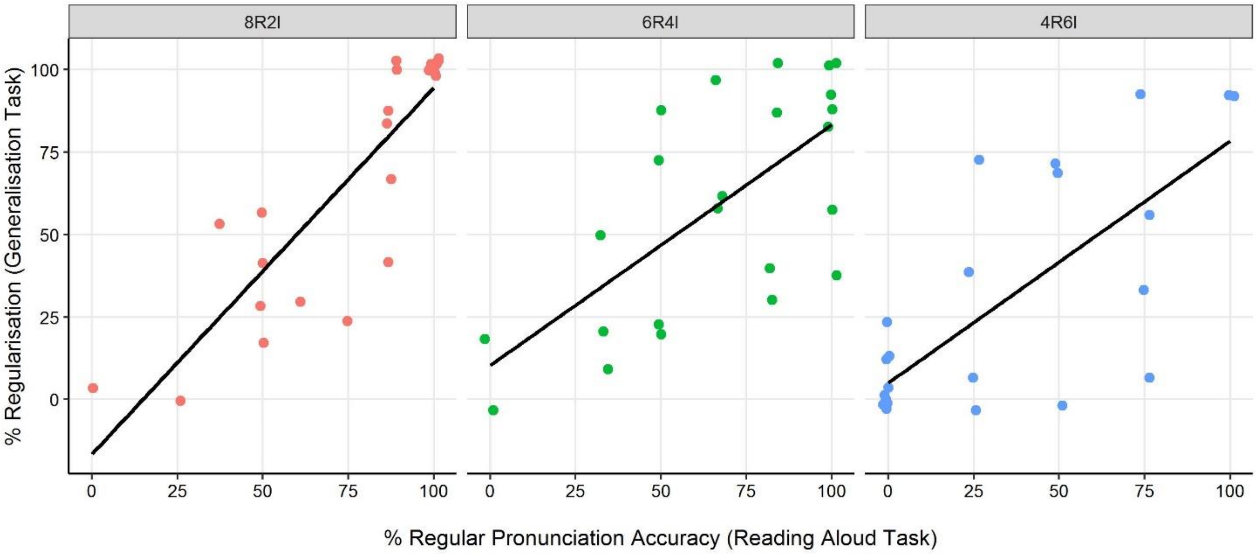
	Estimate	St. error	*z* value	*p* value	Inverse logit (probability)
Intercept	-1.035	0.281	-3.678	<.001	0.262
Irregular vs Regular	1.684	0.332	5.066	<.001	0.843

4.4.4 Relationship between adults' accuracy of regular trained items in the final Reading Aloud task and regularisation in the Generalisation Task

To study the relationship between successful learning of trained regular items and vowel regularisation, Figure 4.3 displays participants' accuracy of trained regular items in the final Reading Aloud task by vowel regularisation in the Generalisation task. Results of a Pearson correlation suggest that there is a high correlation between acquisition of regular items and regularisation during generalisation in the 8R2I condition (*r* (22) = 0.87, *p* <. 001), the 6R4I condition (*r* (22) = 0.69, *p* <. 001), and the 4R6I condition (*r* (22) = 0.76, *p* <. 001).

Figure 4.3

Adult Participants' Accurate Pronunciation of Trained Regular Items (%) in the Final Reading Aloud Task by Vowel Regularisation (%) in Untrained Items in the Generalisation Task, by Condition



4.4.5 Vowel regularisation by children in the Generalisation task

Turning next to results from child participants, children's mean regularisation of the vowel symbol pronunciation for untrained items across all conditions in the Generalisation task was 44.85% (*SE* = 3.24). Children's percentage regularisation of the vowel symbol pronunciation by condition is shown in Figure 4.1 above. Table 4.8 presents results from a model using condition as the fixed effect (rotating each condition as the reference level) and participant as a random effect. Item was not added as a random effect as it resulted in a model with singular fit, due to an overly complex random effects structure. Children's regularisation was significantly lower in the 6R4I condition than the 8R2I condition, contrary to the TP's prediction that regularisation should not differ here as the vowel passes the tolerance test in both conditions. As predicted, regularisation was significantly lower in the 4R6I condition than the 6R4I condition and the 8R2I condition. Children did not display the categorical behaviour predicted by the TP: their percentage vowel regularisation was significantly lower than 100% in the 8R2I condition (*t*(23) = -4.48, *p* < .001) or the 6R4I condition (*t*(23) = -7.55, *p* < .001), but as expected, regularisation in the 4R6I condition was lower than 100% (*t*(23) = -17.43, *p* < .001).

Table 4.8

Output from mixed-effects model comparing the effect of condition on children's vowel regularisation in the Generalisation task

glmer(Regularisation ~ Condition + (1|Participant), family = binomial)

	Estimate	St. Error	*z* value	*p* value	Inverse Logit (Probability)
8R2I vs 6R4I	-0.940	0.220	-4.596	<.001	0.281
8R2I vs 4R6I	-2.966	0.258	-11.512	<.001	0.049
6R4I vs 4R6I	-2.026	0.244	-8.288	<.001	0.117

4.4.5.1 The effect of token frequency and the TP on children's vowel regularisation

The results above do not display the categorical generalisation behaviour predicted by the TP. The following analysis examined whether the TP plays an underlying role in children's regularisation beyond the effect of token frequency of the regular vowel during training. Recall that the token frequency of regular items in the training phase varied across all participants, with items randomly assigned a position on the Zipfian distribution.

A series of mixed-effects logistic regression models investigated the effects of token frequency and passing the TP on children's regularisation in the Generalisation task. A maximal mixed-effect model did not converge, but Table 4.9 presents the output from a model using token frequency of the regular vowel during training a fixed effect and participant as a random effect. This analysis shows that increased token frequency is associated with a greater likelihood of a regularised response.

Table 4.9

Output from mixed-effects model comparing the effect of type frequency on children's regularisation in the Generalisation task

glmer(Regularisation ~ Token + (1|Participant), family = binomial)

	Estimate	St. Error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	-3.782	0.392	-9.657	<.001	0.022
Token	5.792	0.523	11.079	<.001	0.997

A second model examined whether passing the tolerance test had an effect on vowel regularisation above the effect of token frequency. This model used passing the tolerance test (coded as 1 for pass and 0 for fail) and token frequency of the regular vowel pronunciation as fixed effects and participant as a random effect. Item was not included as a random effect as it resulted in a model with singular fit, due to a too complex random effects structure. Table 4.10

presents the output from this model. This analysis shows that the TP does have a significant effect on children's regularisation, meaning that they are significantly more likely to regularise the vowel symbol in items from the conditions that pass the tolerance test. Increased token frequency is also associated with a greater likelihood of a regularised response.

Table 4.10

Output from mixed-effects model comparing the effect of token frequency and the TP on children's regularisation in the Generalisation task

glmer(Regularisation ~ Token + TP + (1|Participant), family = binomial)

	Estimate	St. Error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	-2.580	0.454	-5.678	<.001	0.070
Token	3.304	0.724	4.563	<.001	0.965
TP	1.441	0.313	4.606	<.001	0.809

A chi-square test compared the two models to assess whether the model including the TP offered a better fit to the data than the model using only token frequency as a predictor. The result suggests that including the TP as a predictor significantly improved the model ($\chi^2(1) = 21.775$, $p < .001$), thereby indicating that the TP is able to explain variance in child participants' regularisation behaviour that token frequency alone cannot.

4.4.6 Comparing adults' and children's vowel regularisation in the Generalisation task

The percentage of vowel regularisation in the Generalisation task by adults and children is provided in Figure 4.1. A mixed-effects model compared adult and child participants' vowel regularisation and the relative effect of the vowel pronunciation passing the TP (coded as 1 for pass and 0 for fail) in the Generalisation task. A maximal mixed-effects model did not converge, but Table 4.11 presents the output of the model using an Age x TP interaction as the fixed effect with age group rotated as the reference level, fail as the TP reference level, and item and participant included as random effects. This analysis shows that adults' rate of vowel regularisation was higher than children's for items with vowels that failed the TP, but not for items that passed the TP. For both adults and children, regularisation was higher for items with vowels that passed the TP than for items with vowels that did not. Additionally, the difference between regularisation in vowel pass items and vowel fail items was greater for children than adults; in other words, passing the tolerance test (or not) had a greater effect on children's vowel regularisation than on adults'.

Table 4.11

Output from mixed-effects model comparing the effect of age group and the TP on participants' vowel regularisation in the Generalisation task, with age group and passing the TP rotated as reference levels

*glmer(Regularisation ~ Age*TP + (1|Item) + (1|Participant), family = binomial)*

	Est.	St. Error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	-0.960	0.211	-4.548	<.001	0.277
Age (fail)	-0.887	0.250	-3.551	<.001	0.292
Age (pass)	-0.148	0.142	-1.037	0.300	0.463
TP (adult)	1.675	0.200	8.390	<.001	0.842
TP (child)	2.141	0.234	10.299	<.001	0.895
TP (pass)*Age (child)	0.739	0.282	2.620	0.009	0.677

4.4.7 Children's accuracy in the final Reading Aloud task

Figure 4.4 presents child participants' percentage accuracy by condition in the final Reading Aloud task. Mean accuracy across all conditions was 41.67% (*SE* = 3.15). I used a mixed-effects logistic regression model to examine the effect of condition on participants' accuracy. Accuracy of the vowel pronunciation in each trained item was treated as the binary outcome variable, coded as 1 if the vowel was pronounced correctly and 0 if the vowel was pronounced incorrectly. Condition was used as a fixed effect (rotating each condition as the reference level), with participant and item as random effects. Table 4.12 presents the results from this model. The analysis suggests that children's accuracy was significantly higher in the 8R2I condition than the 6R4I and 4R6I conditions, and significantly higher in the 6R4I condition than the 4R6I condition.

A further mixed-effects model was used to examine the effect of vowel regularity on children's accuracy. Table 4.13 presents the results of a model with accuracy as the binary outcome variable, regularity as the fixed effect (with irregular as the reference level), random intercepts for item, and random slopes and intercepts for participant. The results suggest that across conditions, child participants' accuracy was significantly higher for regular items than for irregular items. A final mixed-effects model was used to examine the effect of age group on participants' overall accuracy. A maximal mixed-effects model did not converge, but Table 4.14 presents the results of a model with accuracy as the binary outcome variable, age as the fixed effect (with adult as the reference level), and item and participant as random effects. This analysis suggests that adults' overall accuracy was significantly higher than children's.

Figure 4.4

Child Participants' Accuracy (%) for Trained Items by Condition in the Final Reading Aloud Task

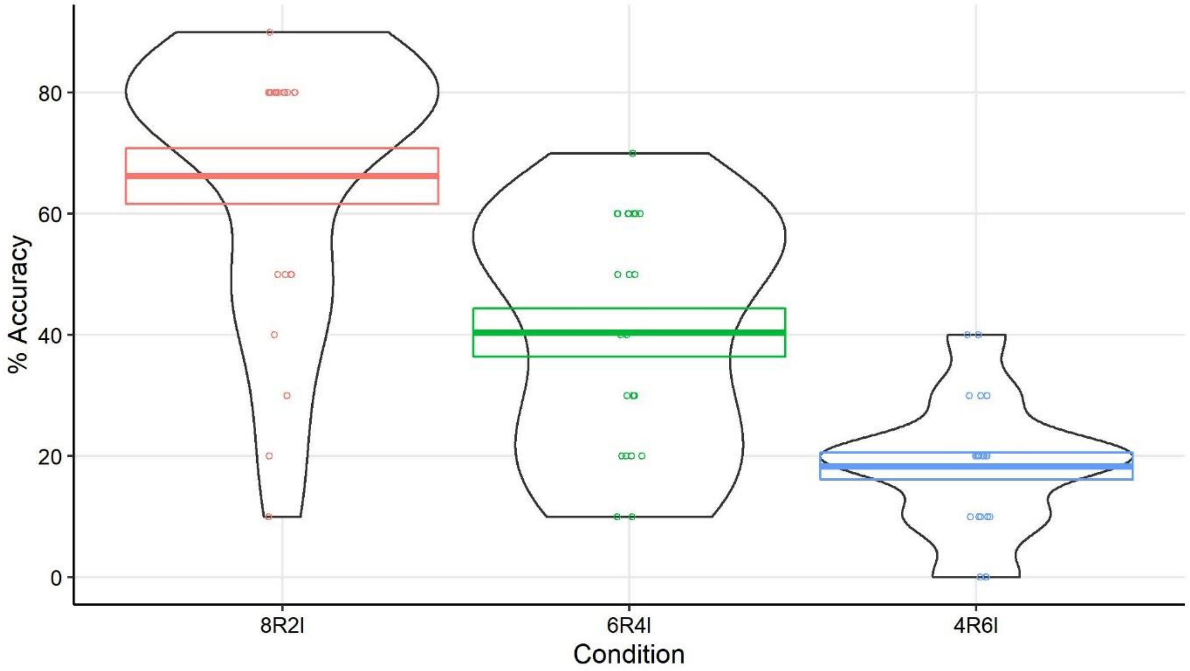


Table 4.12

Output from mixed-effects model comparing the effect of condition on children's accuracy in the Reading Aloud task

glmer(Accuracy ~ Condition + (1|Participant) + (1|Item), family = binomial)

	Estimate	Standard Error	*z* value	*p* value	Inverse Logit (Probability)
8R2I vs 6R4I	-1.298	0.518	2.046	0.012	0.215
8R2I vs 4R6I	-2.447	0.527	-4.644	<.001	0.080
6R4I vs 4R6I	-1.149	0.521	-2.204	0.028	0.241

Table 4.13

Output from mixed-effects model comparing the effect of regularity on children's accuracy in the Reading Aloud task

glmer(Accuracy ~ Condition + (1|Participant) + (1|Item), family = binomial)

	Estimate	St. Error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	-1.809	0.359	-5.044	<.001	0.141
Irregular vs. Regular	2.168	0.442	4.911	<.001	0.897

Table 4.14

Output from mixed-effects model comparing the effect of age on participants' accuracy in the Reading Aloud task

glmer(Accuracy ~ Age + (1|Participant) + (1|Item), family = binomial)

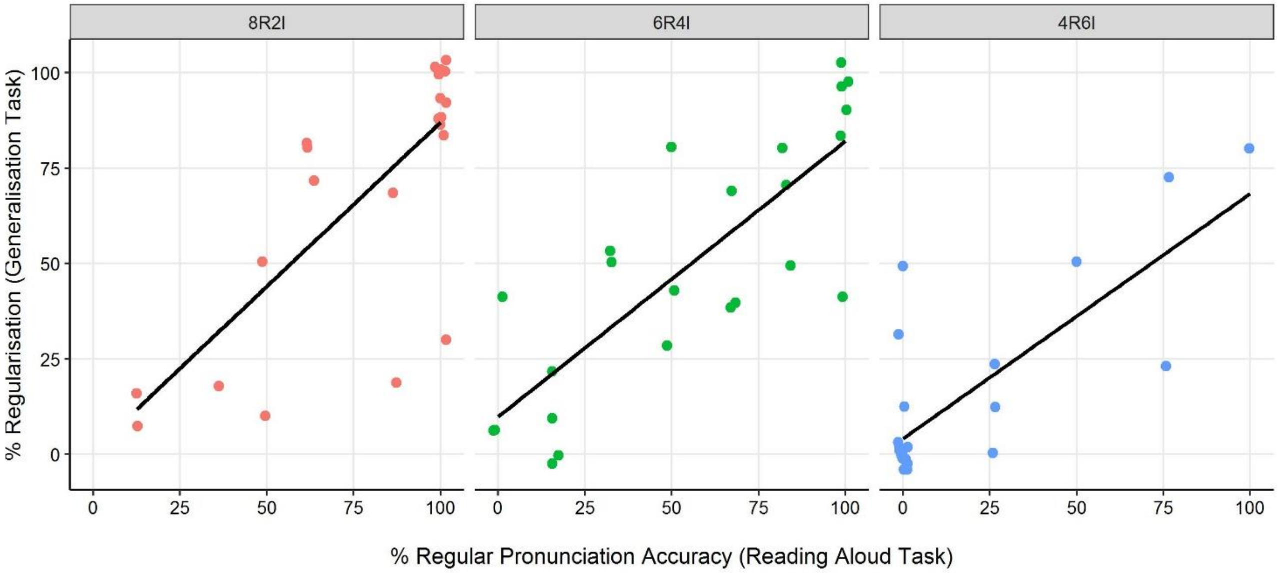
	Estimate	St. Error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	-0.018	0.269	-0.066	0.947	0.496
Adult vs. Child	-0.451	0.189	-2.382	0.017	0.389

4.4.8 Relationship between children's accuracy of regular trained items in the final Reading Aloud task and regularisation in the Generalisation task

Figure 4.5 displays child participants' accuracy of trained regular items in the final Reading Aloud task by vowel regularisation in the Generalisation task. Results of a Pearson correlation suggest that there is a high correlation between acquisition of regular items and regularisation during generalisation in the 8R2I condition (*r* (22) = 0.759, *p* <. 001), the 6R4I condition (*r* (22) = 0.816, *p* <. 001), and the 4R6I condition (*r* (22) = 0.780, *p* <. 001).

Figure 4.5

Child Participants' Accurate Pronunciation of Trained Regular items (%) in the final Reading Aloud Task by Vowel Regularisation (%) in Untrained Items in the Generalisation Task, by Condition



4.5 Discussion

The Tolerance Principle (Yang, 2016) states that leaners should form a productive rule for a particular pattern if the number of exceptions to the rule falls below a critical threshold. This rule can then be applied to novel items, allowing information gathered from the learning environment to be extended beyond a learners' direct experience. If the number of exceptions exceeds the threshold, all items should be memorised individually and no productive rule is formed that can be applied to novel items. Experimental work on the TP thus far has focused on spoken language within an artificial grammar paradigm (Schuler, 2017). Here, I ask for the first time whether the TP can capture rule-learning and generalisation of novel spelling-sound correspondences using an artificial orthography paradigm. Previous studies have found that learners in such paradigms are able to extract sub-word regularities from whole-word forms without explicit instruction and generalise this knowledge to new forms (Taylor et al. 2011; Taylor et al. 2017).

The aim of Experiment 2 was to determine whether the Tolerance Principle could predict when adult and child learners would and would not form a productive pronunciation rule for novel vowel symbols. Further, this experiment aimed to assess whether the TP has a different effect on generalisation for children and adults, in line with findings from a series of artificial grammar studies (Schuler, 2017). As part of the wider investigation into the statistical features of the input which are important for generalisation, I also considered the effects of type and token frequency of the regular vowel pronunciation on participants' vowel regularisation.

In this artificial orthography, three novel vowel symbols were used in 10 nonword items each. For a set of 10 items, the TP predicts that a productive rule can tolerate 4 exceptions. For two of the vowel symbols (in the 8R2I and 6R4I conditions), the number of irregular pronunciations falls below this tolerance threshold, meaning that a productive pronunciation rule using the most common symbol-phoneme mapping is predicted. However, for the third vowel symbol (in the 4R6I condition), the number of irregular pronunciations exceeds the threshold, and therefore the TP predicts that no productive pronunciation rule should be formed. Adult and child learners were trained to read aloud these 30 nonword items using these novel vowel symbols and were then assessed on their generalisation and learning of the spelling-sound correspondences during the testing phase.

4.5.1 Adults' generalisation

The Generalisation task allowed an assessment of whether adults and children had formed productive pronunciation rules for these three vowel symbols through analysis of their use of the regular (most common) pronunciation of the vowels in untrained items. Whilst adults' vowel regularisation did not reach the level of 100% in the 8R2I or 6R4I conditions predicted by the TP, regularisation was higher in these conditions than the 4R6I condition, as predicted. The TP also predicts that regularisation should be similar in the 8R2I and 6R4I conditions, as they both pass the tolerance test. However, it is difficult to conclude whether adults' regularisation was significantly different in these conditions. Therefore, whilst generalisation behaviour was not categorical as predicted, it is possible that the TP may still play an underlying role in generalisation. In future investigations, a power analysis would improve confidence that critical null findings are not due to an underpowered study. Meanwhile, these results do not support the DRC’s type-based approach in which the most common pronunciation of each vowel symbol should always be used productively during generalisation.

An analysis exploring the effects of token frequency and passing the TP on vowel regularisation offered support for an underlying influence of the TP on adults’ generalisation. A statistical learning account (specifically, one involving probability matching) would predict that the distribution of token frequencies in the input would be the basis for generalisation: learners should produce a range of pronunciations matching the token frequency with which they were encountered during training. Indeed, Schuler (2017) found that adults closely matched the token frequency of grammatical markers from the training phase during generalisation. Therefore, I investigated whether token frequency had an effect on learners’ regularisation, and whether the TP had an effect beyond this.

This analysis found that both token frequency and the TP (passing the tolerance test or not) had a significant effect on vowel regularisation. Additionally, the TP was able to explain variance in adults’ regularisation beyond that of token frequency. This finding suggests that whilst adults may be using information about token frequency in some way, they are not simply reproducing this frequency distribution in their generalisation. Therefore, these results do not support a statistical learning account in which adults’ generalisation is simply based on token frequency (i.e. probability matching). They suggest that adult learners are not learning veridically as previous studies in which the outcome of learning reflected the range of input variation have found (Hudson Kam & Newport, 2005, 2009; Schuler, 2017). Neither do they support established statistical models of word reading such as the Triangle and CDP+, which would again expect token frequencies during training to affect grapheme pronunciations.

In contrast, these results offer some support to the TP approach, as the TP had a significant effect on regularisation and crucially, was able to capture variance in adult participants' regularisation behaviour that token frequency cannot. This is striking evidence that the TP may be underlying adult participants' generalisation: there seems to be a shift in learner's behaviour once the consistency of a pattern crosses the tolerance threshold. In accordance with the TP approach to rule-learning and generalisation, this is a categorical rather than continuous effect. Further, this pattern of results provides new evidence that adults are able to impose some additional structure on the input during learning and generalisation: their regularisation of vowel grapheme pronunciations extended beyond the token frequencies of regular pronunciations during training, as predicted by the TP but generally not seen in adult artificial language learning studies (Hudson Kam & Newport, 2005, 2009; Schuler 2017).

4.5.2 Children's generalisation

Contrary to the predictions of the TP, children's vowel regularisation did not reach the level of 100% in the 8R2I or 6R4I conditions, and there was a significant difference in regularisation between the 8R2I and 6R4I conditions. Although regularisation was significantly lower in the 4R6I condition than the other conditions as predicted, these results do not provide evidence that there is a categorical shift in children's regularisation behaviour at the point where consistency of the pronunciation pattern crosses the tolerance threshold. Further, they do not support the DRC's type frequency rule-based approach in which the most common pronunciation of each vowel symbol should always be used productively during generalisation.

Following the findings that the TP does not directly predict children's vowel regularisation, I investigated whether the TP has an underlying effect on generalisation behaviour, in addition to any possible effect of token frequency. A statistical learning account may expect token frequency to be the basis for generalisation behaviour, as discussed above. This analysis found that both token frequency and passing the TP had a significant effect on children's regularisation. Additionally, the TP was able to explain variance in children's regularisation beyond that of token frequency. These results suggest that children's generalisation does not directly reflect the statistical distributions of the input. This pattern of results is in line with previous artificial language learning studies, which found that children tend to impose regular structure on the variability they are exposed to (Hudson Kam & Newport 2005, 2009; Schuler 2017). Indeed, results from this experiment support the hypothesis that children's generalisation behaviour goes beyond input statistics in a way that is consistent with the TP. Like Schuler, I also found a limit to children's regularisation, namely the level of inconsistency found in the spelling-sound correspondence that does not pass the tolerance test from the 4R6I condition. Overall, these findings support the possibility that the TP is a mechanism children use to guide the regularisation process: as for adults, there is evidence that it may play an underlying role in their generalisation. However, participants' regularisation behaviour was not completely categorical as the theory would predict.

4.5.3 Generalisation by children and adults

A comparison of children's and adults' behaviour in the Generalisation task revealed that adults were more likely than children to regularise pronunciations of the vowel grapheme for items with vowels that failed the tolerance test. This evidence of higher regularisation in adults may seem surprising given previous findings that children are more likely to regularise patterns from the input than adults (Hudson Kam & Newport 2005, 2009). However, this finding is a result of children's particularly low level of regularisation in the 4R6I condition, a pattern which the TP would predict as this spelling-sound correspondence does not pass the tolerance test. Indeed, the TP was found to have a greater effect on children's regularisation than adults'. This is in accordance with Schuler's (2017) findings, as well as theoretical discussion of the TP (Yang, 2016) which suggests that children are more likely to generalise using the TP than adults. Further, it gives weight to the TP approach to generalisation, which predicts precisely when input patterns should and should not be regularised, rather than the less specific prediction that children will regularise more often than adults, as previous statistical learning studies have reported (Hudson Kam & Newport 2005, 2009).

Although these results do reveal a difference between adults' and children's generalisation, we do not see the stark categorical difference in behaviour reported in Schuler's (2017) studies. This may be due to the older age of child participants in the current study (9 – 10 years) compared to those in Schuler's studies (6 – 8 years). There may also be an effect of modality, given that the current artificial orthography study investigated learning of written language (spelling-sound knowledge) rather than spoken language (morphological knowledge), as in previous artificial language studies (Hudson Kam & Newport, 2005; 2009).

4.5.4 Adults' and children's accuracy of trained items

In the final Reading Aloud test, participants were presented with the 30 original exposure items and were asked to pronounce each one individually. This task assessed whether participants had successfully learned the pronunciations of trained items. Adults' accuracy was higher in the 8R2I condition than in the 4R6I condition, with no significant differences between the other conditions. Children's accuracy was higher in the 6R4I than the 4R6I condition, and higher still in the 8R2I condition. These findings suggest that a higher number of exceptions to the most common symbol-sound mapping makes pronunciations more difficult to learn. Indeed, both adults and children demonstrated more accurate learning of regular pronunciations than irregular pronunciations across conditions, indicating again that spelling-sound correspondences for irregular items are more difficult to acquire than those for regular items. This in accordance with previous findings that systematic mappings are easier to learn (Rueckl & Dror, 1994).

4.5.5 Relationship between vowel regularisation and accuracy of trained regular items by adults and children

Adult and child learners demonstrated a high correlation between their rate of vowel regularisation in the Generalisation task and accuracy of regular trained items in the final Reading Aloud task. This pattern may be because successful acquisition of trained items is required for rule-learning and subsequent generalisation, or conversely that extraction of a pronunciation rule during training supports accurate learning of regular (but not irregular) trained items. The relationship between acquisition and generalisation will be considered further in Chapters 5 and 6.

4.6 Summary

Experiment 2 explored whether Yangs's (2016) Tolerance Principle could predict adults' and children's regularisation in an artificial orthography learning paradigm. Results from the

Generalisation task found that adults and children both produce a pattern of regularisation which is not as categorical as the TP theory would predict. However, the results did offer some support for the TP: firstly, passing the tolerance test was associated with a greater likelihood of regularisation for both adults and children. Secondly, passing the tolerance test had an effect on regularisation beyond that of the token frequency of the regular pronunciation in the input. Therefore, it seems that participants are not simply reproducing the token frequency distribution of pronunciations, but that the TP's threshold of consistency may guide participants' generalisation behaviour in a categorical way. Indeed, Yang's theory predicts that generalisation should not be based directly on token frequency, but instead reflect the TP's categorical threshold which uses type frequency counts. Experiment 3 in Chapter 5 will explore further the relationship between the TP and token frequency.

Chapter 5: Testing the Tolerance Principle in adults learning an artificial orthography with high frequency irregulars

5.1 Introduction

Results from Experiment 2 revealed that the Tolerance Principle (Yang, 2016) has an effect on adults' and children's generalisation of novel spelling-sound correspondences beyond that of token frequency input statistics. These findings suggested that there is a categorical effect on learners' generalisation behaviour when the consistency of the novel vowel grapheme crosses the tolerance threshold, which would not be expected if learners were simply reproducing the token frequency distribution of regular and irregular pronunciations from the input. However, in this first artificial orthography experiment, the token frequencies of regular and irregular items in the input were randomly allocated across a Zipfian distribution (Zipf, 1949). Experiment 3 investigates whether a high frequency of irregular items during training moderates the effect of the TP on participants' pattern of generalisation observed in Experiment 2.

5.1.1 High frequency of irregular forms

It is often noted that irregular forms tend to be highly frequent in natural language (Schuler, 2017), an observation supported by the widely-studied case of English past tense verbs. High frequency of use may protect irregular forms against regularisation or analogical levelling (Bybee & Slobin, 1982); indeed, word frequency predicts the rate of regularisation of exceptions in language evolution (Leiberman et al., 2007). The path of language acquisition may drive this pattern: for the English past tense, irregular forms that are highly frequent tend to be acquired more accurately by children than lower frequency irregulars (Yang, 2016) suggesting that in general, irregular forms must be heard often in the input in order to be learned robustly (Hooper, 1976). This view is supported by Bybee and Slobin's (1982) empirical study of English irregular past tense verbs: children (aged 1–5 years and 8-10 years) demonstrated that low-frequency irregulars are over-regularised during acquisition, whilst adults also over-regularised low frequency irregulars during production when under time pressure. The authors concluded that frequency was an important variable for both learning and maintaining irregular forms.

In written English, a similar trend between frequency and irregularity can be observed: many words with irregular pronunciations are highly frequent. Indeed, Solity and Vousden (2009) reported that 39 of the 100 most frequent word types in the Early Reading Research programme (Solity et al., 2000; Solity & Shapiro, 2008) could not be pronounced accurately using regular GPCs (i.e., had irregular pronunciations). Solity and Vousden found that these 100 highest frequency words were the same word types as the most frequent items in an adult database, and that in fact, the 39 irregularly pronounced words accounted for 50% of all word tokens in the adult database and 59% of word tokens in the children's books database.[12] This supports the impression that irregular words are highly frequent in natural language, although that is not to say that all irregular words are highly frequent, nor that all of the most frequent words are irregular.

5.1.2 Token frequency, rule productivity and the TP

Can this typological tendency for irregular items to be highly frequent disrupt the productivity of a rule generating a regular pattern? According to Yang's Tolerance Principle, the answer is no. The algorithm that computes the threshold for the number of exceptions that can be tolerated uses only type counts of regular and irregular forms in the input. Yang (2016, p. 67) notes that a productive rule must be supported by accumulation of evidence over a "sufficient number of distinct word types"; a learner would not generalise a grammatical pattern after hearing it many times in a single context.

It should be noted that the TP theory does take into account token frequencies of words in a number of ways. Firstly, the tolerance algorithm presupposes that word frequencies follow a Zipfian distribution and uses this to determine the probability of encountering a target item. From this approximation, the time taken to access a target form either through searching a full lexical listing or by searching a list of exceptions before applying a productive rule can be compared; items which are lexically listed are ranked according to token frequency. However, as a high token frequency of irregulars would affect both access routes, the competition between

[12] The adult database was extracted from the MRC Psycholinguistic Database (Coltheart, 1981) of adult fiction and non-fiction (Kucera & Francis, 1967), with the restriction that they had a Kucera-Francis frequency of at least 1. The children's books database was constructed from the content of 66 children's books (Solity & Vousden 2009, p. 477).

their relative access times would be largely unaffected. Furthermore, Yang (2016, p. 65) notes in relation to an example of high-frequency irregulars: “because the frequencies of words drop off precipitously due to Zipf’s law, most of the computational complexity will be allocated to the top half of the lexical items anyway, such that a few exceptions located in the bottom half hardly make any difference”. Indeed, he later states: “I now believe that during the courses [sic] of rule learning, the empirical frequencies of words are ignored entirely and children *only* keep track of the effectiveness of a rule [the number of exceptions, the number of total items, and the threshold], and nothing more” (Yang, 2016, p. 76). Thus, it is strictly type, and not token, frequency which determines the balance of productivity. However, token frequency may affect which items are learned first, and thus which items children base early rule-productivity on. As a child’s vocabulary grows cumulatively with time, the pattern of productivity may change according to the number of regular and irregular items that have been acquired (as lower frequency items are gradually added to the “tabulation of productivity” (2016, p. 70)).

Some experimental work has also considered the effect of token frequency of exemplars on generalisation and rule-formation, and the relationship with the TP. In the first of a series of experiments (discussed in Chapter 4), Schuler (2017) investigated whether children form productive rules according to the most frequent form in the input counted by type (as the TP would predict) or by token. During training, the regular plural marker had high token frequency in both the 5 Regular, 4 Exceptions (5R4E) condition and the 3 Regular, 6 Exceptions (3R6E) condition, as the regular marker had been assigned to the top of the Zipfian frequency distribution. Therefore, learners who used token counts to form productive rules would regularise the plural marker (i.e. productively use the regular marker) in the 5R4E and 3R6E conditions, as it is the most frequent marker in both input conditions. However, learners who used type counts to form productive rules according to the TP would only regularise the plural marker in the 5R4E condition, where the regular marker occurs in enough item types to pass the TP. They would not regularise the plural marker in the 3R6E condition, where the regular marker does not pass the TP and a productive rule is not supported. In this way, the experimental design pitted type and token frequency against each other. Results from a generalisation test revealed that children regularise the regular plural marker in the 5R4E condition, but use it at chance level in the 3R6E condition – even though this marker appears in a high number of tokens. This finding suggests that children use type frequencies to form productive rules as predicted by the TP, rather than forming productive rules according to token frequency, or using a range of markers matching the token frequency of the input as adult participants did.

In Experiment 2, Schuler investigated the effect of increasing the token frequency of irregular plural markers during training on learners' generalisation and use of the TP. In the original artificial language from Experiment 1, the most highly frequent items all took the regular form. Noting that a more typologically and ecologically valid artificial language will have some high frequency irregular forms, Schuler modified the original language such that exceptions were evenly distributed with regular forms across the Zipfian distribution. This ensured that some (but not all) exceptions and some regular forms were highly frequent during training. This modification does not alter the TP prediction that learners will form a productive rule in the 5R4E condition but will not do so in the 3R6E condition, as the TP algorithm uses type rather than token counts (as discussed above).

As in Schuler's Experiment 1, adult learners' generalisation did not follow the TP's predictions; instead, they seemed to match the token frequency of the regular form they had been exposed to during training, in both conditions. In the 3R6E condition, children behaved as predicted by the TP and seen in Experiment 1, using the regular marker at chance level. However, in the 5R4E condition, children no longer followed the TP as they had in Experiment 1. The average use of the regular marker across participants was significantly lower than 100%, which also seemed to suggest that children's behaviour was no longer categorical. However, an inspection of individual data revealed that most children were in fact still behaving categorically; either using the regular marker 100% of the time, or no more than by chance. To investigate this split in behaviour, Schuler calculated each participants' individual tolerance threshold based on the number of trained items they had accurately learned. This follows from the reasoning that the TP is intended to apply to an individual learners' vocabulary. Most children used the regular marker in accordance with their individual tolerance threshold, which Schuler argues provides evidence that the TP is a "very robust metric of productivity" (2017, p. 67) based on type rather token counts. However, it should be noted that only a subset of participants completed the rating task which measured accuracy of trained items, and the small number of children whose behaviour was not categorical were also removed from this analysis. Therefore, these results are based on data from only a small group of child participants.

5.1.3 The role of type frequency and token frequency in generalisation

Other studies have also investigated the relative effects of type and token frequency on generalisation and rule-learning. Although they did not set out to explore the TP directly, their results can be considered in light of the TP's predictions. For example, Endress and Hauser (2011) carried out a series of experiments examining the effect of type and token frequency on the acquisition of morphological patterns and exceptions. They found that type frequency was the basis of participants' generalisation of a regular pattern, whilst the learning of exceptions was supported by token frequency. This pattern of results supports the TP account of productivity, as discussed further below. Their artificial language paradigm trained adults on one of two counterbalanced inflectional patterns: either regular prefixation with some irregular suffixation or vice versa. Participants were then asked to judge whether trained regular stems, trained irregular stems, and untrained stems were more likely to take a prefix or suffix. Each of the 8 experiment versions manipulated the type or token frequency of regular and irregular affixation during training.

Experiment 1 showed that participants could successfully learn the affixation patterns when there were no exceptions. In Experiment 2, participants failed to learn the four exceptions to the affixation pattern and their learning of the regular pattern also decreased. Experiments 2-4 demonstrated that participants learned the exceptions more successfully as the token frequency of exceptions increased, but their use of the regular pattern decreased. In Experiment 5, the type frequency of exceptions was reduced but their token frequency was increased: here, participants were able to learn both the exceptions and the regular pattern. Experiment 6 suggested that this success was likely due more to the low type frequency of exceptions than their high token frequency. Experiments 7-8 revealed that high token frequency supports learning of exceptions because it provides a high absolute number of occurrences of these items (which is important for memorisation), rather than because exceptions were more frequent relative to regular items. The authors suggest that this work reveals different roles for type and token frequency in the learning of regular patterns and exceptions: type frequency seems to determine the productivity of a pattern, whilst token frequency determines how well exceptions are learned. These results accord with the TP, which predicts productivity to be based on type frequency, whilst token frequency may determine which items are learned first. Yang notes both that "generally speaking, words with higher frequencies tend to be acquired earlier” (2016, p. 70) and that “exceptions tend to be clustered at the high-frequency region of words” (2016, p. 65), which suggests that token frequency will support learning of individual items, particularly irregulars.

More widely, other research has also found little direct effect of token frequency on generalisation, as the TP would predict. Schuler et al. (2017) investigated whether lexical frequency and range of sentential contexts affects adults’ abilities to categorise novel words in an artificial language learning paradigm. They found that learners use the conditional probabilities that words will occur in certain contexts rather than the absolute (token) frequency of words and their contexts to determine generalisation of categories. Perfors et al. (2014) investigated the effect of type and token frequency on adults’ generalisation in linguistic and non-linguistic contexts, specifically whether participants update their generalisations when they encounter more tokens of familiar types. When types were kept constant but token frequency was increased tenfold, generalisation behaviour was not updated by participants in linguistic or non-linguistic contexts. The authors concluded that learners are insensitive to token frequency when extending generalisations to new items and instead base their generalisations about grammaticality on the distribution of types.

It should be noted that token frequency is important in language acquisition even if it does not affect generalisation directly. For example, Kurumada et al. (2013) examined the effects of Zipfian frequency distributions on word segmentation in an artificial language paradigm. They found that adults’ word segmentation in context – specifically, accurate learning of adjacent dependencies – was supported by a high token frequency of words in the input. They claim that a Zipfian distribution facilitates this process, as items from the top of the distribution are encountered more frequently, learned more accurately, and thus provide an efficient entry to word segmentation. I suggest this finding supports the role of token frequency found by Endress and Hauser (2011), where token frequency was important for successful learning of individual items. According to this view, (and in line with the TP), token frequency may influence which item types are learned earliest and most accurately, and early generalisation and rule-formation may be largely based on these high frequency items. However, once items have been acquired successfully, the types are not weighted by tokens in order to determine productivity of a rule.

Results from some artificial language learning studies (e.g. Hudson Kam & Newport, 2005, 2009; Schuler, 2017) also suggest that token frequency may also be important in instances where learners (particularly adults) do not form productive rules after exposure to linguistic input. As discussed in Chapter 4, these studies found that when patterns in the input were not overly complex, adult learners did *not* adopt general rules that regularised the inconsistencies they were exposed to. Instead, adults may perform "probability matching" (Hudson Kam & Newport, 2005) in which their learning and generalisation reflects the token frequency distribution of a range of forms in the input. It is worth noting that the paradigms used by Hudson Kam and Newport (2005, 2009) involved probabilistic, or inconsistent, marking of items, in which one item may sometimes take one marker and at other times take a different one. This is in contrast to the current experiment, in which mappings for individual items are always consistent: items are always pronounced the same way throughout training. However, this type of consistency was also used in Schuler's (2017) experiments, where adults were also found to perform probability matching.

To summarise, irregular forms can be highly frequent in spoken and written language but this does not necessarily disrupt the productivity of rules. A successful account of rule-learning and generalisation must be able to accommodate this typological phenomenon, whilst allowing for findings that suggest rule-formation is relatively unaffected by token frequency. One possibility is that token frequency affects which item types are learned first and most accurately, but once acquired, these types are not weighted by token frequency in terms of their influence on rule productivity. The TP is based on such an account of productivity, according to which rules are determined by types rather than tokens, and therefore high token frequency irregular forms will not directly affect patterns of generalisation.

5.3 Experiment 3

The current experiment explores whether highly frequent irregular items affect rule-formation for novel spelling-sound correspondences and consequently whether the TP is able to predict generalisation in this context. In Experiment 2, regular and irregular items from an artificial language were randomly positioned along a Zipfian distribution (Zipf, 1949) to determine their frequency of occurrence during the training phase. Using the same artificial orthography paradigm, here token frequency is manipulated during training such that irregular items are assigned to the highest positions on the distribution. Therefore, the highest frequency items are always irregular. Meanwhile, the type frequency of regular and irregular items from Experiment 2 are kept constant using the same 8 Regular/2 Irregular (8R2I), 6 Regular/4 Irregular (6R4I) and 4 Regular/6 Irregular (4R6I) conditions. The TP predicts the same patterns of generalisation in both Experiment 2 and Experiment 3, as described below. Analysis of participants' generalisation of spelling-sound correspondences to untrained items will test whether these predictions are met, or whether a high token frequency of irregular items can affect generalisation. This experiment tested adult participants only, firstly because adults and children demonstrated broadly similar generalisation behaviour in Experiment 2, and secondly because it was anticipated that the high frequency irregular version of the task may be challenging for children in this age group.

5.3.1 Predictions

Yang's Tolerance Principle (2016) states that the productivity of a rule is determined by the number of regular and irregular items in the input; the token frequency of these items will not directly affect rule productivity. Thus, according to the TP, increasing the token frequency of irregular items in the input will not affect learners' generalisation of a rule.

In this experiment, the token frequency of irregular items was increased during training, by assigning these items to the top of the Zipfian frequency distribution. Therefore, irregular items were highly frequent in the input. During the Generalisation task, adult participants read aloud a set of 30 untrained items to reveal whether they had formed a productive rule for each novel vowel symbol. If a participant responded using the most common pronunciation of the vowel grapheme by type (i.e., they "regularised" the vowel) then it suggests they had formed a productive pronunciation rule for that novel symbol.

As the number of regular and irregular item types in each condition was identical to Experiment 2, the TP makes the same predictions as the previous study. Namely, learners will regularise their pronunciation of the vowel for 100% of generalisation items from the two conditions in which a productive rule should be formed as the vowel grapheme passes the tolerance test (8R2I and 6R4I), and will use the regular pronunciation no more than chance level (25%) for items from the condition in which a productive rule should not be formed as the vowel grapheme does not pass the tolerance test (4R6I). The TP is predicted to have the same effect on vowel regularisation across the two studies, and token frequency is not predicted to have an effect on vowel regularisation.

In contrast, a probability matching account from statistical learning frameworks (Hudson Kam & Newport, 2004, 2005) expects learners to reproduce the statistical distributions from their learning environment, matching the token frequency of regular and irregular pronunciations from their input during generalisation. Therefore, this approach would predict token frequency of the regular vowel during training to have an effect on vowel regularisation.

Finally, to explore the importance of successful acquisition for rule-learning, the relationship between accurate learning of trained regular items assessed in the Final Reading Aloud test and the rates of vowel regularisation for untrained items in the Generalisation task is considered.

5.4 Method

5.4.1 Participants

25 adult participants (mean age 20; 19 females) were recruited from the student body of Royal Holloway, University of London. Participants were monolingual, native English speakers, with a Southern British accent and no known language or learning difficulties. Participants had normal or corrected-to-normal vision. Each participant received £10 for their involvement in the study. One participant was excluded due to not fulfilling eligibility requirements. Therefore, data from 24 participants were included in our analysis. The study received approval from the procedures of the Ethics Committee at Royal Holloway, University of London.

5.4.2 Stimuli and design

The stimuli and design replicated that of Experiment 2, except that items were not randomly assigned a position on the Zipfian distribution (Zipf, 1949) to determine their frequency of occurrence during training. Instead, irregular items (from all conditions) were assigned to the six positions at the top of the frequency distribution. In a Zipfian distribution, frequency is inversely proportional to rank, meaning that a small number of items occur with high token frequency and a large number of items occur with low token frequency. Allocating irregular items to the top of this distribution ensured that the high frequency items were always irregular. The allocation of 6 out of the 12 irregular items to the top six positions on our Zipfian distribution was varied for each participant, such that the allocation of the highest token frequency (24 repetitions during training) was evenly distributed across irregular items from all three conditions. This minimised an uneven weighting of extremely high token frequencies in any particular condition across participants. Allocation of irregular items to the remaining five highest token frequencies in our distribution (12, 8, 6, 5 and 4 repetitions) was randomised for each participant. All regular items and the remaining six irregular items were allocated to positions at the bottom of the distribution, and thus occurred three times each during the training phase (once each during the Exposure, Reading Aloud and Spelling tasks). An example of one participant's token frequency of exposure to each item during training is shown in Table 5.1. The token frequencies of exposure to items for all participants during the training phase are available here: https://osf.io/7xe4f/view_only=32f5e1d0b3fd437997c78990cf676566.

Table 5.1

The orthography, pronunciation, regularity, condition and token frequency of trained items for one participant. Top six frequencies in bold.

Orthography	Pronunciation	Regularity	Condition	Token Frequency
mϙl	**/mel/**	**Irregular**	**8R2I**	**24**
bʕp	**/buːp/**	**Irregular**	**4R6I**	**12**
vʕd	**/ved/**	**Irregular**	**4R6I**	**8**
tÞb	**/teb/**	**Irregular**	**6R4I**	**6**
nϙm	**/nuːm/**	**Irregular**	**8R2I**	**5**
dʕv	**/dæv/**	**Irregular**	**4R6I**	**4**
gÞv	/gæv/	Irregular	6R4I	3
kÞf	/kæf/	Irregular	6R4I	3
pÞg	/puːg/	Irregular	6R4I	3
kʕk	/kek/	Irregular	4R6I	3
lʕt	/læt/	Irregular	4R6I	3
fʕn	/fuːn/	Irregular	4R6I	3
pϙb	/pɪb/	Regular	8R2I	3
bϙp	/bɪp/	Regular	8R2I	3
kϙg	/kɪg/	Regular	8R2I	3
gϙn	/gɪn/	Regular	8R2I	3
tϙv	/tɪv/	Regular	8R2I	3
lϙf	/lɪf/	Regular	8R2I	3
fϙd	/fɪd/	Regular	8R2I	3
vϙk	/vɪk/	Regular	8R2I	3
lÞn	/lɒn/	Regular	6R4I	3
nÞp	/nɒp/	Regular	6R4I	3
vÞk	/vɒk/	Regular	6R4I	3
fÞd	/fɒd/	Regular	6R4I	3
mÞt	/mɒt/	Regular	6R4I	3

dÞm	/dɒm/	Regular	6R4I	3
pʕb	/pi:b/	Regular	4R6I	3
mʕg	/mi:g/	Regular	4R6I	3
gʕl	/gi:l/	Regular	4R6I	3
tʕf	/ti:f/	Regular	4R6I	3

5.4.3 Procedure

The procedure replicated that of Experiment 2 (with adults) detailed under *Section 4.3.3*. The training phase began with an exposure to the set of 30 items, in which participants were presented with the written form of each item one at a time on a screen, and simultaneously heard a pre-recorded pronunciation of the item. Participants then carried out a Reading Aloud task and a Spelling task, during which they received feedback on their responses. The token frequency of items in these tasks followed a Zipfian distribution (as described above).

The testing phase immediately followed the training phase. In the Generalisation task, participants were asked to read aloud 30 novel untrained items. This task assessed participants' generalisation of the novel spelling-sound correspondences encountered during training. Next, participants carried out an Old/New task in which they were presented with the written form of 30 trained items and 30 novel untrained items that they had not previously encountered. The novel untrained items were formed using the same consonant and vowel characters used for the trained items, and using the same CVC structure. Participants were asked to indicate whether or not each item had been encountered previously during training. This task assessed participants' recognition memory of the 30 original trained items. The testing phase concluded with a final Reading Aloud test of the original 30 trained items. This task assessed how accurately participants had learned the pronunciation of each of the 30 trained items. No feedback was given to participants during the testing phase.

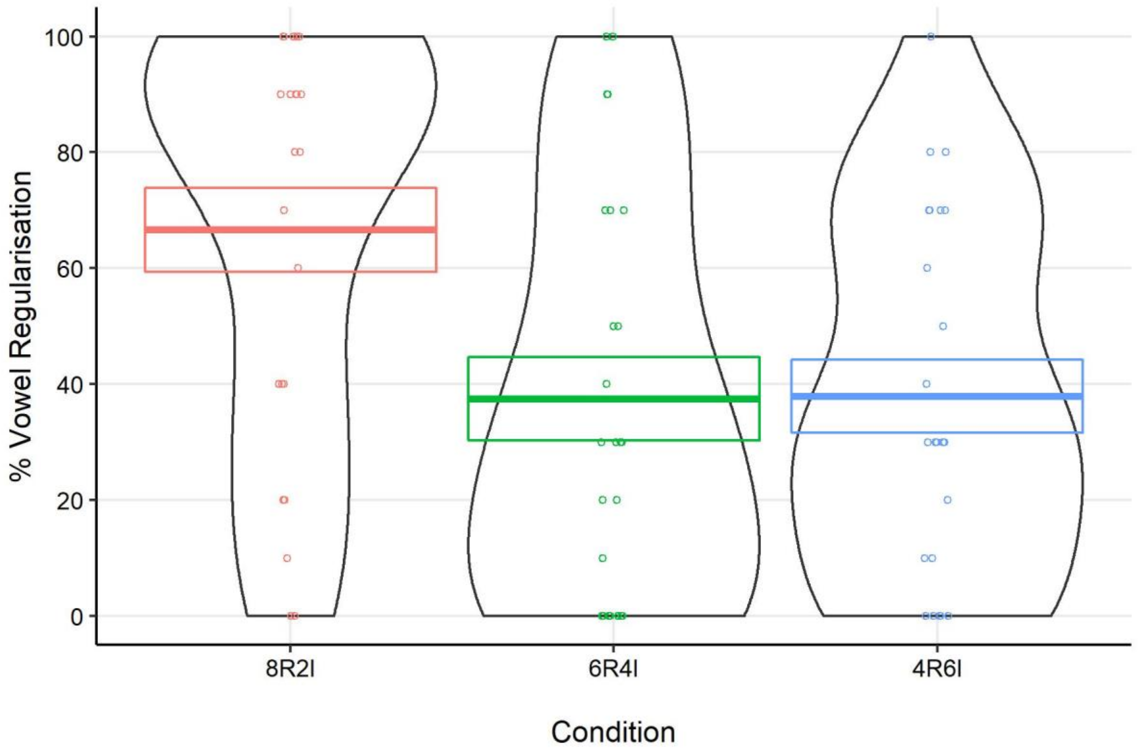
5.5 Results

5.5.1 Vowel regularisation in the Generalisation task

In the Generalisation task, each participant read aloud 30 untrained items. Figure 5.1 presents participants' percentage regularisation of the vowel symbol (use of the most common vowel phoneme by type) in pronunciations of these untrained items, by condition. Participants' mean vowel regularisation for untrained items across all conditions was 47.42% (*SE* = 3.77). For the analysis of participants' vowel regularisation, a series of mixed-effects logistic regression models was used. Regularisation of the vowel in the pronunciation of each item was used as the binary outcome variable, coded as 1 if the vowel was regularised and 0 for any other response. To compare regularisation by condition, I used a model with condition as the fixed effect (rotating each condition as the reference level) and participant as a random effect. Item was not added as a random effect as it resulted in a model with singular fit, due to an overly complex random effects structure. Table 5.2 presents the results from this model. This analysis suggests that regularisation was significantly lower in the 4R6I condition than the 8R2I condition, but was also significantly lower in the 6R4I condition than the 8R2I condition. There was no significant difference in regularisation between the 6R4I and 4R6I conditions. These results are not in line with the TP prediction that participants' regularisation will not differ in the 8R2I and 6R4I conditions which both pass the tolerance test, and will be significantly lower in the 4R6I condition which does not. Further, participants' regularisation is not categorical as the TP would predict: regularisation was significantly lower than 100% in the 8R2I condition (*t*(23) = -4.60, *p* < .001), the 6R4I condition (*t*(23) = -8.74, *p* <. 001) and the 4R6I condition (*t*(23) = -9.93, *p* < .001); this is only predicted for the 4R6I condition.

Figure 5.1

Participants' Vowel Regularisation (%) in Untrained Items by Condition in the Generalisation Task



Note. In this figure and subsequent figures, the horizontal line represents the mean, the box around the mean represents standard error, data points represent individual participants, and the borders around data points are smoothed density curves.

Table 5.2

Output from mixed-effects model comparing the effect of condition on vowel regularisation in the Generalisation task

glmer(Regularisation ~ Condition + (1|Participant), family = binomial)

	Estimate	Standard Error	*z* value	*p* value	Inverse Logit (Probability)
8R2I vs 6R4I	-1.364	0.207	-6.596	<.001	0.204
8R2I vs 4R6I	-1.351	0.206	-6.546	<.001	0.206
6R4I vs 4R6I	0.013	0.200	0.066	0.948	0.503

5.5.1.1 The effect of token frequency and the TP on vowel regularisation

The analysis above suggests that the TP does not guide participants' vowel regularisation during generalisation when irregular items are highly frequent in the input. Further, the finding that there is no significant difference between regularisation in the 6R4I condition and the 4R6I condition suggests that participants are also not simply matching the type frequency of the regular form in each condition. To explore what statistical information from the input participants might instead be using to inform their generalisation, the following analysis investigated the effect of token frequency of regular pronunciations during training on participants' regularisation, plus the effect of the TP beyond this input frequency variable.

For this analysis, I ran series of mixed-effects logistic regression models using regularisation of the vowel as the binary outcome variable. Table 5.3 presents the output from a model using token frequency of the regular vowel pronunciation during training as the fixed effect, and participant and item as random effects. This analysis shows that increased token frequency is associated with a greater likelihood of a regularised response.

Table 5.3

Output from mixed-effects model examining the effect of token frequency on vowel regularisation

glmer(Regularisation ~ Token + (1|Participant) + (1|Item), family = binomial)

	Estimate	Standard Error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	-1.612	0.313	-5.159	<.001	0.166
Token	3.310	0.535	6.187	<.001	0.965

A second model examined whether passing the tolerance test had an effect on regularisation above the effect of token frequency. This model used passing the tolerance test (treated as a factor with two levels) and token frequency as fixed effects, and participant and item as random effects. Table 5.4 presents the output from this model. This analysis suggests that the TP does not have a significant effect on regularisation, whilst an increase in token frequency is associated with a greater likelihood of regularisation.

Table 5.4

Output from mixed-effects model comparing the effects of token frequency and passing the TP on regularisation

glmer(Regularisation ~ Token + TP + (1|Participant) + (1|Item), family = binomial)

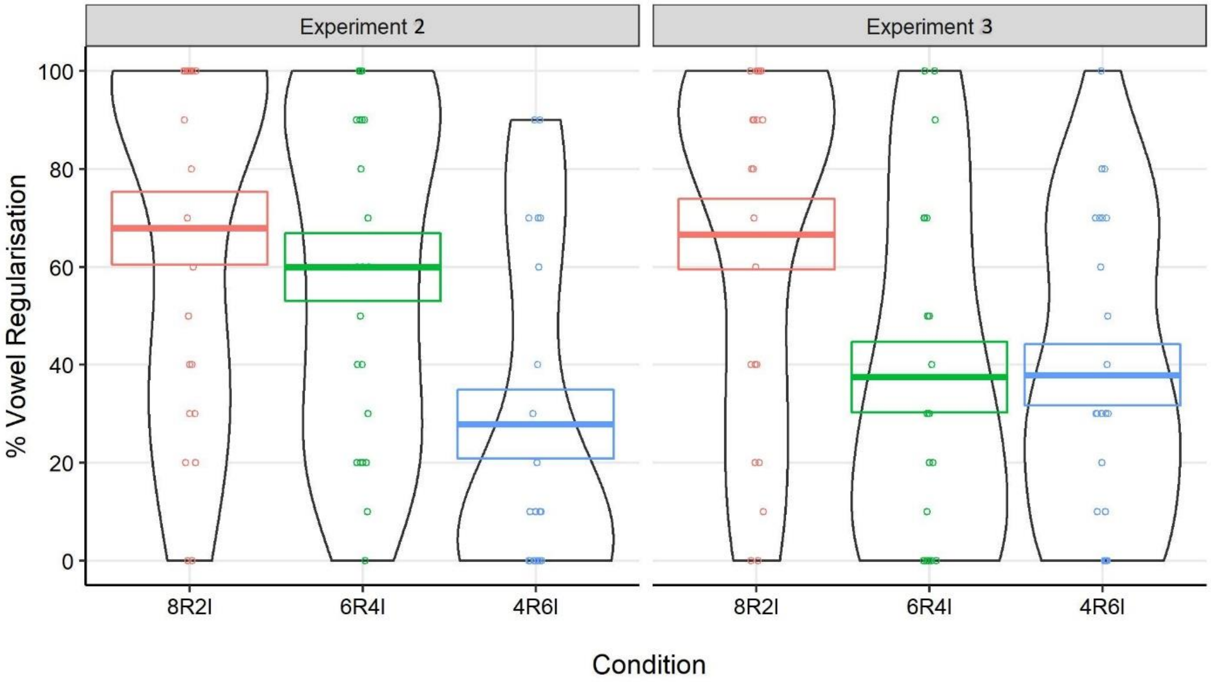
	Estimate	Standard Error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	-1.912	0.361	-5.298	<.001	0.129
Token	4.150	0.745	5.569	<.001	0.984
TP	-0.439	0.268	-1.643	0.100	0.312

5.5.1.2 Comparing adults' vowel regularisation in the Generalisation task in Experiment 2 and Experiment 3

To compare adult participants' behaviour in Experiments 2 and 3, Figure 5.2 presents participants' percentage of vowel regularisation responses by condition in both versions of the study. The analysis above showed that the TP was not associated with an increased likelihood of regularisation in Experiment 3 as it was in Experiment 2.

Figure 5.2

Participants' Vowel Regularisation (%) in Untrained Items by Condition in the Generalisation Task in Experiments 2 and 3



To explore this further, a mixed-effects logistic regression model examined whether there was an interaction between the effect of the TP and the version of the experiment on vowel regularisation. A maximal mixed-effects model did not converge, but Table 5.5 presents a model using a TP x Experiment interaction as the fixed effect (with reference levels rotated), with participant and item as random effects. The results suggest that passing the TP had a significant effect on regularisation in both experiments. Regularisation was significantly higher in Experiment 2 than Experiment 3 for items that pass the tolerance test, but there was no significant difference in regularisation between Experiments for items that fail the tolerance test. Further, the difference between regularisation for items that pass the tolerance test and items that fail the tolerance test was significantly greater in Experiment 2 than Experiment 3. In other words, passing the TP had a significantly larger effect on regularisation in Experiment 2 than Experiment 3.

Table 5.5

Output from mixed-effects model investigating the interaction between the TP and Experiment on participants' vowel regularisation

*glmer(Regularisation ~ TP*Experiment + (1|Participant) + (1|Item), family = binomial)*

	Estimate	St. Error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	-0.227	0.226	-1.006	0.314	0.443
TP (fail vs. pass) in Exp 2	1.880	0.245	7.670	<.001	0.868
TP (fail vs. pass) in Exp 3	0.673	0.227	2.973	0.003	0.662
Exp 2 vs. Exp 3 (TP fail)	0.578	0.345	1.674	0.094	0.641
Exp 2 vs. Exp 3 (TP pass)	-0.623	0.308	-2.041	0.041	0.349
TP (pass)*Exp (2)	1.207	0.262	4.604	<.001	0.770

5.6 Interim Discussion

5.6.1 Investigating the reduced effect of the TP in Experiment 3: the role of token frequency

The analyses above suggests that the TP has a significantly different effect on regularisation across the two experiments: regularisation was higher for items that pass the tolerance test in Experiment 2 than in Experiment 3, and passing the TP had a greater effect on regularisation in Experiment 2 than Experiment 3. This contradicts the TP theory, as the only difference between the two experiments was an increase in token frequency of irregular items during training, and token frequency is expected to have no direct bearing on regularisation. Nevertheless, this manipulation of token frequency reduced the predicted effect of the TP on vowel regularisation in Experiment 3 compared to Experiment 2; and, after controlling for token frequency in Experiment 3, the TP was not found to predict regularisation at all.

There are (at least) two possible ways in which token frequency may be playing a role in the disruption of the TP effect on regularisation. Firstly, it is possible that information about token frequency in the input is used specifically in the middle 6R4I condition where the number of exceptions reaches – but does not cross – the tolerance threshold of 4 exceptions. It is possible that the balance between regular and irregular items is more susceptible to influence from token frequency as the number of irregular items approaches the tolerance threshold, where calculating productivity according to type could be less clear-cut. In Experiment 3, participants' regularisation in this 6R4I condition was not significantly different from the 4R6I condition, even though the spelling-sound correspondence crosses the tolerance threshold in the latter but not the former. If token frequency affects regularisation when using type counts to calculate productivity becomes less informative, it is possible that a distribution of token frequencies in which irregular items occur very frequently results in an unpredicted pattern of regularisation in this middle condition, as observed in Experiment 2. This disruption of rule-productivity by token frequency is not predicted by the TP theory as it contradicts the characterisation of generalisation as a categorical process which is governed by a specific point of consistency, beyond which use of a productive rule becomes computationally inefficient. Further, token frequency is not predicted to have any effect on rule productivity or learners' generalisation, as discussed above.

A second possibility allows a less direct role of token frequency on regularisation. Instead, token frequency during exposure may determine which items are successfully acquired by participants, as items with high token frequency are encountered repeatedly during training. Generalisations may subsequently be made on the basis of these highly-frequent, successfully-learned items. In Experiment 3, the most frequent items were always irregular. Although irregular items are harder to learn (Rueckl & Dror, 1994), their acquisition may be boosted through repetition during training, as reported by Endress and Hauser (2011). The token frequency manipulation in Experiment 3 may therefore affect an individuals' pattern of learning. This is important to note, as the TP predictions for regularisation in each condition were calculated by inputting the total number of trained items to the tolerance algorithm, under the assumption that learners will generalise using of knowledge of the full training set. However, as results from the final Reading Aloud task demonstrate, participants did not successfully acquire spelling-sound knowledge of all items.

As discussed in *Section 5.2.1*, Schuler (2017) found that regularisation by children did not follow the original predictions of the TP in Experiment 2, where the token frequency of irregular plural markers was increased during training. These predictions were also calculated assuming knowledge of the full training set. However, noting that individual children were still using the regular marker in a categorical way (i.e. either all of the time, or at chance), she hypothesised that children were not following the original TP prediction because they had not successfully learned all of the noun-marker pairings in the category of nouns. Depending on which pairings they had learned, each participant may or may not form a productive rule. Therefore, she calculated a personal tolerance threshold for each child, using the number of trained regular and irregular items they had successfully acquired. This individual threshold successfully predicted children's generalisation behaviour, and is in accordance with the notion that the TP should apply to an individual learner's vocabulary.

In the following *Further results section 5.6*, I will explore these two hypotheses for the role of token frequency in adults' regularisation in Experiment 3, namely whether it directly influences participants' regularisation behaviour at the tolerance boundary in the 6R4I condition, or whether it plays a more indirect role by influencing which items from the training set are learned most successfully and provide the specific basis of generalisation for each participant. To investigate this, I will use a revised TP variable by calculating an individual threshold for each participant, and use this to predict whether they should form a productive pronunciation rule for each vowel symbol. If the first hypothesis is correct (that token frequency directly affects regularisation), this personalised TP variable will not have a significant effect on regularisation, but token frequency will. If the second hypothesis is correct (that token frequency affects learning which provides the basis for generalisation), then the personalised TP variable will have a significant effect on regularisation, but token frequency will not. Further, this personalised TP variable will predict vowel regularisation more successfully than the original TP variable.

5.6.2 Calculating individual tolerance thresholds

The original tolerance threshold and predictions for productivity were calculated using the total number of items and the number of irregular items in each condition, according to the tolerance algorithm. This approach assumes full knowledge of all trained items in all conditions. In order to calculate a new, individual tolerance threshold for each participant for each of the three novel vowel symbols based specifically on their own acquired knowledge of trained items, I used data from the final Reading Aloud test. In this task, participants were asked to read aloud each trained item; responses were scored correct if their pronunciation of the vowel symbol matched the trained pronunciation. An accurate response was used to assume successful acquisition of that item. The total number of accurate responses by the participant in each condition was used as the value for *N* in the tolerance algorithm ($\theta N \leq N/ln(N)$), which provided every participant with an individual tolerance threshold for each of the three vowel symbols. This threshold is the number of exceptions the TP predicts a productive rule should tolerate. Next, I counted how many of a participant's correct responses to each symbol were for irregular items, which provided the number of exceptions (*e*) the participant had acquired for each vowel symbol. This value (*e*) could then be compared with the participant's own tolerance threshold (θN) for that vowel symbol. If the number of acquired exceptions (*e*) exceeded the threshold (θN), then the participant was not predicted to form a productive rule for the pronunciation of that vowel symbol. If (*e*) did not exceed the threshold, then the participant is predicted to form a productive rule.

For example, if a participant successfully acquired 6 out of the 10 trained items which use one vowel symbol, their individual tolerance threshold for this symbol is 3.25 (6/*ln*(6)). This means that 3 exceptions can be tolerated by a productive rule. If 2 out of their 6 acquired items were irregular, the number of exceptions does not exceed the threshold and the participant is predicted to form a productive pronunciation rule for this vowel symbol. In this way, I could calculate for each participant an individual TP prediction (form a productive pronunciation rule or not) for each of the three vowel symbols.

5.7 Further Results

5.7.1 The effect of token frequency and the individual TP on vowel regularisation in the Generalisation task in Experiment 3

I ran a mixed-effects logistic regression model to investigate the effects of the individual TP variable and token frequency on adults' regularisation in the Generalisation task in Experiment 3. The model used the individual TP variable (treated as a factor with two levels) and token frequency of the regular vowel pronunciation during training as fixed effects, with participant and item as random effects; a maximal mixed-effects model did not converge. Regularisation of the vowel in the pronunciation of each untrained nonword item was used as the binary outcome variable. For this analysis, I removed data from one participant whose performance on the final reading aloud test (13.3% accuracy) suggested that they should not form a productive pronunciation rule for any vowel symbol.

Table 5.6 presents the results from this model, suggesting that the individual TP variable has a significant effect on regularisation. This means that participants are more likely to use the regular pronunciation of a vowel symbol when the pronunciation passes the tolerance test according to the number of items in their acquired vocabulary. Meanwhile, token frequency was not associated with an increased likelihood of regularisation in this model.

Table 5.6

Output from mixed-effects model comparing the effects of token frequency and the individual TP on vowel regularisation

glmer(Regularisation ~ Token + Individual TP + (1|Participant) + (1|Item), family = binomial)

	Est.	Standard Error	*z* value	*p* value	Inverse logit (probability)
Intercept	-1.779	0.352	-5.054	<.001	0.144
Token	0.280	0.729	0.384	0.701	0.570
Individual TP	2.380	0.294	8.101	<.001	0.915

5.7.2 Comparing the effects of the original and individual TP on vowel regularisation

Following the results suggesting that the individual TP successfully predicts participants' vowel regularisation, a final series of mixed-effects models investigated the effects of both the original and individual TP variables on vowel regularisation. Table 5.7 presents the results of a model using the original TP as a fixed effect (treated as a factor with two levels), with participant and item as random effects. Results suggest that the original TP does have a significant effect on regularisation in this model. Table 5.8 presents the results of a model using both the original TP and the individual TP as fixed effects (each treated as a factor with two levels), with participant and item as random effects. Results from this model suggest that the individual TP, but not the original TP, has a significant effect on regularisation. A chi-square test compared the two models to assess whether including the individual TP offers a better fit to the data. The result suggests that adding the individual TP significantly improved the model (χ^2 (1) = 113.18, p < .001), indicating that the individual TP is able to explain variance in participants' regularisation behaviour that the original TP cannot.

Table 5.7

Output from mixed-effects model examining the effect of the original TP on vowel regularisation

glmer(Regularisation ~ Original TP + (1|Participant) + (1|Item), family = binomial)

	Estimate	Standard error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	-0.496	0.259	-1.911	0.056	0.378
Original TP	0.705	0.274	2.573	0.010	0.669

Table 5.8

Output from mixed-effects model comparing the effects of the original TP and individual TP on vowel regularisation

glmer (Regularisation ~ Original TP + Individual TP + (1|Participant) + (1|Item), family = binomial)

	Estimate	Standard Error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	-1.547	0.299	-5.165	0.049	0.176
Original TP	-0.344	0.287	-1.196	0.232	0.415
Individual TP	2.555	0.274	9.329	<.001	0.928

5.7.3 Recognition memory in the Old/New task

In the Old/New task, participants' mean recognition accuracy for trained ("old") items was 72.92% (*SE* = 2.48) and for untrained ("new") items, 55.00% (*SE* = 2.77). The mean *d'* value for participants' recognition accuracy across all items was 0.73 (*SD* = 0.08). There were no significant differences between participants' *d'* values for items in the 8R2I condition (*M* = 0.637, *SD* = 0.139) and the 6R4I condition (*M* = 0.758, *SD* = 0.118), (*t*(23) = -0.798, *p* = .433), the 6R4I condition and the 4R6I condition (*M* = 0.801, *SD* = 0.146), (*t*(23) = -0.347, *p* = .732), or the 8R2I condition and the 4R6I condition (*t*(23) = -1.285, *p* = .211).

5.7.4 Performance in the final Reading Aloud task

Figure 5.3 presents participants' percentage accuracy for pronunciation of trained items by condition in the final Reading Aloud test. Responses were scored correct if the pronunciation of the vowel symbol matched the trained pronunciation for that item. Mean accuracy for trained items across all conditions was 49.58% (*SE* = 2.84). I used a mixed-effects logistic regression model to examine the effect of condition on participants' accuracy. Accuracy of the vowel pronunciation in each trained item was treated as the binary outcome variable, coded as 1 if the vowel was pronounced correctly and 0 if the vowel was pronounced incorrectly. Condition was used as a fixed effect (rotating each condition as the reference level), with participant and item as random effects. Table 5.9 presents the results from this model. The analysis suggests that accuracy in the 8R2I condition was higher than the 6R4I condition, but this difference approached the significance threshold of 0.05. Accuracy in 8R2I condition was significantly higher than the 4R6I condition, whilst accuracy in the 6R4I condition and the 4R6I condition was not significantly different. A further mixed-effects model was used to examine the effect of vowel regularity on adults' accuracy across conditions. Table 5.10 presents the results of a model with accuracy as the binary outcome variable, regularity as the fixed effect (with irregular as the reference level), and participant and item as random effects. The results suggest that across conditions, participants' accuracy was significantly higher for regular items than for irregular items.

Figure 5.3

Participants' Accuracy (%) for Trained Items by Condition in the Final Reading Aloud Task

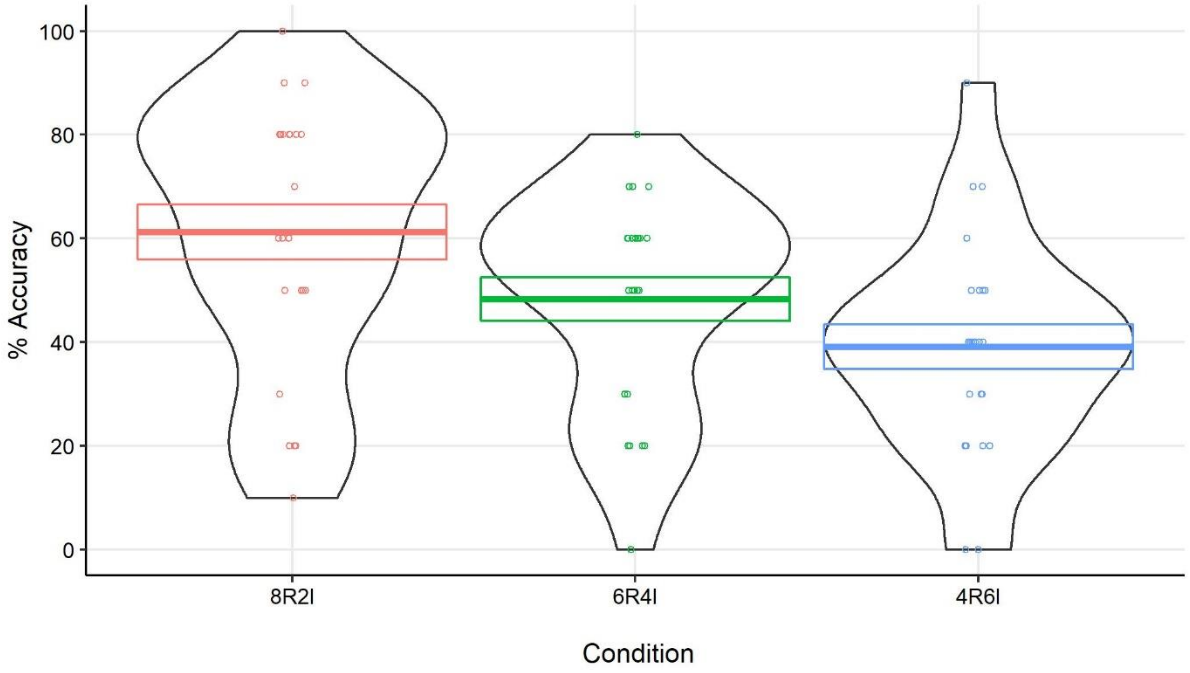


Table 5.9

Output from mixed-effects model comparing the effect of condition on adults' accuracy in the Reading Aloud task

glmer(Accuracy ~ Condition + (1|Participant) + (1|Item), family = binomial)

	Estimate	St. Error	*z* value	*p* value	Inverse Logit (Probability)
8R2I vs 6R4I	-0.584	0.296	-1.971	0.049	0.358
8R2I vs 4R6I	-0.999	0.298	-3.350	<.001	0.475
6R4I vs 4R6I	-0.416	0.296	-1.406	0.160	0.397

Table 5.10

Output from mixed-effects model examining the effect of regularity on adults' accuracy in the Reading Aloud task

glmer(Accuracy ~ Regularisation + (1|Participant) + (1|Item), family = binomial)

	Estimate	St. Error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	-0.701	0.192	-3.658	<.001	0.332
Irregular vs. Regular	1.128	0.205	5.507	<.001	0.755

5.7.5 Comparing accuracy in the final Reading Aloud task in Experiments 2 and 3

A mixed-effects model was used to examine the effect of experiment on adults' accuracy across conditions in the final Reading Aloud task. A maximal mixed-effects model did not converge, but Table 5.11 presents the results of a model with accuracy as the binary outcome variable, experiment as the fixed effect (with Experiment 2 as the reference level), and item and participant as random effects. This analysis suggests that there was no significant difference between participants' accuracy of trained items in Experiment 2 and Experiment 3.

Table 5.11

Output from mixed-effects model examining the effect of experiment on participants' accuracy in the Reading Aloud task

glmer (Accuracy ~ Experiment + (1|Participant) + (1|Item), family = binomial)

	Estimate	St. Error	*z* value	*p* value	Inverse Logit (Probability)
Intercept	-0.018	0.213	-0.083	0.934	0.496
Exp.2 vs. Exp.3	-0.004	0.194	-0.020	0.984	0.499

A final mixed-effects model examined the effect of experiment on adults' accuracy of regular and irregular trained items. Table 5.12 presents the results of a model with accuracy as the binary outcome variable, an Experiment x Regularity interaction as the fixed effect (with Experiment 2 as the experiment reference level, and regular and irregular rotated as the regularity reference level), and item and participant as random effects. This analysis suggests there was no significant difference between Experiment 2 and Experiment 3 in adults' accuracy of trained regular items or trained irregular items.

Table 5.12

Output from mixed-effects model examining the effect of Experiment x Regularity on participants' accuracy in the Reading Aloud task

glmer (Accuracy ~ Experiment + (1|Participant) + (1|Item), family = binomial)

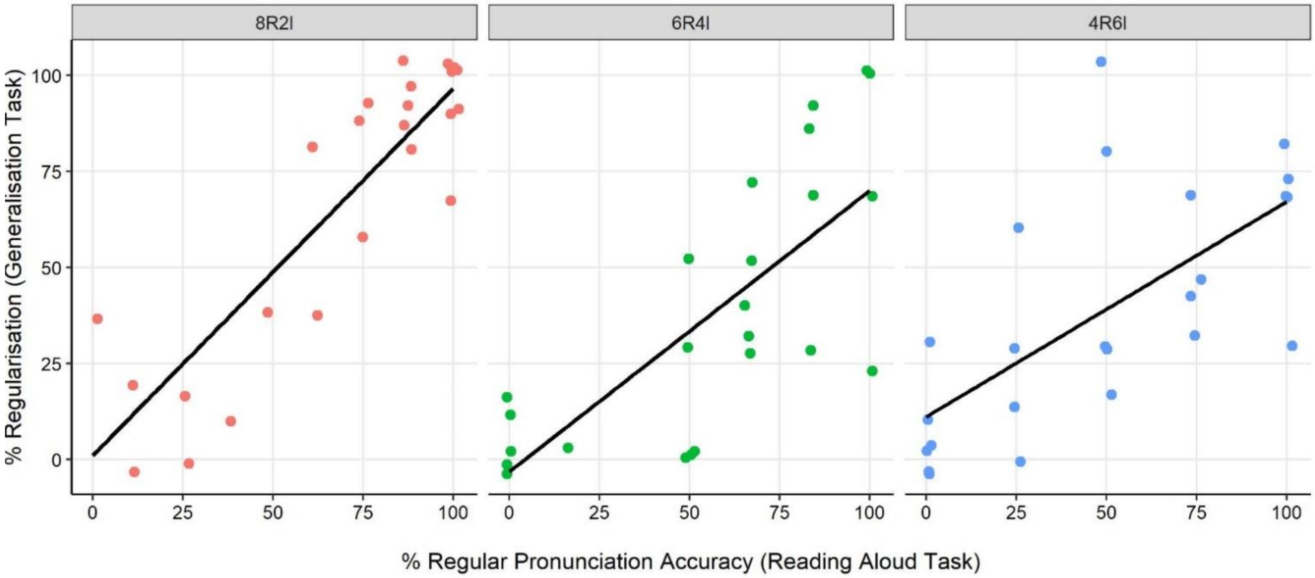
	Est.	St. Error	*z* value	*p* value	Inverse Logit (Probability)
Regular (Exp. 2 vs Exp. 3)	-0.178	0.215	-0.828	0.408	0.456
Irregular (Exp.2 vs. Exp.3)	0.279	0.245	1.136	0.256	0.569

5.7.6 Relationship between accuracy of regular trained items and regularisation in the Generalisation task

To observe the relationship between successful learning of trained regular items and vowel regularisation, Figure 5.4 displays participants' accuracy for trained regular items in the final Reading Aloud task by vowel regularisation in the Generalisation task. Results of a Pearson correlation suggest that there is a high correlation between acquisition of regular items and regularisation during generalisation in the 8R2I condition (*r* (22) = 0.888, *p* < .001), the 6R4I condition (*r* (22) = 0.737, *p* <. 001), and the 4R6I condition (*r* (22) = 0.686, *p* <. 001).

Figure 5.4

Participants' Accuracy (%) for Trained Regular Items in the Final Reading Aloud Task by Vowel Regularisation (%) in Untrained Items in the Generalisation Task



Note. Data points represent individual participants.

5.8 Discussion

The Tolerance Principle (Yang, 2016) states that learners should form a productive rule for a particular pattern if the number of exceptions to a rule falls below a critical threshold. The token frequency (frequency of occurrence) of these items in the input should not affect the productivity of a rule; the tolerance algorithm uses only type frequencies of regular and irregular items to determine the threshold. Nevertheless, it is important to consider the role of token frequency when investigating how learners extract information from the learning environment, as word frequency in natural languages follows a characteristic Zipfian distribution (Zipf, 1949): relatively few words are very frequent, whilst a large number of words occur rarely. Further, irregular words (i.e., those which do not follow a majority pattern) can be highly frequent in both the grammatical systems of spoken languages (Schuler, 2017; Bybee & Slobin, 1982) and the orthographic systems of written languages (Solity & Vousden, 2009).

Previous experimental work on the TP using an artificial grammar paradigm suggests that children do make use of type rather than token frequencies in their patterns of generalisation, (Schuler, 2017). Endress and Hauser (2011) found that adults also use type frequencies to determine productivity, whilst token frequency seems to be important for the learning of individual (and irregular) items. Meanwhile, other artificial language learning studies suggest that adult learners rarely form productive rules that impose structure beyond input frequencies at all; instead, they reproduce the pattern of variation they were exposed to during training (Hudson Kam & Newport, 2005, 2009; Schuler, 2017).

The aim of Experiment 3 was to investigate the effect of high token frequency of irregular pronunciations during training on adult learners' generalisation of novel spelling-sound correspondences in an artificial orthography. In Experiment 2, the TP's prediction for rule-productivity was found to have a significant effect on adult and child learners' vowel regularisation (use of the most common vowel pronunciation) beyond the effect of token frequency when regular and irregular items were randomised across the Zipfian distribution. Here, I considered whether the effect of the TP on adult participants' vowel regularisation revealed in Experiment 3 was maintained when irregular items were assigned to the highest positions in the Zipfian distribution, rather than being randomly distributed amongst regular items.

The TP theory predicts that the token frequency of regular and irregular items should have no direct bearing on rule productivity, which should instead be determined simply by the number of regular and irregular items as described above. As the type frequencies of regular and irregular items in each of the three conditions did not vary across Experiments 2 and 3, the TP predicts the same pattern of generalisation in the second experiment as in the first. For the two vowel symbols used in items from the 8R2I and 6R4I conditions, participants should form a productive pronunciation rule using the most common symbol-phoneme mapping. However, for the third vowel symbol used in items from the 4R6I condition, a productive pronunciation rule should not be formed as the number of irregular pronunciations exceeds the tolerance threshold. Adult learners were trained to read aloud 30 nonword items (10 from each condition) and were then assessed on their generalisation and learning of spelling-sound correspondences during the testing phase.

5.8.1 Generalisation of spelling-sound knowledge to untrained items

The Generalisation task assessed whether adults had formed productive pronunciation rules for the three vowel symbols through analysis of their use of the regular (most common) pronunciation of the vowels in untrained items. My initial analysis of the data from this task showed that vowel regularisation in the 8R2I condition was high, but did not reach the predicted level of 100%, suggesting that participants' behaviour is not categorical as the TP predicts. In the 4R6I condition, vowel regularisation was significantly lower than the 8R2I condition, as predicted by the TP. However, vowel regularisation in the 6R4I condition was significantly lower than the 8R2I condition, and not significantly different from the 4R6I condition, contrary to the TP prediction. Therefore, it seems that participants' generalisation behaviour in this middle condition does not follow the TP prediction when irregular items have high token frequency: even though the number of irregular item types does not cross the critical threshold, learners do not form a productive rule. This pattern contrasts with results in Experiment 2, where regularisation in the 4R6I condition was significantly lower than in the 6R4I condition. However, a power analysis in future investigations would improve confidence that critical null findings are not due to an underpowered study. More broadly, the finding that regularisation in line with the TP is disrupted by high token frequency irregulars mirrors that of Schuler's (2017) initial analysis of Experiment 2.

5.8.2 The effect of token frequency and the original TP on vowel regularisation

The suggestion that learners do not form productive rules as predicted by the original TP variable in Experiment 3 following a manipulation of token frequency during training was supported by the analysis exploring the effects of token frequency and passing the TP on regularisation. Whilst the TP does not predict an effect of token frequency, a statistical learning framework would predict that frequency variables from the input will have an effect on generalisation. Results from the Generalisation task suggested that token frequency did have a significant effect on regularisation, whereas passing the TP did not (after controlling for token frequency of the regular form). Critically, a comparison across Experiments 2 and 3 found a significant interaction between the effect of the TP and Experiment on vowel regularisation, with the TP having a smaller effect on regularisation in Experiment 3.

This result suggests that the manipulation of token frequency in Experiment 3 reduced the predicted effect of the TP on regularisation and triggered a change to adult learners' generalisation behaviour seen in Experiment 2, particularly in the middle 6R4I condition where the number of exceptions reaches – but does not cross – the tolerance threshold of 4 exceptions. In the *Interim Discussion,* I hypothesised two ways in which token frequency could be playing a role: the distribution of token frequencies could directly affect regularisation, particularly when the number of irregular types approached the tolerance threshold (as in the 6R4I condition) and productivity based on type counts is less clear-cut. Alternatively, token frequency could play a more indirect role by determining which items are acquired successfully during training, with these items then forming the basis of generalisation. By calculating an individual tolerance threshold for each participant according to the number of regular and irregular items trained items they had successfully acquired, I used the TP to predict whether each participant should form a productive pronunciation rule for each vowel symbol according to their individual vocabulary knowledge. My analysis showed that this personalised TP had a significant effect on regularisation beyond that of the original TP, whilst token frequency no longer had a significant effect.

This finding accords with Schuler's (2017) analysis of child generalisation data in Experiment 2: the TP can predict regularisation behaviour based on type counts when an individual learners' acquired vocabulary is taken into account. It is also consistent with the theoretical assumptions of the TP: according to Yang, "productivity is determined by two integer values [the total number of items and the number of exceptions], which are obviously matters of individual vocabulary variation" (2016, p.70); "children's productivity calculation depends on their effective vocabulary, which would be a particular subset of the input" (2016, p. 87). Therefore, it follows that the best predictor of regularisation in Experiment 3 is the application of the TP to an individual learner's vocabulary.

My analysis suggests an underlying role of token frequency in regularisation, whereby it supports learning of particular items on which generalisation is based, but does not determine productivity *per se*. This account is also consistent with Yang's theory, albeit here in relation to learning and generalisation of orthographic knowledge by adults rather than the development of spoken language by children. He notes that during language acquisition, children will use a relatively small vocabulary of high frequency items to determine rule productivity (2016, p. 70). As a child’s vocabulary grows accumulatively, their tolerance thresholds will be updated and the pattern of productivity can change accordingly. The suggestion that token frequency is important for learning individual items, particularly exceptions, but does not directly affect the productivity of a pattern, is consistent with this approach. It is also consistent with Endress and Hauser’s (2011) finding that token frequency supported learning of individual items, whilst type frequency supported rule productivity. However, the indirect role of token frequency outlined above does not align with a statistical learning approach in which adult learners probability match during generalisation, i.e. reproduce the distribution of token frequencies they are exposed to (Hudson Kam & Newport, 2005, 2009).

Overall, these results reveal that learning is an important part of the generalisation process; learners generalise beyond their input statistics, but critically they do this using information about the forms they have successfully acquired, not simply the forms they are exposed to. The statistical distributions in this input – such as token frequency - are important, but generalisation is not simply a mirror of them. Instead, they play a role in determining the patterns we are able to learn. According to the analysis above, the TP offers a successful account of the way in which learners use these patterns productively, on the basis of the specific forms they have acquired.

5.8.3 Accuracy of trained items

In the final Reading Aloud test, participants were presented with the 30 original trained items and were asked to pronounce each one individually. This task allowed an assessment of whether participants had successfully learned the pronunciations of the exposure items they had encountered during the training phase. Accuracy in the 8R2I condition was significantly higher than the 4R6I condition; the difference between the 8R2I and 6R4I conditions fell just within the significance threshold; and there was also no significant difference between accuracy in the 6R4I condition and the 4R6I condition. Across conditions, accuracy for regular items was higher than for irregular items despite the high frequency of irregular items during training, indicating that spelling-sound correspondences for irregular items are more difficult to acquire than those for regular items even when irregular items occur more frequently (although it should be noted that not all irregular items had a higher token frequency than regular items due to the number of positions available at the top of the Zipfian distribution.) These results are in accordance with previous findings showing that systematic mappings are easier to learn than mappings without a systematic relationship (Rueckl & Dror, 1994). Further, accuracy for regular and irregular items was not significantly different across Experiments 2 and 3. These results suggest that in general, the high token frequency of irregular items did not affect overall learning of all items, nor learning of irregular items, compared to Experiment 2. However, as discussed above in relation to results from the generalisation task, it is possible that individual participants' pattern of learning was affected by the specific token frequencies they were exposed to during training.

5.8.4 Relationship between vowel regularisation and accuracy of trained regular items

Adult learners demonstrated a high correlation between their rate of vowel regularisation in the Generalisation task and accurate pronunciation of regular trained items during the final Reading Aloud task in all three conditions. Despite the fact that regularisation followed a different pattern in Experiment 3 compared to Experiment 2, this result suggests that use of a productive rule still corresponds with successful acquisition of trained regular items in this context. It also adds additional support to the claim that successful learning of rule-following items is important in order to form productive rules, as discussed above.

5.9 Summary

Experiment 3 set out to explore the effect of high token frequency of irregular items during training on learners' generalisation, and in particular whether this frequency distribution moderates the effect of the TP reported in Experiment 2. Initial analysis suggested that the effect of the TP was reduced in Experiment 3, in contrast to the TP prediction that token frequency should not determine generalisation. However, subsequent analysis (following Schuler, 2017) found that an individual tolerance threshold based on each learners' successfully-acquired items offered a better predictor of generalisation than either the original TP or the token frequency of the regular form during training. This result indicated two important findings: firstly, that token frequency may indirectly affect patterns of generalisation by determining which items are learned accurately and thus used as the basis for generalisation. Secondly, that to understand an individual learner's pattern of generalisation, we must consider what that individual has gleaned from their input, rather than simply considering the distributions they have been exposed to. These findings offer support to the TP theory, where productivity is predicted according to an individual's vocabulary. They also have important implications for statistical learning research, which has previously found a more direct role for token frequency in adults' generalisation (Hudson Kam & Newport, 2005, 2009). Overall, Experiment 3 has highlighted the importance of understanding the relationship between the acquisition and generalisation of quasi-regular patterns.

Chapter 6: Testing the recursive application of the Tolerance Principle in adults learning an artificial orthography

6.1 Introduction

The aim of Experiment 4 was to examine whether the recursive application of the Tolerance Principle (Yang, 2016) could predict adults' generalisation of unfamiliar context-sensitive spelling-sound "sub-rules". I used an artificial orthography paradigm in which participants were first trained to read aloud nonword items using novel vowel symbols, and then asked to pronounce untrained items in order to capture their generalisation of the novel spelling-sound correspondences. This experiment manipulated whether the spelling-sound consistency of the vowel symbols passed the tolerance test in any consonantal context, or only in the context of a specific word-final consonant. Therefore, I could investigate whether adults follow the recursive application of the TP to form more specific "sub-rules" when predicted to do so.

6.1.1 Context sensitivity of orthography-phonology correspondences

Written English uses an alphabetic writing system in which there are systematic correspondences between letters and sounds. However, it is a deep orthography which does not just have simple one-to-one mappings between phonemes and graphemes. Instead, many graphemes can be pronounced in a number of ways. Some of this variation is associated with the co-occurrence of adjacent letters, such that the pronunciations of some graphemes can be characterised as "context-sensitive". For example, the most common pronunciation of the grapheme "oo" in English words is /u:/, however when this grapheme is followed by "k", the letter sequence is almost always pronounced /ʊk/ as in "book" and "look". Although these context-sensitive patterns are often characterised in terms of graded consistency (Seidenberg & McClelland, 1989; Plaut et al., 1996; Harm & Seidenberg, 2004), it may also be possible to categorise single grapheme-phoneme correspondences (GPCs) as the most general pronunciation rules, and these context-sensitive pronunciations as more specific "sub-rules".

Both adults and children seem to generalise knowledge of context-sensitive spelling-sound correspondences when reading aloud unfamiliar or nonword items (Glushko, 1979; Ryder

& Pearson, 1980; Treiman et al., 1990; Coltheart & Leahy, 1992; Andrews & Scarratt, 1998). For example, participants may pronounce the nonword "clead" to rhyme with "head" rather than "bead", even though the most common pronunciation of "ea" in English monosyllabic words is /i:/. Further, readers may not have been explicitly taught these context-sensitive patterns during instruction at school, but instead acquire this knowledge implicitly during their reading experience (Laxon et al., 1991).

6.1.2 The role of statistical learning in acquiring knowledge of context-sensitive orthography-phonology correspondences

The range of evidence demonstrating adults' and children's acquisition and use of context-sensitive pronunciation patterns without explicit instruction prompts questions about how readers use information from their text experience to pronounce unknown words. Many researchers suggest that readers are sensitive to statistical distributions in text. For instance, Taylor et al. (2011) trained adults in an artificial orthography paradigm. The consistency of the four novel vowel characters in the orthography varied: two were consistent, using a one-to-one grapheme-phoneme mapping; whilst two were inconsistent, pronounced in one way when preceded by a particular consonant character (inconsistent-conditioned) and in a different way when preceded by any other consonant character (inconsistent-unconditioned). After training, learners demonstrated higher accuracy for items with consistent vowels than inconsistent-conditioned vowels, and lowest accuracy for items with inconsistent-unconditioned vowels. This finding offers evidence that adults are able to extract context-sensitive sub-word regularities from exposure to whole-word items, and that learners are sensitive to the consistency of such mappings; inconsistent mappings are more difficult to acquire. Overall, the authors noted that learners have an impressive ability to track the statistical distributions of their input.

Further, some researchers have suggested that readers use statistical learning mechanisms to extract letter-sound patterns. For example, Apfelbaum et al. (2013) investigated which principles of statistical learning may support the development of phonics knowledge (i.e., knowledge of GPCs), and found that variability of consonant frames around a target vowel during training of GPCs was beneficial for children's learning. The authors suggest that this variability helps learners to identify the relevant elements and patterns in their input as part of a statistical learning process. Similarly, Samara et al. (2019) explored whether statistical learning processes are involved in spelling development. They found that children aged 6-7 were sensitive to graphotactic constraints in nonwords (i.e. contingencies between a medial vowel and word-initial or word-final consonants) following exposure without explicit instruction. They suggest that children are able to learn and generalise these constraints using information about statistical distributions in the input rather than through instruction.

Arcuili and Simpson (2012) explored the relationship between statistical learning and reading more broadly, finding that performance on a visual statistical learning task predicted word reading ability for adults and children aged 6-12, beyond the effects of age and attention. This result prompted them to highlight a potential role for statistical learning in reading development: some mappings between letters and sounds may be learned implicitly as readers develop sensitivity to contextual cues such as the co-occurrence of letters. Although their study did not explore this possibility directly, the authors suggest that this is one way in which statistical learning may support readers to acquire knowledge of the regularities between letters and sounds, which is in line with results from Apfelbaum et al. (2013) and Samara et al. (2019).

Other researchers have considered more specifically how readers use statistical learning mechanisms to extract context-sensitive spelling-sound patterns from text input, and further, how these patterns are used productively in nonword pronunciations. In a discussion of spelling development, Kessler (2009) suggests that children may pay more attention to contextual information about individual letters when there is no clear orthography-phonology correspondence that is salient across contexts. Thus, there may be a pay-off in which learning a context-sensitive rule is only motivated when a grapheme has several inconsistent pronunciations. Similarly, Steacy et al. (2019) suggest that during reading development, pronunciations of vowel graphemes are determined by a trade-off between the frequency of a context-free GPC and the strength of a context-dependent pronunciation. Therefore, it seems that readers are not simply reproducing the range of variation or the most common mappings that appear in their input. Instead, they are undertaking a process in which use of a particular pronunciation is determined by an interaction between multiple sources of statistical information.

This suggestion that readers do not simply reproduce the input statistics in their pronunciations is consistent with a finding by Treiman and Kessler (2019), who demonstrated that both child and adult readers make less use of contextual information in their pronunciation of initial consonants than would be expected given the consistency of certain patterns in the English spelling system. Their investigation focused on pronunciations of word-initial “c” and “g” which are often influenced by an adjacent “i” or “e” in English words, for example in “centre” and “generate”. They examined the pronunciation of nonwords with these initial letters by participants aged 6 to 23 years. Participants’ use of context-sensitive front pronunciations of these consonants when followed by “i” and “e” increased gradually with reading skill, which the authors suggest may be supported by the increased token frequency of these spelling-sound correspondences in words seen by older readers. The results support the view that contextual information is useful to readers, but that it takes many years of reading experience for this to approach a level that would be anticipated given the contextual effects of surrounding letters on pronunciations in written English. Furthermore, readers do not seem to be ideal statistical learners who optimally match their reading behaviour to the structure of the input, but explicit instruction may allow them to take advantage of these more complex consistencies when decoding words. Similarly, Treiman et al. (2003) note that adult participants do not use consonantal context in their pronunciations of vowels in nonwords as much as one might expect given the frequency of these patterns in real English words. Further, they found that none of the ten computational models of word reading they considered (including dual-route, single-route, rule-based and connectionist models) provided a successful account of human performance on nonwords with contextual conditioning.

Overall, research on context-sensitive grapheme pronunciations suggests that whilst readers may make use of statistical learning mechanisms to acquire these patterns through text experience, use of the context-sensitive pronunciations in nonword reading does not simply reflect the statistical distributions in the input, and is not successfully predicted by computational models of word reading (as discussed in detail in Chapter 3).

6.1.3 Beyond reading: can learners acquire sub-categories and nested structures?

Limited experimental work has investigated whether learners are able to acquire structurally-embedded rules or sub-categories in artificial language learning studies, akin to the context-specific pronunciation rules discussed above in the domain of word reading. Reeder et al. (2017) investigated whether adults are able to use distributional information to form sub-categories of items in an artificial grammar learning paradigm. The distributional cues were word co-occurrences, without any correlated semantic or phonological cues. The design of their Experiment 1 had perfect subcategory boundaries which had complete overlap of distributional contexts across words within a subcategory, and no overlap of distributional contexts for words across subcategory boundaries. Results showed that adults were able to form sub-categories under these conditions: they generalised appropriately to novel, grammatical word strings within a sub-category. Further, they were able to restrict generalisation across sub-categories by rejecting novel strings that crossed the sub-category boundary. In their Experiment 2, one exception was added to the artificial grammar in the form of a single string that crossed the subcategory boundary. Here, learners were again able to successfully form sub-categories and maintain the sub-category boundaries; the unique string was treated as an exception which did not affect learners' generalisation when compared to Experiment 1. The authors suggest that learners were able to interpret the absence of certain strings across sub-categories as purposeful omissions (due to their ungrammaticality), whilst interpreting less systematic gaps within subcategories as accidental sampling absences. Overall, they suggest that their results reveal adults' ability to use distributional information in a sophisticated way to create and generalise sub-categories, even when an exception is presented.

Udden et al. (2009) investigated adults' implicit learning of recursive sequence structures in an artificial grammar learning paradigm. Learners were exposed to letter sequences which featured recursively embedded structures (i.e. $A_1A_2A_3B_3B_2B_1$, a "nested structure") and recursive structures with cross-dependencies (i.e. $A_1A_2A_3B_1B_2B_3$, a "crossed structure"). They found that over 9 days of learning, adults were able to demonstrate learning of long-distance dependencies in both nested and crossed structures. Together, these studies offer some evidence that adult learners are able to learn more complex rules such as sub-categories and embedded structures without explicit instruction in artificial language learning paradigms.

6.1.4 The recursive application of the TP

If hierarchical structures of nested rules exist in spoken and written language systems – which they do – and learners are able to implicitly acquire knowledge of them, then any successful theory of the language acquisition process must be able to account for how learners form these complex, productive rule systems that go beyond the simple input statistics. Yang's theory offers one possibility: that using the TP in a recursive fashion allows learners to acquire "nested" rules. These are the sub-regularities that are a common feature of natural language and perhaps of opaque orthographic systems such as written English. By following the *Maximise Productivity* strategy to "pursue rules that maximise productivity" (2016, p. 72), Yang suggests failure to find a productive rule for a set of items will prompt learners to search for productivity within subsets of the items. The tolerance threshold could then apply to the number of regular and exception items within each subset. Indeed, in his discussion of spoken language acquisition, Yang supposes that learners become attuned to specific features of rules precisely because productivity will not arise without doing so.

6.1.5 Can the TP explain how learners use statistical information to form context-sensitive pronunciation sub-rules?

As discussed above, readers are sensitive to context-specific letter-sound mappings, but their productive use of these pronunciation patterns in nonwords does not necessarily match the frequency of these mappings in English words (Treiman et al. 2003; Treiman & Kessler, 2019). However, neither do readers simply use the most frequent GPCs in all instances (Glushko, 1979; Ryder & Pearson, 1980; Treiman et al., 1990; Coltheart & Leahy, 1992; Andrews & Scarratt, 1998). It is possible that there is a balance to be struck between using the most frequent, context-free individual grapheme-phoneme mappings and less frequent (but potentially more consistent) context-dependent pronunciations (Kessler, 2009; Steacy et al., 2019). Considering the evidence which suggests that some form of statistical learning is used to acquire more complex spelling-sound mappings (e.g. Arcuili & Simpson, 2013; Steacy et al., 2019; Treiman & Kessler, 2019), perhaps there is mechanism by which learners use statistical information from their text input to form a productive system of more general, context-free pronunciation rules, and more specific, context-dependent pronunciation rules. The Tolerance Principle offers one account of this process: that learners will search for the most general pronunciation rule (i.e. a context-free GPC, such as "ea" - /i:/), but if the level of inconsistency for this pronunciation breaches the tolerance threshold, the TP will be applied recursively and a more specific pronunciation rule (i.e. a context-sensitive mapping, such as “ead” – /ɛd/) will be sought within subsets of items (e.g. all items which have the letter sequence “-ead”).

In this way, readers are making use of statistical information about the frequency and consistency of alternative pronunciations in their text experience, and using this information to determine which pronunciations should be used productively when presented with unknown items to read aloud. According to Yang’s theory, this process is motivated by computational efficiency; the system of productive rules a learner develops will be the most efficient way to access and apply the patterns they have gleaned from their input. In the current Experiment, I test the recursive application of the TP in the field of reading by using an artificial orthography paradigm.

6.2 Experiment 4

Experiment 4 explored the recursive application of the TP in reading acquisition and generalisation. This was investigated using an artificial orthography paradigm in which context-free and context-sensitive spelling-sound correspondences offered learners either productive pronunciation “rules” or “sub-rules”. Adult participants learned to read a set of nonword items which used novel vowel symbols. The spelling-sound consistency of these symbols and the conditioning of their pronunciation by consonantal context was manipulated. Following training, participants were tested on trained items to assess the spelling-sound knowledge they had acquired from this artificial orthography, and critically were also tested on untrained items which used the trained vowel symbols to assess their generalisation of this knowledge.

This artificial orthography used two novel vowel symbols with inconsistent pronunciations. The consistency of these pronunciations in trained items was manipulated by condition. In the Vowel Rule condition, the consistency of the vowel grapheme pronunciation passed the tolerance test, such that the TP predicts learners should form a productive pronunciation rule for the vowel in any consonantal context. In the Body Rule condition, the consistency of the vowel grapheme pronunciation across all items did not pass the tolerance test, such that the TP predicts learners should not form a context-free productive pronunciation rule for the vowel symbol. However, the consistency of the vowel symbol pronunciation in a specific final consonantal context did pass the tolerance test, such that the TP predicts a productive context-sensitive pronunciation “sub-rule” for the vowel when it occurs in this particular word body (i.e., the orthographic unit containing a medial vowel and word-final consonant). If learners apply the TP recursively to subsets of items when a more general pronunciation does not pass the tolerance test, then they should form a productive pronunciation for this word body.

6.2.1 Hypotheses

The TP makes the following four hypotheses for participants’ generalisation behaviour in the Vowel Rule and Body Rule conditions. Predictions for generalisation in each condition made by the rule-based DRC model of word reading (Coltheart et al., 2001) are also included.[13] Firstly, participants should use the phoneme /i:/ to pronounce the vowel symbol in 100% of generalisation items in the Vowel Rule condition, as this pronunciation passes the tolerance test in trained items and therefore should be used as a productive rule in any consonantal context. Similarly, the DRC predicts that participants will use this pronunciation in 100% of these items, as this approach maintains that the most common pronunciation of a grapheme (by type) will always be used productively.

Secondly, participants should use the phoneme /u:/ to pronounce the vowel symbol in less than 100% of generalisation items in the Body Rule condition, as this pronunciation does not pass the tolerance test in trained items and therefore should not be used as a productive rule across all consonantal contexts. Further, this pronunciation should be used less often than the most common pronunciation /i:/ used in the Vowel Rule condition, where a productive rule *is* predicted.

Thirdly, participants should use the phoneme /ɛ/ to pronounce the vowel symbol in 100% of a subset of generalisation items from the Body Rule condition which end in word-final –v. This pronunciation passes the tolerance test in this specific consonantal context when the TP is applied recursively, and therefore forms a productive body sub-rule.

[13] Predictions for generalisation made by the Triangle and CDP+ models of word reading are not assessed in Experiment 4, as such predictions are not easily available for the stimuli set used.

Finally, participants should use the sub-rule /ɛ/ pronunciation in generalisation items from the Body Rule condition which have any other final consonant less often than they do in for items in the word-final –v subset, as this pronunciation does not pass the tolerance test in any other consonantal context.

In contrast, the DRC predicts that participants should use the phoneme /u:/ in 100% of generalisation items in the Body Rule condition, as it is the most common pronunciation of this symbol in the training set. There should be no difference between use of this pronunciation in subset or non-subset items.

Three further questions are explored as part of our wider investigation of adult readers' learning and use of spelling-sound mappings: i) whether participants display sensitivity to the statistical properties of the learning environment (e.g. type frequency) by matching these statistics in their generalisation or ii) whether their individual generalisation behaviour is instead categorical, and iii) to what extent generalisation is supported by accurate learning of items from the training set.

6.3 Method

6.3.1 Participants

27 adult participants aged 18-35 (14 females) were recruited from the online participant platform Prolific.[14] Participants were monolingual, native English speakers, with a Southern British English accent and no known language or learning disorders. Each participant received £6 for their involvement in the study. Three participants were excluded due to online recording technical difficulties. Therefore, data from 24 participants were included in our analysis. The study received approval from the Ethics Committee at Royal Holloway, University of London.

6.3.2 Stimuli and design

To assess whether participants learning to read aloud an artificial orthography follow the predictions of the TP, a novel artificial language consisting of 26 three-letter nonword items was

[14] Data collection was carried out online rather than in person due to the COVID-19 pandemic restrictions.

designed. Each item had a consonant-vowel-consonant (CVC) structure. In the orthography of this artificial language, the consonant graphemes were 11 familiar letters from the English alphabet (D, T, P, B, K, G, M, N, V, F, L, S, Z) which corresponded consistently to their regular English phonemes (/d/, /t/, /p/, /b/, /k/, /g/, /m/, /n/, /v/, /f/, /l/, /s/, /z/ respectively). The vowel graphemes were two "novel" letters: "δ" and "Ռ", the forms of which were borrowed from the Greek and Armenian alphabets respectively. In our artificial orthography, these two graphemes had inconsistent vowel pronunciations, with a one-to-many grapheme-phoneme mapping. The consistency of the pronunciations of the two vowel graphemes was manipulated to form two conditions: one condition in which the TP predicts that learners should form a productive rule for the pronunciation of the vowel grapheme in any context, and one condition in which the TP predicts that learners should not form a productive rule for the pronunciation of the vowel in any context. Instead, learners should apply the TP recursively, and form a productive rule for the pronunciation of this vowel grapheme only in the context of a specific final consonant, i.e. forming a productive rule for the pronunciation of a specific word body.

Both conditions consisted of 13 nonword items. Each vowel grapheme was used in only one condition, appearing in the medial position of all 13 items in that condition. We used the TP algorithm to calculate the number of exceptions a productive rule can tolerate for a set of 13 items; the predicted threshold of tolerated exceptions is 5 items. This allowed us to form one condition in which, according to the TP, the pronunciation of the vowel grapheme is sufficiently consistent to form a productive rule (passing the tolerance test), and one condition in which the pronunciation of the vowel grapheme is not sufficiently consistent to form a productive rule (failing the tolerance test).

In the Vowel Rule condition, the vowel grapheme was pronounced using the phoneme /i:/ in 10 items, /æ/ in two items, and /u:/ in one item. As the number of exceptions to the most common /i:/ pronunciation (three) falls below the tolerance threshold (five), the TP predicts that learners form a productive rule for this pronunciation of the vowel grapheme.

In the Body Rule condition, the vowel grapheme was pronounced /u:/ in six items, /ɛ/ in five items, /æ/ in one item and /ɪ/ in one item. Here, the number of exceptions to any of these pronunciations exceeds the tolerance threshold, meaning that the TP predicts that learners will not form a categorical productive rule for the pronunciation of this vowel grapheme. However, the TP predicts that this will trigger a recursive search for more specific rules within subsets of the items (see Chapter 2, *Section 2.5.4* for discussion of the recursive application of the TP). In this condition, six of the 13 items use “v” in word-final position, creating a six item word body subset. Of this six item subset, the vowel grapheme is pronounced /ɛ/ in five items and as /ɪ/ in one item. As the number of exceptions tolerated by a set of six items is three, the most common pronunciation /ɛ/ passes the tolerance test within this subset. Therefore, the TP predicts learners will form a productive pronunciation rule for the vowel grapheme in the context of word-final -v (i.e. a productive body rule).

Frequency of the word-initial consonant graphemes was balanced across conditions, such that each of the 13 consonant items was used in word-initial position once in each condition. In contrast, word-final consonants could be duplicated within conditions, such as word-final “-g” in the Vowel Rule condition. This ensured that word-final “-v” in the Body Rule condition subset was not the only duplicated final consonant in the exposure set. Use of the two vowel graphemes in each condition was rotated so that two mappings (A and B) were counterbalanced across participants (see Table 6.1).

Table 6.1

Counterbalanced mappings of the two novel vowel graphemes across two conditions

Condition	Mapping A	Mapping B
Vowel Rule	Ղ	δ
Body Rule	δ	Ղ

The full artificial language composed of two conditions and 26 nonword items, with pronunciations and orthographic representations using Mapping A, is presented in Table 6.2.

Table 6.2

The orthographic and phonological form of 26 items from the exposure set of the artificial language, using vowel Mapping A

Orthography	Phonology	Condition
TՐG	/ti:g/	Vowel Rule
ZՐG	/zi:g/	Vowel Rule
FՐG	/fi:g/	Vowel Rule
NՐG	/ni:g/	Vowel Rule
KՐG	/ki:g/	Vowel Rule
PՐB	/pi:b/	Vowel Rule
SՐF	/si:f/	Vowel Rule
LՐT	/li:t/	Vowel Rule
DՐK	/di:k/	Vowel Rule
VՐN	/vi:n/	Vowel Rule
GՐG	/gu:g/	Vowel Rule
BՐZ	/bæz/	Vowel Rule
MՐZ	/mæz/	Vowel Rule
SδV	/sev/	Body Rule
ZδV	/zev/	Body Rule
BδV	/bev/	Body Rule
GδV	/gev/	Body Rule
MδV	/mev/	Body Rule
NδV	/nɪv/	Body Rule
LδD	/lu:d/	Body Rule
VδN	/vu:n/	Body Rule
DδL	/du:l/	Body Rule
PδB	/pu:b/	Body Rule
FδG	/fu:g/	Body Rule
KδS	/ku:s/	Body Rule
TδS	/tæs/	Body Rule

Participants were exposed to this set of items from the artificial language throughout the training phase (discussed further in *Procedure* below). During training, the frequency of nonword items varied approximately along a Zipfian distribution (Zipf, 1949), according to which the frequency of a word is inversely proportional to its rank. The training phase consisted of 119 total presentations of the 26 unique nonword items; the most frequent item appearing 24 times and the least frequent items appearing 3 times each. Items were randomly assigned a position on the Zipfian distribution for each participant, meaning that items were encountered a different number of times by each participant. This avoided the application of a consistent but arbitrary assignment of items to frequencies for the entire language across all participants. The token frequencies of exposure to items for all participants during the training phase are available here: https://osf.io/3bhrx/?view_only=493351a77f0e4bee92977b81ad8e2526.

A set of 20 new, untrained nonword items was used during the Generalisation Task in the testing phase (discussed under *Section 6.3.3*). These items were also three-letter nonwords with a CVC structure, using the 13 consonants and 2 vowels of the artificial language. Each vowel symbol appeared in 10 items, thus corresponding to either the Vowel Rule or Body Rule conditions from the exposure set. The generalisation items used some duplicate bodies corresponding to those used in the exposure set: 5 out of 10 generalisation items corresponding to the Vowel Rule condition used word-final “-g”, and 5 out of 10 generalisation items corresponding to the Body Rule condition used word-final “-v”. This ensured that the distribution of repeated word bodies in the generalisation items was similar to that used in the exposure items, and further allowed us to assess generalisation behaviour in these specific consonantal contexts. Table 6.3 presents the full set of 20 generalisation items.

Table 6.3

Orthographic form of the generalisation stimuli and corresponding condition from the exposure set

Item (Orthography)	Condition
SՂG	Vowel Rule
MՂG	Vowel Rule
DՂG	Vowel Rule
BՂG	Vowel Rule
VՂG	Vowel Rule
KՂB	Vowel Rule
GՂD	Vowel Rule
NՂM	Vowel Rule
TՂP	Vowel Rule
LՂL	Vowel Rule
KδV	Body Rule
PδV	Body Rule
LδV	Body Rule
FδV	Body Rule
TδV	Body Rule
DδT	Body Rule
GδM	Body Rule
NδM	Body Rule
VδK	Body Rule
ZδN	Body Rule

6.3.3 Procedure

Participants were briefed on the nature of the task by being informed that they would be trained to read items from an artificial language using an artificial script. They were informed that some letters from the script would be familiar, English letters and others would be novel symbols for them to learn. They were informed that they would be trained to read this artificial script by carrying out reading aloud and spelling activities on their computer. Participants then completed speaker and microphone checks. The procedure was run online using Gorilla software.

The training phase began with an exposure to the set of 26 exposure items. Participants were presented with the written form of each item one at a time on their computer screen for a duration of 6 seconds, and also heard a pre-recorded pronunciation of the item commencing after 2 seconds of the visual presentation. Participants were asked to try to remember the pronunciation of the items. Each item was presented once and items were presented in a randomised order.

After completing the exposure to each of the 26 items, participants carried out a reading aloud task. During this task, the written form of each item was presented to the participant on the screen and the participant was asked to say aloud the pronunciation of each item. Each item was presented one at a time on the screen for a maximum of 9 seconds, or until the participant had made their response and selected "continue". Following the participant's response, which was audio recorded by the Gorilla software, the participant heard a pre-recorded correct pronunciation of the item. The 26 items were presented in a random order and their frequency followed a Zipfian distribution (as described in *6.3.2 Stimuli and design*).

Following the reading aloud task, participants carried out a spelling task. During this task, each participant was presented with a pre-recorded pronunciation of each exposure item and was asked to spell the written form of the item by using mouse clicks to select letters from a matrix on the screen. The matrix contained all letters from the artificial language (13 consonants and 2 vowels). There was no time limit for entering the response. Selected letters appeared at the top of the screen, and after three letters had been selected for the spelling of each item, the correct written form of the item was presented. Feedback on whether the selected letters were correct or not was also presented on the screen. The 26 exposure items were presented in a random order and their frequency followed a Zipfian distribution. This task concluded the training phase of the experiment.

The testing phase of the experiment immediately followed the training phase. This phase began with a Generalisation Reading Aloud task. During this task, the participant was presented with the written form of 20 novel, untrained items and asked to say aloud the pronunciation of each item. Each item was presented one at a time on the screen for a maximum of 9 seconds, or until the participant had made their response and pressed the spacebar in order to proceed to the next item. The items were presented in a random order. No feedback was given to participants during the task. This task tested the participants' ability to generalise their newly-acquired knowledge of the novel vowel letter pronunciations to untrained items, offering an opportunity to assess their rule-learning.

The testing phase concluded with a final Reading Aloud test of the 26 trained items. Participants were presented with the written form of each item and asked to read aloud the correct pronunciation. Items were presented on the screen one at a time for a maximum of 9 seconds, or until the participant had made their response and clicked "continue" in order to proceed to the next item. No feedback was given to participants during the task. This task assessed how accurately each participant had learned the pronunciation of each of the trained items.

6.4 Results

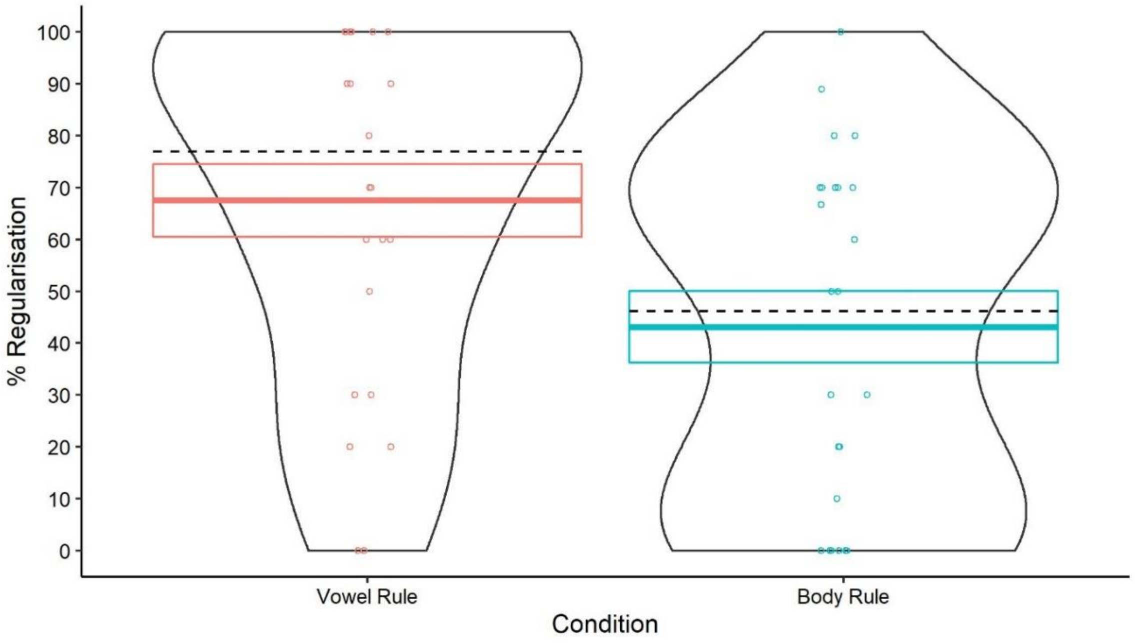
6.4.1 Generalisation

6.4.1.1 Use of the most common vowel pronunciation

In the Generalisation task, each participant read aloud 20 untrained items. Participants failed to provide a response in 0.4% of trials; these trials were excluded from the analysis. Figure 6.1 displays participants' percentage vowel regularisation (i.e. use of the most common vowel pronunciation from the training set) in their pronunciations of untrained generalisation items from the Vowel Rule and Body Rule conditions. Participants' mean vowel regularisation for all untrained items across both conditions was 55.32% (*SE* = 5.18). In the Vowel Rule condition, participants' use of the most common pronunciation (/i:/) was significantly less than 100% ($t(23) = -4.643$, $p < .001$), contrary to the first hypothesis of the TP and also the prediction of the DRC. In the Body Rule condition, participants' use of the most common pronunciation (/u:/) was less than 100% ($t(23) = -8.241$, $p < .001$), and also significantly lower than for generalisation items in the Vowel Rule condition ($t(23) = -3.204$, $p = .004$). These results support the TP's second hypothesis regarding use of this pronunciation, and contradicts the DRC's prediction that use of the most common vowel pronunciation should be high in both conditions.[15]

Figure 6.1

Participants' Percentage Vowel Regularisation in Untrained Items During the Generalisation Task



Note. Vowel regularisation is use of the most common pronunciation of the vowel grapheme from the training phase. Dashed lines represent type frequency of exposure to this phoneme during training.

Across conditions, these results indicate that participants are not producing the categorical behaviour predicted by either the TP or the DRC in this task. In fact, participants' average use of the most common pronunciation of the vowel graphemes during generalisation

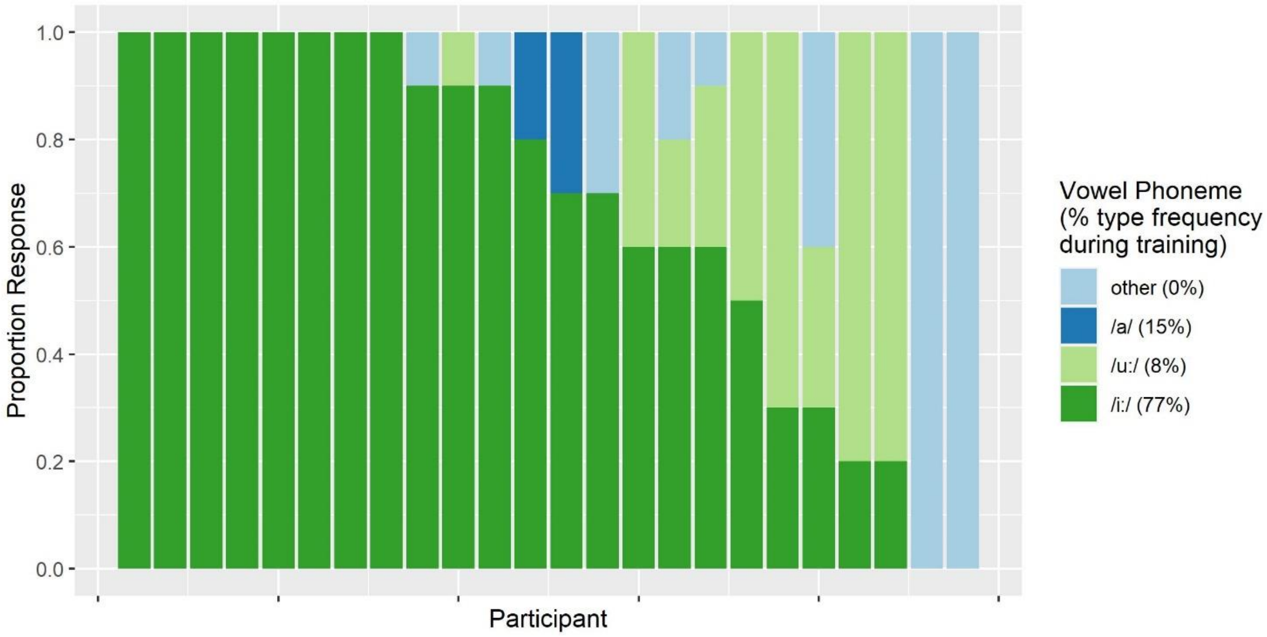
[15] An analysis comparing the effect of type and token frequency on generalisation was not included due to the high collinearity of these variables.

did not differ significantly from the type frequency of exposure to these pronunciations during the training phase, both for the grapheme in the Vowel Rule condition (*t*(23) = -1.343, *p* = 0.192) and in the Body Rule condition (*t*(23) = -0.442, *p* = 0.662).

However, Figure 6.2 presents the proportion of phonemes each participant used in their responses to generalisation items from the Vowel Rule condition. This presentation of the data suggests that some participants are in fact behaving more categorically than analysis of the group data would suggest, and are not matching the distribution of pronunciations during training according to type frequency. For example, 11 participants used the regular vowel pronunciation in at least 90% of generalisation items. Further analysis of pronunciations of items from the Body Rule condition is presented below.

Figure 6.2

Proportion of Pronunciation Responses to Items from the Vowel Rule Condition in the Generalisation Task Using Each Vowel Phoneme by 24 Participants



Note. Participants are ranked by proportion of responses using the regular /i:/ pronunciation (decreasing left to right). The percentage in parenthesis for each phoneme in the figure legend represents the type frequency of this pronunciation across trained items from the Vowel Rule condition.

6.4.1.2 Use of a body pronunciation sub-rule in the Body Rule condition

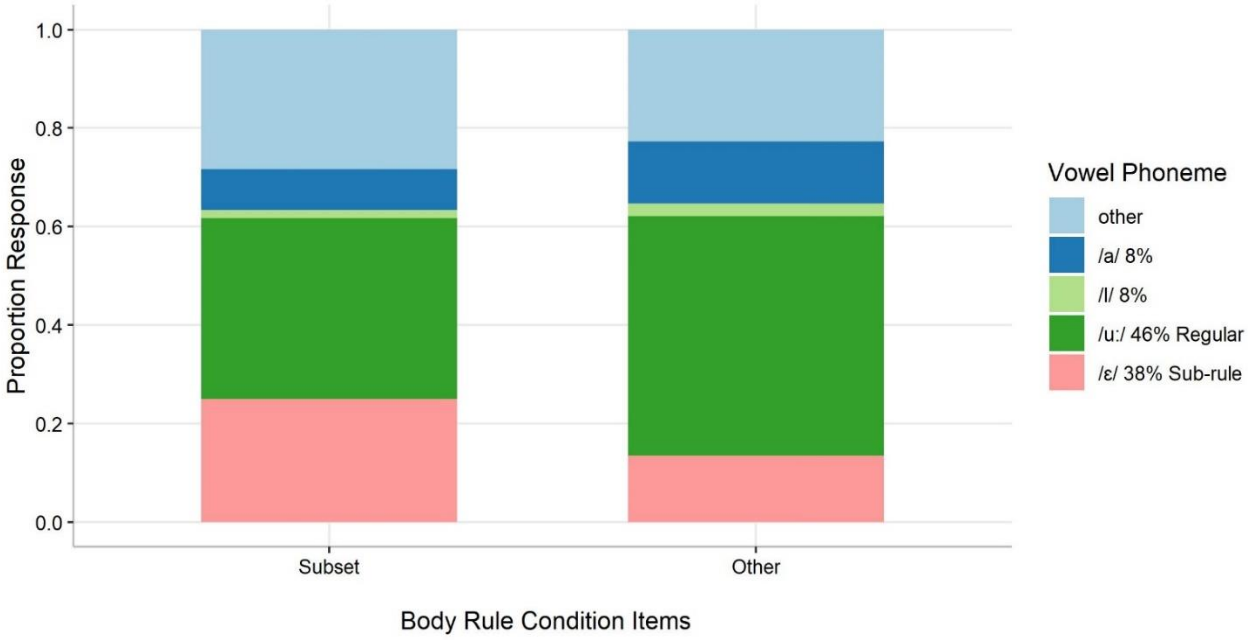
Five out of ten untrained generalisation items from the Body Rule condition used a word-final –v, forming a body subset. Figure 6.3 presents the proportion of participants' responses using each vowel phoneme in the body subset (five items with word-final –v) and other untrained items (five items with other word-final consonants) during the Generalisation task. Participants' use of the sub-rule vowel phoneme /ɛ/ to pronounce items from the body subset during this task was significantly less than 100% ($t(23) = -9.161$, $p < .001$), contrary to the

TP’s third hypothesis. Use of the most common pronunciation across all trained items from this condition (/u:/) for these generalisation subset items was also lower than 100% (*t*(23) = -8.250, *p* < .001), in contrast to the DRC prediction. Participants’ deviance from the TP prediction and from the DRC prediction for pronunciations of these items was not significantly different (*t*(23) = 0.813, *p* = .424). These results suggest that participants are not behaving in the categorical ways that either the TP or the DRC predict for these subset items. Instead, participants’ average use of the /ɛ/ phoneme for subset items did not differ from the type frequency of exposure to this phoneme across all items in this condition during training (i.e. the proportion of trained items which used this pronunciation) (*t*(23)= -1.637, *p* = 0.115). Similarly, use of the /u:/ phoneme for subset items did not differ from the type frequency of exposure to this phoneme across all items in this condition during training (*t*(23)= -1.242, *p* = 0.227).

Participants’ use of the sub-rule pronunciation /ɛ/ in other generalisation items (5 items without word-final –v) is not significantly different from use of this pronunciation in subset items (*t*(23) = 2.024, *p* = 0.055), contrary to the TP’s fourth hypothesis that this pronunciation should be reserved for subset items only. Participants’ use of the most common vowel pronunciation /u:/ for these items was less than 100% (*t*(23) = -7.092, *p* < .001) contrary to the DRC prediction, although this was not significantly different from use of this pronunciation for subset items (*t*(23) = -1.567, *p* = 0.131), as the DRC would predict.

Figure 6.3

Proportion of Responses Using Each Vowel Phoneme in The Subset and Items with Other Word-final Consonants from the Body Rule Condition in the Generalisation Task



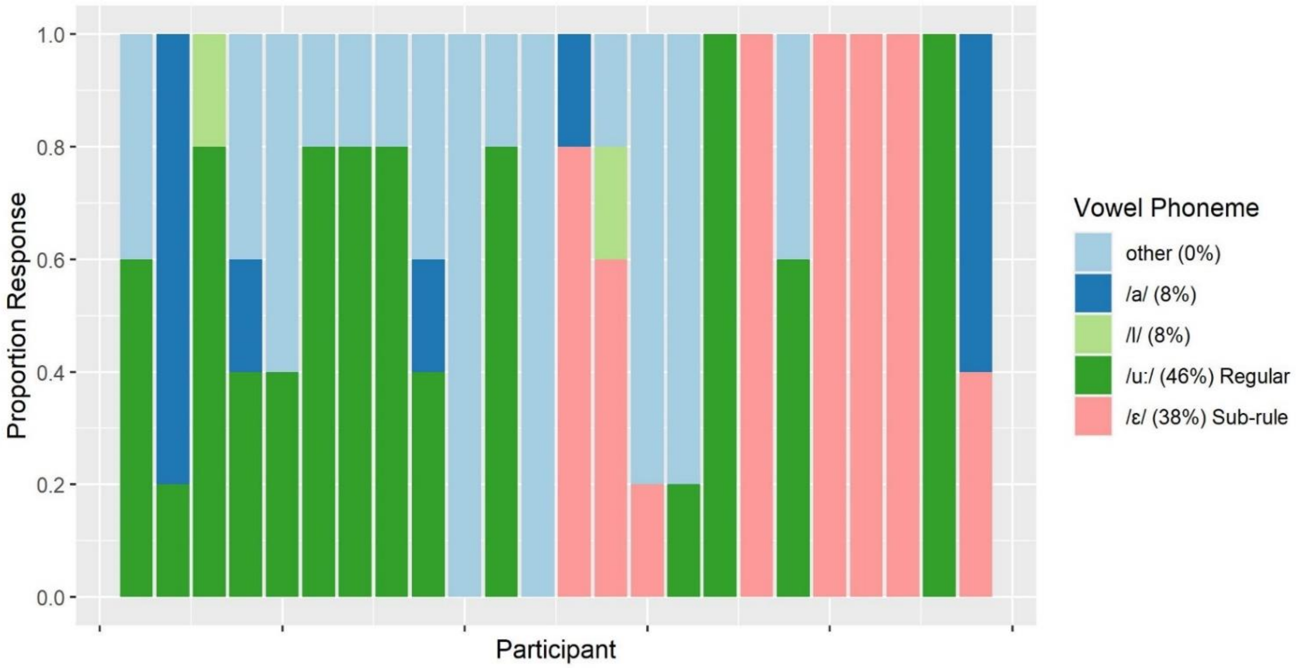
Note. Five items in each group (Subset and Other). The percentage in parenthesis for each phoneme in the figure legend represents the type frequency of this pronunciation across all trained items from the Body Rule condition. “Regular” denotes the most common pronunciation across training items by type; “Sub-rule” denotes the pronunciation predicted by the TP for subset generalisation items.

6.4.1.3 Individual differences in generalisation of subset items

The group-based analysis above seems to suggest that in pronunciations of generalisation items, including subset items, participants use a range of vowel phonemes at a rate close to the type frequency of these phonemes in the training set. Turning to investigate individual participant’s generalisation behaviour in the Body Rule subset, Figure 6.4 presents each participant’s use of the phonemes /a/, /ɪ/, /ɛ/, /u:/ (each of which was heard as a pronunciation of this vowel symbol in trained items from the Body Rule condition) and “other” in their pronunciations of these generalisation items. In this figure, participants are ranked by their number of accurate responses to trained subset items in the final reading aloud test (increasing accuracy from left to right). This presentation of the data reveals that, in fact, participants behave more categorically than the group-based analysis would suggest, with four participants using the sub-rule /ɛ/ pronunciation for all subset items, indicating formation of a categorical sub-rule as predicted by the TP. The accuracy ranking of participants indicates an association between accuracy of trained subset items and use of the sub-rule: these four participants scored in the top 30% of participants for accuracy of trained subset items. Further, every use of the sub-rule /ɛ/ pronunciation for subset items in the Generalisation task was made by participants scoring in the top 50% for accuracy on trained subset items.

Figure 6.4

Proportion of Pronunciation Responses to Subset Items in the Generalisation Task Using Each Vowel Phoneme by 24 Participants.



Note. Participants are ranked by number of accurate responses to trained subset items in the final reading aloud test (increasing in accuracy from left to right). The percentage in parenthesis for each phoneme in the figure legend represents the type frequency of this pronunciation across all trained items from the Body Rule condition (both subset and other items). "Regular" denotes the most common pronunciation across training by type; "Sub-rule" denotes the pronunciation predicted by the TP for subset items.

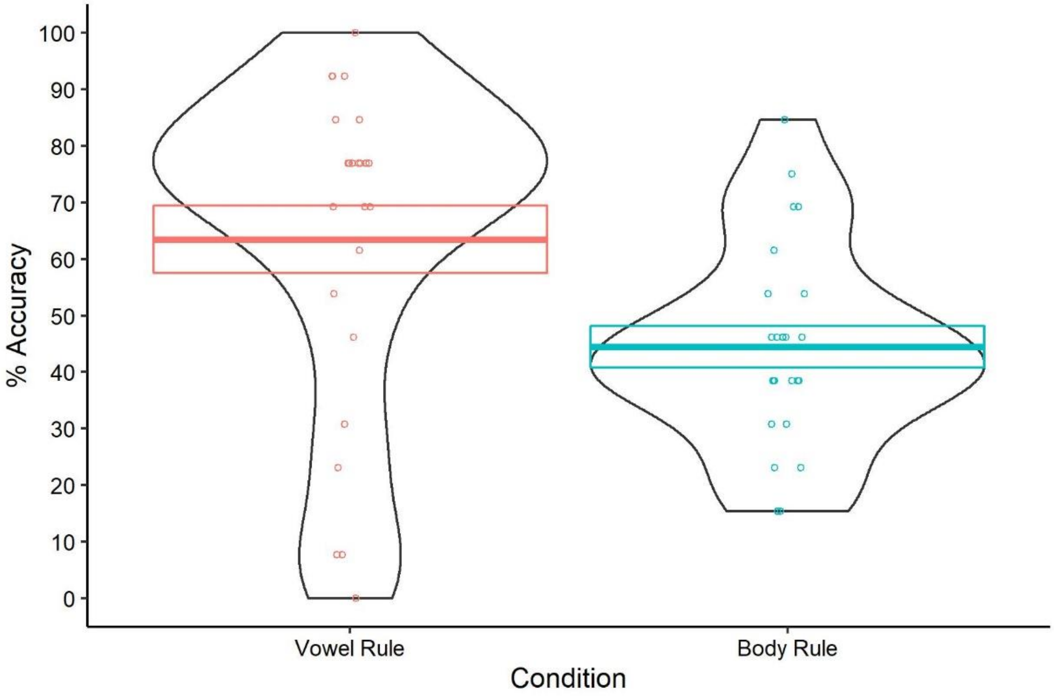
6.4.2 Accuracy of trained items

6.4.2.1 Performance in the final reading aloud task

Figure 6.5 presents participants' percentage accuracy in the final reading aloud task for the 26 trained items from the two conditions in the training set. Responses were scored as correct if pronunciation of the vowel symbol matched the trained pronunciation for the vowel in that item. Participants failed to provide a response in 0.4% of trials; these trials were excluded from the analysis. Participants' accuracy in the Body Rule condition was significantly lower than in the Vowel Rule condition ($t(23) = 3.264$, $p = .003$); the average overall accuracy across both conditions was 54.0% (SE = 3.99).

Figure 6.5

Participants' Accuracy (%) for Trained Items by Condition in the Final Reading Aloud Task

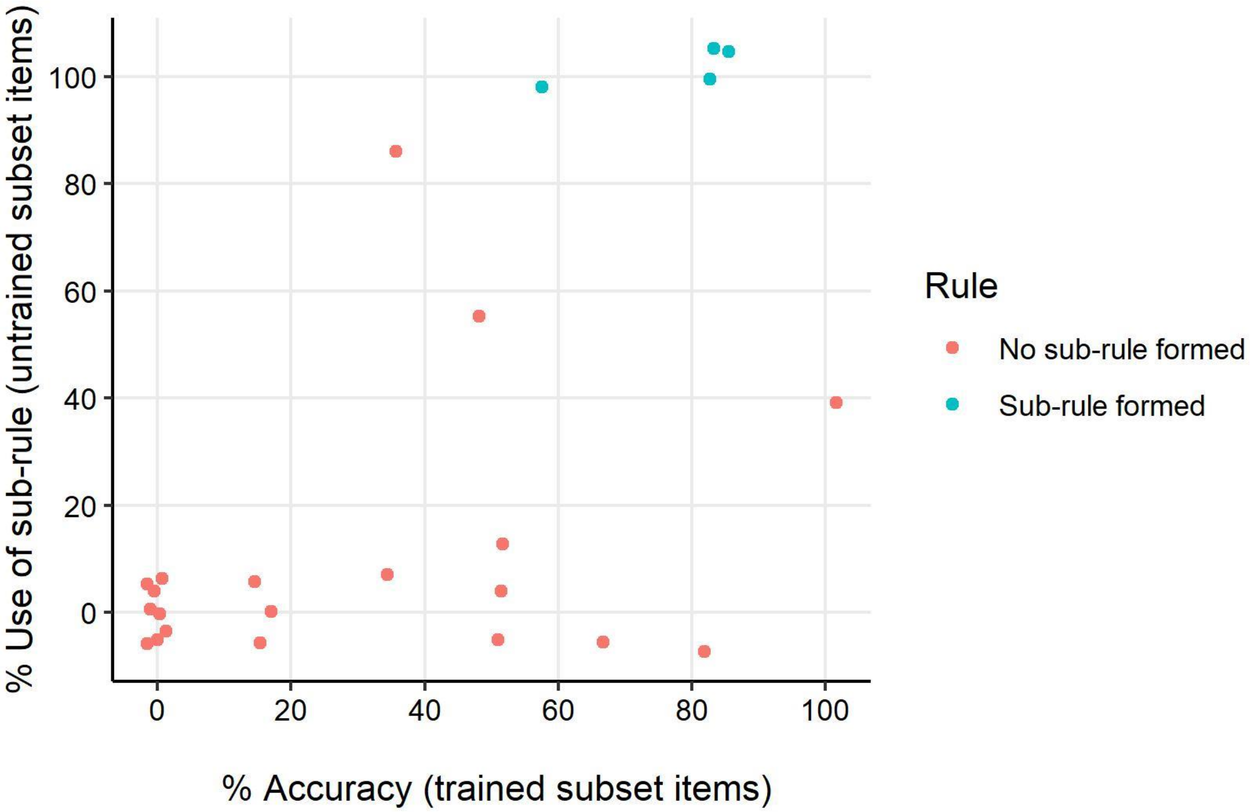


6.4.2.2 The relationship between accuracy of trained items and generalisation of untrained items in the body subset

Analysis of the generalisation task above revealed that four participants formed a categorical body subset rule which they used to pronounce 100% of generalisation items ending in word-final –v (the five-item body subset). Exploring what factors may support these participants' extraction and use of this sub-rule, I considered the relationship between accuracy of trained subset items and use of the sub-rule during generalisation. Figure 6.6 displays all participants' percentage use of the sub-rule for subset items in the Generalisation task as a function of their pronunciation accuracy on trained subset items during the final reading aloud task, with the four participants who formed a categorical sub-rule coded in blue. This presentation of the data suggests that the four participants who formed the sub-rule were successful learners with high levels of accuracy on these trained items. However, some other participants also had high levels of accuracy on these items but did not form a productive sub-rule, suggesting that even successful learning of subset items is not necessarily sufficient for forming a productive pronunciation sub-rule.

Figure 6.6

24 Participants’ Use (%) of the /ɛ/ Sub-rule Pronunciation for Subset Items in the Generalisation Task by Accuracy (%) on Trained Subset Items in the Final Reading Aloud Task



Note. Participants are coded according to whether they used the /ɛ/ sub-rule pronunciation in 100% of subset item responses (blue) or less than 100% of items (red).

6.5 Discussion

Yang's *Maximise Productivity* principle states that learners should pursue a search for productive rules in their input; when the consistency of a rule crosses the tolerance threshold, the TP should apply recursively to find productive regularities within smaller subsets. This, he argues, is necessary in order for children to acquire the nested structures of grammatical patterns in natural language. Some experimental work has found that adults too are able to acquire knowledge of nested structures in artificial grammar learning studies (Reeder et al., 2017; Udden et al., 2009). Meanwhile, a body of research has investigated how readers of English use information from their text experience to develop sensitivity to context-specific pronunciation patterns. This knowledge goes beyond the most common but often inconsistent mappings between graphemes and phonemes, as demonstrated by their pronunciations of nonwords (Glushko, 1979; Ryder & Pearson, 1980; Treiman et al. 1990; Coltheart & Leahy, 1992; Andrews & Scarratt, 1998). Some researchers have highlighted the role of statistical learning mechanisms in readers' ability to acquire this complex knowledge without explicit instruction (Arcuili & Simpson, 2013; Steacy et al., 2019), although it has also been noted that readers do not simply reproduce the distribution of different pronunciations that occur in English words (Treiman et al., 2003; Steacy et al., 2019; Treiman & Kessler, 2019).

In Experiment 4, I explored whether the TP could account for the way in which readers use certain pronunciation patterns more or less often than one might expect from the statistical distribution of the input, and specifically readers' ability to extract and use context-specific pronunciation patterns without explicit instruction. The TP offers a promising approach to these questions as it is a rule-based theory which makes categorical predictions on the basis of statistical information. Using the recursive application of the algorithm, Experiment 4 assessed whether the TP can predict learners' acquisition and generalisation of both a context-free "vowel rule" and a context-specific "body rule" to pronounce novel symbols in an artificial orthography. As context-specific spelling-sound mappings can be characterised as "sub-rules" applying to a subset of items in the input, the experiment investigated whether a highly inconsistent, context-free pronunciation pattern would trigger the search for a more consistent context-specific pronunciation rule that could be used productively, as the TP would predict. Specifically, I hypothesised that learners would form a context-specific pronunciation rule for a vowel symbol in a subset of items when the pronunciation of this symbol across all items did not pass the tolerance test.

In this artificial orthography, two novel vowel symbols were used in nonword items from two conditions. In the Vowel Rule condition, the pronunciation of the novel vowel symbol passed the tolerance test, so the TP predicts that a context-free pronunciation rule for this symbol should be formed. In the Body Rule condition, the pronunciation of the novel vowel symbol did not pass the tolerance test so a context-free pronunciation rule is not predicted. However, pronunciation of the vowel symbol in a subset of items ending in –v did pass the tolerance test, meaning that the TP predicts a productive body rule for pronunciation of the vowel symbol in this body context. Adult learners were trained to read aloud 26 items (13 in each condition) and were then assessed on their generalisation and learning of spelling-sound correspondences during the testing phase.

6.5.1 Generalisation to untrained items

6.5.1.1 Use of the most common pronunciation of the vowel symbol

The Generalisation task assessed whether adults had formed pronunciation rules for the two vowel symbols through analysis of their pronunciations of these symbols in 20 untrained items. Contrary to the categorical predictions of both the TP and DRC, participants' use of the most common pronunciation of the vowel symbol during training counted by type (i.e., vowel regularisation) in generalisation items from the Vowel Rule condition was significantly lower than 100%. However, use of the most common pronunciation of the vowel symbol in the Body Rule condition was significantly lower than the Vowel Rule condition; this is predicted by the TP as this pronunciation does not pass the tolerance test and thus should not be used productively, but does not support the DRC which would predict regularisation to be high in both conditions. These results suggest that participants are more likely to regularise the pronunciation of the vowel when the consistency of this pronunciation passes the tolerance test, which is in accordance with results from Experiment 2 and Schuler (2017) where regularisation was also found to be higher in conditions which pass the tolerance test.

Group-level analysis suggested that participants may be matching the type frequency of the most common pronunciation from training in their generalisation in both conditions. However, examination of the distribution of pronunciation responses by individual participants in the Vowel Rule condition suggests that generalisation may be more categorical than group-level analysis would suggest; for instance, 11 participants used the regular vowel pronunciation in at least 90% of their generalisations. Firstly, this hints at the importance of considering individuals' behaviour that may be obscured by group-level patterns when assessing participants' generalisation. Further, it suggests that individual learners' generalisation behaviour may be more categorical than the statistical distribution of forms they were exposed to. For example, some participants regularised the most common pronunciation rather than producing the range that they heard during training, in line with the TP and a similar pattern of results from Experiment 2. Similarly, Treiman and Kessler (2019) found that readers are not ideal statistical learners who precisely match the distributions of their input.

6.5.1.2 Use of a body pronunciation sub-rule

In the Generalisation task, five untrained items from the Body Rule condition used a word-final –v, forming a body subset. The TP predicts that learners should form a body sub-rule for the pronunciation of the vowel symbol in this consonantal context, rather than use the most common pronunciation of the vowel across all items from this condition as the DRC would predict. However, group-level analysis of participants' use of both the subset pronunciation and the regular pronunciation for these items was significantly lower than 100%, suggesting that they follow the categorical predictions of neither the TP nor the DRC. Further, participants' use of both pronunciations during generalisation did not differ significantly from their type frequency during training, and use of the subset pronunciation in non-subset items did not differ from that in subset items; a power analysis would improve confidence that these critical null findings are not due to an underpowered study. Although the use of a range of pronunciations is consistent with earlier findings that readers do not simply generalise the most common grapheme-phoneme correspondences in their input (e.g. Glushko 1979; Andrews & Scarratt 1998), these results do not offer evidence that participants formed a context-specific sub-rule as predicted by the TP.

Instead, they suggest that learners match the type frequency of pronunciation distributions in the input.

However, as in the Vowel Rule condition, examination of individual participants' generalisation for subset items revealed that participants were again behaving more categorically than the group-level analysis would suggest. Individual participants largely did not use a range of pronunciations matching the distribution of the trained items. However, they also do not all behave in the same way. For instance, four participants used the context-sensitive pronunciation to pronounce all untrained subset items, suggesting that they did form the predicted sub-rule. This behaviour matches the predictions of the TP, and is in line with the suggestion by Steacy et al. (2019) that the pronunciation of a vowel grapheme with possible alternative pronunciations is determined by a trade-off between the strength of context-free and context-dependent pronunciations. Additionally, these participants all scored in the top 30% when participants were ranked in order of accuracy on reading aloud trained subset items. Further discussion of the relationship between productive pronunciation rules and successful learning of trained items is detailed below.

6.5.2 Accuracy of trained items

In the final Reading Aloud test, participants were presented with the 26 original trained items and were asked to pronounce each one individually. This task allowed assessment of whether participants had successfully learned the pronunciations of the exposure items they had encountered during the training phase. Accuracy in the Vowel Rule condition was significantly higher than the Body Rule condition, suggesting that spelling-sound mappings with higher inconsistency are more difficult to learn, even if there are embedded sub-regularities. This result is in line with similar findings by Taylor et al. (2011): adult learners in an artificial orthography learning paradigm demonstrated highest accuracy for consistent vowel characters, followed by inconsistent-conditioned characters which had contextual sub-regularities, and lowest accuracy for inconsistent-unconditioned characters without such sub-regularities.

6.5.3 Relationship between pronunciation rules and accuracy of trained items

As outlined above, only four participants formed the predicted pronunciation sub-rule, but those that did had relatively high accuracy on trained subset items compared to other participants, and scored at least 50% accuracy on these items in the final Reading Aloud task. This suggests that using a productive sub-rule for untrained items is specifically associated with knowledge of trained items that use the sub-rule, highlighting that producing the predicted pattern of generalisation is supported by successful learning of relevant items. At a general level, this finding is consistent with results from Experiment 3, which demonstrated the importance of taking into account an individual learners' knowledge of the forms they have acquired in order to understand their patterns of generalisation, rather than simply considering the forms they have been exposed to.

Regarding the acquisition and generalisation of context-sensitive spelling-sound mappings more specifically, these results are also consistent with findings from Treiman and Kessler (2019). Firstly, most participants do not make extensive use of the contextual information that is available in the input; less than one might expect given the distribution or consistency of the patterns they have been exposed to. Critically, Treiman and Kessler found that participants who did use the contextual information in nonword pronunciations were more skilled readers, offering evidence that use of contextual information increases throughout development of reading ability. This is akin to our finding that participants who were able to use contextual information to form a sub-rule had relatively high levels of accuracy on the trained subset items. As noted by Treiman and Kessler, productive use of more complex contextual information in orthography-phonology mappings is likely to be difficult, and may only be possible when this knowledge has been secured through extensive experience or explicit instruction. However, knowledge of the subset items may be a necessary but not sufficient condition for use of a sub-rule, as some other participants in the current experiment learned the subset items well but did not use the sub-rule productively. It is therefore possible that explicit knowledge and instruction is required to enable some readers to use more complex spelling-sound mappings.

In summary, during generalisation in Experiment 4, individual participants did not use the range of pronunciations matching the statistical distributions they were exposed to, as the group-level behaviour would suggest. Instead, they seem to behave more categorically, although they did not all do so in the same way. Most participants did not provide evidence that they had applied the tolerance threshold recursively to form a context-sensitive pronunciation sub-rule as predicted by the TP, but those that did had acquired some knowledge of the trained subset items. This finding highlights the importance of taking into account which items a learner has successfully acquired, which they may then use as the basis for generalisation.

Given that overall accuracy for all trained items was 54.0%, it is possible that more participants would have developed the sub-rule given further training that increased their accuracy on trained items. This would be consistent with findings from Treiman and Kessler (2019), who found that use of contextual information in nonword pronunciations written in English orthography increased gradually with reading skill. It also runs parallel to earlier work on reading development which also demonstrated that knowledge of more complex spelling-sound mappings increases with reading ability, potentially through text experience (Treiman et al., 1990; Laxon et al., 1991; Coltheart & Leahy, 1992). The current findings extend this research by suggesting that adult learners in an artificial orthography paradigm may follow a similar trajectory to younger readers of English as they develop the ability to use context-sensitive information in their pronunciations. In particular, secure knowledge of items that use these complex mappings may be required in order to use these patterns productively.

Chapter 7: Discussion

7.1 Summary of the thesis

This thesis set out to investigate whether a newly-proposed theory of rule-learning in language acquisition, the Tolerance Principle (Yang, 2016), could account for the acquisition and generalisation of spelling-sound correspondences in reading. Specifically, it examined how the presence of exceptions affects the acquisition and use of productive spelling-sound patterns by adults and children. It presented a series of nonword reading aloud and artificial orthography learning studies which assessed the TP's predictions for generalisation of such inconsistent spelling-sound mappings in unfamiliar word items. This research provides a novel contribution to the literature on word reading, as the TP offers a rule-based approach which incorporates statistical information, thus offering a middle ground between previous rule-based and statistical accounts. In doing so, it addresses long-standing questions in this field of research regarding the nature of orthographic-phonological knowledge and readers' use of different orthographic grain sizes. Additionally, it contributes to the wider literature on reading development and statistical learning by demonstrating what information from the input is extracted and used productively by learners.

In Chapter 3, Experiment 1 presented an initial, exploratory investigation of whether the TP could be applied to reading in an empirical setting. Specifically, it addressed whether the TP could predict the generalisation of spelling-sound correspondences from the English writing system in the pronunciations of nonword items by adults and children aged 8-9. The results found that the TP could predict participants' use of vowel- and body-level correspondences in nonword pronunciations relatively well, although participants' reading behaviour was less categorical than predicted. Additionally, results showed that the TP could predict participants' pronunciations of vowel graphemes in these nonword items more successfully than could three computational models of word reading. Further analysis suggested that the role of consistency in the TP theory underlies the TP's relative success in predicting nonword pronunciations within a quasi-regular writing system. Notably, the TP offers a type- frequency-based, categorical metric of consistency which can be applied recursively to multiple grain sizes.

In Chapter 4, Experiment 2 examined whether the TP could predict adults' and children's (aged 9-10) generalisation of novel, inconsistent spelling-sound correspondences in an artificial orthography learning paradigm. The TP had an effect on adults' and children's vowel regularisation beyond that of token frequency. This result suggests that during generalisation, learners are able to impose some additional structure on the input statistics in a way that is predicted by the TP. Overall, the effect of the TP on children's regularisation was greater than for adults. Although participants' generalisation was not completely categorical, this experiment provides evidence that the TP underlies adults' and children's generalisation during reading in some way.

In Chapter 5, I investigated whether token frequency can moderate the effect of the TP on the pattern of adult participants' generalisation observed in Experiment 2. Experiment 3 manipulated the token frequency of items during training, such that irregular items were highly frequent in the input. Initial analysis suggested that under these conditions, the TP no longer had an effect on adults' vowel regularisation. However, when individual participants' acquired vocabulary was taken into account, the TP did have an effect on regularisation, whilst token frequency did not. These results support the TP, and further suggest that although token frequency does not affect regularisation directly, it has an important role to play: highly frequent items may be acquired most successfully, and subsequently form the basis of generalisation. This finding highlights the importance of considering an individual's pattern of acquisition in order to understand their generalisation behaviour, offering an important contribution to the statistical learning literature.

Finally, the aim of Chapter 6 was to explore the recursive application of the TP. To do so, Experiment 4 asked whether the TP could predict adults' generalisation of novel, context-sensitive orthography-phonology correspondences in an artificial orthography. According to the TP, learners should apply the tolerance threshold recursively to find more specific sub-rules when a general rule does not pass the tolerance test. The results suggested that only four out of twenty-four participants demonstrated evidence of forming a categorical sub-rule for the symbol in a particular word body as predicted by the TP. Importantly, those that did form a context-sensitive pronunciation rule had acquired some knowledge of the relevant trained subset items. These findings support earlier research indicating that learning complex spelling-sound mappings is difficult, and they are consistent with suggestions that this process progresses throughout development. The results also highlight the importance of considering individual learners' trajectory of acquisition and generalisation, rather than relying on group-level data.

This brief summary offers an overview of the intentions and findings of the four experiments conducted here. Together, these experiments addressed a number overarching research aims, including: assessing the applicability of the TP to word reading; increasing our understanding of skilled and developing readers' orthography-phonology knowledge; and identifying which statistical distributions of the input are important for acquisition and generalisation. In the following section, a comprehensive evaluation of the TP in the context of word reading will be presented. This discussion will be followed by details of specific contributions from the current findings to the study of word reading, reading development and instruction, and statistical learning.

7.2 Evaluation of the Tolerance Principle in the context of word reading

The primary motivation for this thesis was to explore whether a theory of spoken language acquisition, the Tolerance Principle (Yang, 2016), applies more generally to other quasi-regular domains such as spelling-sound mappings in reading. To do so, I examined whether the TP could predict readers' generalisation of inconsistent spelling-sound correspondences. Whilst parallels between the quasi-regularity of morphological systems in spoken language and alphabetic systems in written language make this an area ripe for investigation, there are important factors to consider when applying the TP to a different modality from the one for which it was developed.

7.2.1 The issue of serial search

One previously-raised criticism of the TP involves the centrality of the serial search mechanism to the theory (see *Section 2.5.1* for an initial description of this mechanism). For instance, Kapatinkski (2018) suggests that serial search is not compatible with the current understanding in psychology regarding distributed representations and parallel processing;

Wittenberg and Jackendoff (2018) similarly highlight the implausibility of using this mechanism during lexical access. Whilst Yang (2018) refutes these arguments in the context of morphological processing ("is a parallel brain really incompatible with serial behavioural effects?" (2018, p.797)), the following discussion lays out how they might not be so easily dismissed in a discussion of word reading.

Serial search models have been developed previously in the study of visual word recognition (e.g. Forster, 1976; Murray & Forster, 2004). In Forster's original (1976) model, lexical access is conceived as a two-stage process involving the serial search of a single access file, which subsequently provides a point of access to an entry in the lexicon. Specifically, there are three alternative types of access file - orthographic, phonological and semantic - with information about a written word reached using the orthographic access file. The access file is organised into "bins", each containing orthographic representations of a subset of the lexicon. The input word is mapped to a specific bin using a hash-code, which provides an abstract representation of the word's features. Within the bin is a list of similar words that all share a hash-code with the input word, ranked by their frequency. A search procedure then commences through the list of words in the bin, setting out to find a match between the input word and a bin entry. After a match is found, a corresponding entry in the master file (i.e., in the lexicon) can be located and accessed, providing full lexical detail about the input word. This model accounts neatly for frequency effects found in lexical decision tasks (Yelland, 1994), whereby reaction times to more frequent words are shorter than for less frequent words (e.g. Rubenstein et al., 1970), as a match will be located faster for a more frequent word nearer the top of the list. Indeed, it has been argued that rank-order offers a better fit to lexical access data than log frequency (Murray & Forster, 2004), in line with the specific predictions of a serial search model.

Yang (2016) notes that this model accommodates the lack of age effects found in skilled lexical processing, whereby older adults do not display faster reaction times than younger adults in word-naming tasks (e.g. Cerella & Fozard, 1984).[16] Yang concludes that the model's use of rank frequency can account for that fact that cumulative experience with words over time does

[16] In fact, Balota and Duchek (1988) found that older adults were *slower* than younger adults to initiate their productions during pronunciation tasks.

not increase processing speeds during adulthood, because the relative rank of words stays constant even though cumulative frequency continues to grow.[17]

However, serial search models have faced several points of criticism. For instance, they cannot account for error responses being made faster than correct responses (Ratcliff et al., 2004). Neither can they explain instances in which responses are faster for nonwords than real words, as according to the model, nonword responses are produced when no match is found in the corresponding bin. Findings that neighbourhood size affects latencies for nonwords but not real words in lexical decision tasks (e.g. Coltheart et al., 1977) pose a similar problem. Further, Adelman and Brown (2008) question Murray and Forster's (2004) analysis of lexical decision data which claimed to demonstrate that rank order was a better prediction of decision times than logarithmic frequency transformations. Instead, Adelman and Brown present analysis suggesting that the picture is in fact more complicated; that there is variability in lexical decision times which is systematically related to frequency but is not accounted for by a linear function of rank frequency. They propose the factor of contextual diversity, which is confounded with word frequency, could be a plausible explanation for the observed frequency effect.

Critically for the current discussion, most of the assessments of the serial search model have involved lexical decision tasks and masked-unmasked priming, rather than production tasks such as nonword reading aloud tasks which involve generalisation. Indeed, it is not clear in Forster's serial search model how orthography-phonology knowledge (which is stored only at the lexical level in the model) would be generalised in order to produce pronunciations for unfamiliar written items as I have investigated in this thesis.

Setting aside previous serial search models to focus on Yang's (2016) theory, one thing is clear: productivity is the TP's *raison d'être* and the serial search mechanism is its cornerstone; if the TP is to offer a successful account of generalisation then in some way the two must be reconciled. Certainly in its current instantiation, the TP is bound to the serial search mechanism

[17] Zevin and Seidenberg (2004) found an effect of cumulative frequency on naming latency in skilled readers, with high frequency words named faster than low frequency words. However, cumulative frequency was treated as a two-level factor (early vs. late), and an effect of rank frequency was not investigated. Further, the cumulative frequency values were log frequencies from Zeno (1995), summed across all grades from kindergarten to college. Additionally, participants were all undergraduate students. Therefore, we cannot draw conclusions from this study about whether reading speed over the lifespan continues to be affected by cumulative experience, as participants were likely all young adults, and frequency values were summed over years in education rather than measured across a lifetime of reading experience.

because it uses the time of lexical access using this procedure to assess the growing cost of exceptions to a productive rule, and thus to determine the tolerance threshold. Unlike Forster's model in which *all* orthographic, phonological and semantic knowledge is lexically listed and serially searched, Yang's theory uses this mechanism for only part of the access procedure. Specifically, Yang implements the Elsewhere Condition using the serial search mechanism, according to which exceptions to a rule are stored in a frequency-ranked lexical list, and searched for a match with the target. If a match is not found, then the productive rule is applied. When the number of exceptions to a productive rule exceeds the tolerance threshold, *all* items are lexically listed and will be subject to the serial search procedure. The TP states that a learner will use the route that is the fastest to carry out during online processing (either exceptions-plus-rule or everything-listed).

The reason why the serial search mechanism in particular plays a vital role in the derivation of the TP is that it is used to approximate the time of lexical access according to each route, and thus provide the point at which they are equal (i.e., the tolerance threshold). Yang describes the expected time of rule access to be $T(N,e)$ and the time of access when all items are listed to be $T\ (N,N)$; according to the TP, a rule is productive if $T(N,e) < T\ (N,N)$ (Yang, 2016 p. 61). The closed form solution to the equation $T(N,e) = T\ (N,N)$ to find e (the number of exceptions) involves a calculation of the probability of occurrence of a target item in a Zipfian frequency distribution (using the Nth harmonic number). This calculation is used to approximate the time it would take to access the target using a serial search of a frequency-ranked list within each route: either for all items in $T\ (N,N)$, or for only exceptions before rule application in $T(N,e)$ (2016, p. 63 – 64; see also Appendix A). Without the serial search mechanism, the threshold that is derived (i.e. e¸ when $T(N,e) = T\ (N,N)$) becomes meaningless.

It is in fact the Elsewhere Condition that protects the serial search mechanism from immediately negating use of the TP for reading. As the experiments in this thesis have demonstrated, a successful theory of reading must be able to capture *generalisable knowledge*; specifically, it must include an abstract representation of spelling-sound correspondences that can be applied to read aloud unfamiliar words. A theory which used a serial search of lexical items alone would not provide this, as highlighted above. However, under the Elsewhere Condition there is an abstract productive rule system that operates separately (and subsequently) to the serial search of lexical items. Therefore, not all knowledge is stored lexically; there are also generalisable rules, which we have seen to be required for reading unfamiliar lexical items. The rules themselves do not involve serial search, and therefore offer a process whereby readers can read aloud unknown items by using generalisable knowledge. Beyond this, the TP offers a novel solution to the question of how a learner decides which patterns should be used productively, and which should not be generalised beyond items attested in the input, but instead stored in a list. In doing so, the TP can incorporate both distributed and lexical information, offering a method of justification according to which orthography-phonology knowledge is deemed valuable enough to achieve a distributed, abstract structure. Namely, once a spelling-sound correspondence attested in the input is sufficiently consistent to pass the tolerance test, this information can be stored as a productive rule involving individual orthographic units such as graphemes and bodies, rather than being stored at the lexical level.

The use of a serial search mechanism under the Elsewhere Condition makes specific predictions about frequency effects: not only should rank frequency predict processing speed more accurately than absolute frequency, but all else being equal, irregular forms should be accessed faster than regulars (as they are searched before the rule is applied). Yang reports an analysis of reaction times in a lexical decision task from the English Lexicon Project (Balota et al., 2007) which is in line with this prediction; rank frequency was a “slightly” better fit for reaction time data of irregular past tense verbs than the logarithm of lexical frequency (2016 p. 51). Additionally, Yang offers a detailed discussion of cross-linguistic evidence demonstrating why the “puzzlingly cumbersome” Elsewhere Condition is a “fundamental principle of linguistic organisation where specificity and generality come into conflict” (2016 p. 52 – 60).

This corollary that exceptions are accessed before the application of a rule, presents a potential hurdle for the application of the TP to reading, as it would seemingly predict that irregular words are read faster than exception words. Such a prediction would be problematic as the regularity effect on word naming (and its interaction with word frequency) is well established (e.g. Seidenberg et al., 1984; Paap & Noel, 1991; Coltheart & Rastle, 1994). As such, words pronounced using regular GPCs are in fact *faster* to read aloud than irregular words, although this effect is stronger for low frequency than high frequency words.

However, this fact about reading behaviour is not necessarily an issue in the current conception of the TP for reading aloud, as the TP does *not* necessarily predict that irregular words will always be read faster than regular words. The reason for this is that the TP mechanism described here applies to individual orthographic units rather than to whole words. Specifically, the process to pronounce a written item using the TP (whether it be a regular word, an irregular word, or a nonword) takes place using a serial, left-to-right process starting with the smallest individual unit[18]. For example, when presented with the word "bull" to read aloud, the first step is *not* to search through a frequency-ranked list of all exception words for an entry corresponding to the pronunciation of that entire word. Instead, the first step is to search through a list of words in which "b" does not follow the regular pronunciation rule (such as *climb, tomb* etc). As the target is not amongst the list, and "b" has a pronunciation that passes the tolerance test, the productive rule "b" -> */b/* can be applied. The next step is to produce a pronunciation of the grapheme "u", and so on (see *Section 7.3.2* below for a full exemplification of the process to produce a pronunciation for an entire lexical item). Crucially, assembling the pronunciation for a word in this way involves identifying productive rules for individual orthographic units in turn. Therefore, it is not irregular words that will be accessed before regular words, but irregular pronunciations of orthographic units before regular pronunciations of orthographic units. The overall naming latency of a word will thereby depend on the regularity and frequency of its constituent orthographic units, not whether it is a regular or exception word *per se*.

7.2.2 Using the TP mechanism for word reading

Whilst there may be theoretical motivation for applying the TP to reading, how feasible is the use of the TP mechanism (including serial search, the Elsewhere Condition, and the recursive rule structure) for online processing of orthographic knowledge during skilled reading? To exemplify this process, consider the steps that would be undertaken by a reader to pronounce the nonword *boup*:

[18] This reasoning follows Yang's *Maximise Productivity* principle, which states that learners pursue rules that maximise productivity (2016, p. 72). Therefore, the most general rule that passes the tolerance test - using the smallest orthographic unit available – will be prioritised.

1) Start with the initial grapheme "*b*", which follows the exception-plus rule route as the consistency of the *b -> /b/* rule in English words passes the tolerance test.

2) Search through a frequency-ranked list of exceptions in which "*b*" does not follow the productive rule *b -> /b/,* such as *climb, tomb,* etc.

3) No match for the target word amongst the list of exceptions is found, so the productive rule *b* -> /b/ (which is stored in the reader's orthography-phonology rule system) can be applied.

4) Move next to the vowel grapheme "*ou*". This grapheme does not have a productive rule stored in the readers' orthography-phonology knowledge system, as no pronunciation of this grapheme in English words is consistent enough to pass the tolerance test. Therefore, the reader does not follow an exception-plus-rule route for this grapheme. However, the grapheme forms part of a number of more specific pronunciation rules in the reader's rule-system, triggering the reader to consider the consonantal context surrounding the vowel grapheme in the target item, in this instance the word body "*oup*".

5) The word body "*oup*" is associated with a productive pronunciation rule in the reader's orthography-phonology rule system, so the reader can follow the exception-plus-rule route for this body.

6) Search through a frequency-ranked list of exceptions that do not follow the "*oup*" -> /u:p/ rule, such as *coup.*

7) No match for the target is found amongst the list of exceptions, so the "*oup*" -> /u:p/ rule can be applied.

8) Finally, the pronunciation *boup* -> /b / + /u:p/ ->/bu:p/ can be assembled.

It would be reasonable to argue that this multi-step process involving lexical listing, the serial search procedure, and a hierarchical structure of rules, is simply not a feasible account of online processing for skilled readers. Nevertheless, results from Experiment 1 suggested that the TP was a better predictor of both adult and child participants' nonword pronunciations than the three computational models of word reading also assessed. I argued that the reason why the TP performed more successfully than the extant rule-based or statistical approaches was that it uses a type frequency-based categorical threshold of consistency that applies recursively to capture orthography-phonology correspondences at different grain sizes. Even if we are to discount the TP in its current instantiation as a realistic online processing mechanism for skilled word reading, we need not dismiss everything that has been gleaned from this investigation at once.

In its original instantiation, the TP mechanism is used both to acquire productive rules, and to apply them during online skilled processing. During the acquisition process, learners encounter items in their input which are learned individually. Throughout cumulative experience, productive rules will emerge (or, indeed, disappear) over time, depending on the items in the linguistic evidence a learner has received. Rules are formed when the number of items which are exceptions to a majority pattern falls below the tolerance threshold; otherwise, all items are stored lexically in a frequency-ranked list. The balance between the number of regulars and exceptions is updated as a learner's experience grows. The underlying motivation for this process is to identify the most computationally efficient (i.e., quickest) route of access during online processing, as this same system of rules and frequency-ranked exceptions developed during acquisition is later employed by the skilled language user in their real-time productions. In this way, the TP sets out to offer a full and homogeneous account of the way linguistic knowledge is acquired, stored and processed.

However, one possibility for word reading is that the full TP mechanism (involving the Elsewhere Condition and serial search) is used only for the *acquisition* of orthography-phonology correspondences, and not as an online process during skilled reading. In this way, it could be employed by learners to assess which pronunciation patterns should be abstracted to form generalisable rules; a mechanism by which to extract useful information from the input. Those rules that are formed during development according to the TP's consistency metric (motivated by computational efficiency) may then become frozen within a separate, skilled reading system. Therefore, a serial search of a frequency-ranked list of exceptions would not need to be undertaken before application of a productive rule during online, skilled reading, and instead readers would use a hierarchical structure of productive pronunciation rules that has been left as an artefact of the acquisition process.

The concept of rule-freezing may seem incompatible with the TP's underlying motivation to determine the most computationally efficient route of access during real-time, online processing. However, rule-freezing is included in Yang's original theory, which states that learners "stop looking" and "freeze the rules in place at a value of *N* no more than a few hundred" (Yang 2018b, p. 803). This rule-freezing enables decisions about productivity to be made when a learner's vocabulary is relatively small. Yang's theory does not address why exception items acquired after rule-freezing – which presumably must still be stored lexically, increasing the serial search time during online processing – would not disrupt the carefully measured balance between the real-time processing speeds of alternative access routes. After all, the TP theory was intended to offer an account of online lexical access. Nevertheless, if this seemingly counterintuitive feature is permitted by Yang's theory, perhaps it is reasonable here to go further and propose that - for reading at least - the TP offers a suitable mechanism for extracting productive rules from the input during development, but *not* a model of online skilled reading.

Certainly, there is evidence of a close relationship between what is learned from the input and what can be generalised to novel items, in favour of an approach according to which patterns acquired during development form the basis of a skilled, fully productive system. This evidence includes results from Experiment 3 in Chapter 5, in which the high token frequency of irregular items during training disrupted the pattern of generalisation observed in Experiment 2. In particular, I argued that the distribution of regular and irregular items that a learner has actually been able to acquire (with those occurring in the input at higher token frequency being easier to learn) goes on to determine which patterns are used productively. Taking these results together with the current discussion suggests that future research must look more closely at the overarching process whereby knowledge of individual items is acquired by learners, the consistency of the patterns those items fall into is assessed, and qualifying patterns are subsequently generalised using productive rules. These stages certainly seem to be closely related and their inter-relationship should be understood further. However, perhaps a theory which uses a single mechanism to capture in one fell swoop the acquisition of items by learners, the forming of productive rules, and the storage and access of knowledge by skilled users, is too ambitious. Instead, a more discrete, multi-stage process, whereby evidence is accumulated during development to form productive patterns, which are subsequently frozen to create a skilled productive rule system, may be more appropriate for reading.

I suggest that the concept of the freezing point demands further consideration. Yang's discussion of rule-freezing remains relatively vague: "it is plausible to conjecture that language acquisition never uses *N*'s beyond a certain value, probably just a few hundred. It is doubtful that anyone can keep track of large values of *N* and *e*; perhaps learners will simply freeze the rule once they have seen enough data, i.e., a sufficiently large value of *N*" (2021, p. 5). It is conceivable that rule-freezing could emerge as an organic property of the acquisition process; that once sufficient evidence to support a rule has been encountered, any items acquired subsequently are unlikely to disrupt the balance. In this way, the rule would be effectively frozen as consequence of the distribution of the input, rather than by a shift in the nature of the learning process.

However, Yang's argument (2016) states that in some cases, a productive rule (such as the stress rule in English, or the noun-determiner rule in English) is *only* learnable with a small vocabulary. Were a larger vocabulary to be used, the number of exceptions would breach the threshold and a productive rule would not be supported (particularly given that the tolerance threshold becomes proportionally lower as *N* increases). Therefore, rule-freezing in these instances cannot be explained by the distribution of the input alone, because later acquired items *would* disrupt the balance. Yang (2016, p.225) indicates that this (sometimes surprisingly) early rule-freezing may unfold alongside developmental changes in children's processing capacity, such as those described under the "Less is More" hypothesis (Elman, 1993; Newport, 1990). However, he suggests elsewhere that the TP operates for both child and adult learners (2018b, p. 801), leaving rule-freezing by adults unexplained. In summary, it is conceivable that some type of rule-freezing takes place, but the exact reason for this process, and the precise point at which it occurs, requires further investigation.

There is also second possible conclusion that could be drawn for word reading in response to the work presented here. The application of the TP to reading is novel because it offers a threshold of consistency that determines when a grapheme-phoneme rule should be used productively, and when a more specific word body rule should be sought instead. Results from Experiment 1 suggested that this threshold, based on type frequency counts, predicted participants' nonword pronunciations more successfully than computational models of reading that use alternative processing mechanisms. However, just because the threshold provided by the TP was a better predictor of nonword pronunciations than extant models, it is not necessarily the best or only threshold that could be used. Indeed, participants did not display the categorical behaviour that was predicted by the TP in any experiment reported here; there is variability in readers' behaviour that remains to be explained. This second possibility circumvents any long-standing issues with the serial search mechanism that may remain, as an alternative consistency threshold may not involve serial search at all. Further, it opens the door for future research to investigate alternative tipping points of consistency between productive and unproductive rules which can also predict use of grapheme vs. word body correspondences. Importantly, as will be discussed in *Section 7.3.*, what is clear from the current investigation is that an approach to word reading which predicts use of a variety of grain sizes according to a threshold of consistency appears to be a valuable extension beyond extant models.

To summarise the applicability of the TP to word reading at a broad level, the theory offers a promising account of the way in which readers generalise spelling-sound correspondences as demonstrated in the experiments presented here. However, there remain mechanistic issues raised above which suggest that in its original implementation, the TP may not be an optimal account of skilled word reading. Nevertheless, I have proposed ways in which these could potentially be overcome or expanded from in future.

7.3 Contributions to wider literature

The experiments reported in this thesis addressed a number of wider research aims beyond an examination of the TP itself, and their findings contribute to multiple fields of study. These include research on skilled word reading, reading development and instruction, and statistical learning. The following sections present the ways in which the current findings inform this broader range of research.

7.3.1 Aims of the thesis

7.3.1.1 Aims of the thesis related to word reading

The word reading literature has been dominated by debate over the most appropriate ways to characterise readers' spelling-sound knowledge, and how to predict generalisation of this knowledge in nonword pronunciations. Specifically, there is a dichotomy between rule-based models such as the DRC (Coltheart et al., 2001) which predict use of the most common GPCs for nonword pronunciations, and statistical models such as the Triangle model (Seidenberg & McClelland, 1989; Harm & Seidenberg, 2004) and the CDP+ model (Perry et al., 2007) which allow the graded consistency of multi-letter sequences to inform nonword pronunciations. These approaches also differ in how they use frequency: the DRC uses *type* frequency to count the most frequent GPCs, whilst the statistical models also use *token* frequency to weight the strengths of connections between units.

Nonword reading aloud studies have been widely used to adjudicate between computational models, and to gain insight into readers' orthography-phonology knowledge more generally (including Seidenberg et al., 1994; Andrews & Scarratt, 1998; Treiman et al., 2003; Pritchard et al., 2012; and Mousikou et al., 2017). Results from these studies can be construed in different ways, largely because reading behaviour is not categorical. However, these overall trends seem clear: adult readers use GPCs to pronounce nonwords most often, but also make use of alternative pronunciation patterns, suggesting sensitivity to the consistency of larger orthographic units beyond the individual grapheme. Such responses cannot easily be accounted for by the DRC, but neither is reading behaviour captured sufficiently well by statistical models, which tend to predict context-sensitive pronunciations and lexicalisations at a higher rate than used by participants. Indeed, readers' use of context-sensitive pronunciations tends to fall below the rate expected by corpus statistics (Treiman et al., 2003; Treiman & Kessler, 2019), suggesting that readers do not simply reproduce the distributions of pronunciations they are exposed to. Overall, there exists a gap in our understanding of readers' orthography-phonology knowledge and its generalisation, which is not well accounted for by extant models.

Some researchers have indicated possible ways to address these issues, such as determining a trade-off between the frequency or consistency of an individual grapheme and the cost of learning a context-sensitive pronunciation (Kessler, 2009; Steacey et al., 2019), or by assessing the utility of different spelling-sound correspondences in a quantitative way (Vousden, 2008). Thus far, however, no fully-developed account has been proposed. Therefore, one aim of this thesis was to investigate whether a theory of spoken language acquisition and generalisation, the Tolerance Principle (Yang, 2016), could aid our understanding of these outstanding matters in the field of reading and bridge the gap between existing computational models.

Indeed, the TP offers a promising set of tools with which to investigate these long-standing questions. The TP’s type-frequency based metric of consistency makes specific predictions on the basis of a quantitative assessment of spelling-sound correspondences in the input. This metric determines a categorical threshold of inconsistency beyond which a more general spelling-sound correspondence (e.g. at the grapheme-phoneme level) should not be generalised. Crossing this tolerance threshold triggers the recursive application of the TP, in a search for a more specific productive pattern within a subset of the input (e.g. at the body-rime level). In these ways, the TP offers potential to resolve previous debates; the experiments reported in this thesis set out to assess its success in doing so.

Importantly, the current application of the TP to reading set out to provide an initial investigation into whether the algorithm supplied by Yang (2016, p. 9) offers suitable predictions for readers’ generalisation behaviour. The theory itself deals only with numbers: a type-based tally of regulars and irregulars used to generate a numeric threshold of productivity. Therefore, the TP has been employed here simply as a mathematical model, *not* a mechanistic model of reading akin to the fully-developed DRC, Triangle or CDP+ models. Yang’s theory has nothing to offer regarding the process of reading *per se*, and therefore the TP cannot be directly compared with these established computational models of word reading in terms of their ability to capture all aspects of the reading system. Instead, I set out to assess whether the predictions of the TP algorithm offered a more successful account of readers’ nonword pronunciations than the outcomes of three computational models. Potential integration of the TP mechanism within a more established framework, such as development into the non-lexical route of a dual-route model of reading, lies outside the remit of this thesis but could be taken up in future work.

7.3.1.2 Aims of the thesis related to reading development and instruction

Research on reading acquisition reveals a consistent trend whereby the earliest readers rely on the most frequent GPCs in novel word pronunciations, but make increasing use of alternative (i.e. context-sensitive) mappings throughout reading development (Marsh et al., 1981; Treiman et al., 1990; Laxon et al., 1991; Coltheart & Leahy, 1992; Brown & Deavers, 1999). However, pinpointing precisely when readers make use of different correspondences according to orthographic context or reading skill level has been difficult to achieve (Brown & Deavers, 1999; Treiman et al., 2003; Kessler, 2009; Steacy et al., 2019). Accordingly, one aim of this thesis was to improve our understanding of the trajectory towards skilled reading by comparing adults' and children's use of spelling-sound correspondences at different orthographic levels with the TP's specific predictions.

Further, I set out to investigate the specific process by which learners extract and use more complex (i.e. context-specific) spelling-sound mappings without explicit instruction. Adult learners in an artificial orthography paradigm have demonstrated sensitivity to context-sensitive spelling-sound mappings during learning and generalisation without instruction (Taylor et al. 2011). However, research using English orthography has highlighted that both developing and skilled readers generalise context-sensitive correspondences less often than would be predicted by their consistency in English words (Treiman et al., 2003). Participants' use of these context-sensitive correspondences increased gradually with reading skill. Here, I considered under which orthographic conditions learners are able to extract and generalise such patterns from the input.

The quantitative approach used here to examine which spelling-sound correspondences are most useful (or most difficult) for developing readers may have additional value when considering whether particular mappings should be highlighted during reading instruction. For instance, there may be correspondences which efficiently capture the statistical distributions of text according to the TP, but which take time for developing readers to be able to generalise and could thus benefit from targeted instruction.[19] Previous research has similarly applied a quantitative analysis to the statistical distributions of text, aiming to identify the most efficient ways to capture spelling-sound correspondences in English, and thereby inform reading

[19] After all, the TP was developed as a theory of spoken language acquisition, whilst learning to read is a substantively different process that requires explicit teaching.

instruction (e.g. Vousden, 2008; Vousden et al., 2011). The current work aimed to extend this type of research by using a quantitative analysis of text input to predict the productive use of specific correspondences based on an interaction between the consistency of mappings at different orthographic levels.

My investigation into reading development involved two additional considerations. Firstly, to bear in mind the role of an individual reader's text experience as they build a system of spelling-sound knowledge, particularly as the TP theory is built on the assumption a learner's productive rule system is the product of their specific input. Therefore, we might expect readers with similar reading experience to behave more closely, and readers with more disparate experience to behave more variably. Secondly, children learning to read English in UK schools undergo a systematic phonics instruction programme. As this method explicitly teaches the most frequent GPCs, it was unknown how it may interact with the TP predictions for readers' use of grapheme- or body-level correspondences.

7.3.1.3 Aims of the thesis related to statistical learning

At a broad level, this thesis set out to explore how readers make use of the statistical distributions in their input to build productive knowledge of spelling-sound correspondences. Examining the TP's predictions could offer new insights into learners' use of the patterns they are exposed to, by assessing the importance of different statistical properties during this process. This line of enquiry has specific relevance for the study of statistical learning - both within and beyond reading - by building on previous literature to address a number of research aims.

Principally, the research presented here aimed to refine our understanding of generalisation within a quasi-regular system. Learners are exposed to patterns which vary in consistency, but are nonetheless able to build productive systems on the basis of this input. Precisely how consistency of the input affects the outcome of generalisation has been a matter of ongoing investigation. Statistical learning research has found that following exposure to inconsistent patterns, children are more categorical in their generalisation, whilst adults tend to reproduce the variation they are exposed to (Hudson Kam & Newport, 2005, 2009; Schuler 2017). Meanwhile, research on reading has considered the applicability of statistical learning mechanisms and the importance of input distributions in this domain (e.g. Taylor et al., 2011; Arcuili & Simpson, 2012). The current studies aimed to build on this body of research by using a quantitative measure of consistency to examine the nature of adults' and children's generalisation of inconsistent familiar and novel spelling-sound patterns.

Beyond consistency, these studies also investigated the effect of other input variables on generalisation, namely the type and token frequency of regular and irregular items during training. Previous research has suggested that type rather than token frequency of input items affects rule productivity directly (Endress & Hauser, 2011; Perfors, et al., 2014; Schuler, 2017); the TP also maintains that the calculation of productivity should involve only type frequencies. However, token frequency may be more informative in instances where productive rules are not formed (Hudson Kam & Newport, 2005, 2009; Schuler, 2017), or could play an alternative role by determining which items are learned most quickly and accurately (Endress & Hauser; 2011; Kurumada et al., 2013). The contribution of this investigation therefore is two-fold: both informing our understanding of input frequencies for acquisition and generalisation, and allowing a rigorous test of the TP theory.

As discussed above, readers are able to make use of more complex spelling-sound mappings, but do so less often than predicted by the distribution of these mappings in text (Treiman et al., 2003; Treiman & Kessler, 2019). From a statistical learning perspective, this finding suggests there is some process readers undertake in order to extract and use these patterns, rather than simply reproducing the input distributions. I set out to assess whether this process is successfully captured by the TP mechanism. In so doing, I could examine whether the TP provides a missing link not only for research on reading, but also for statistical learning.

Finally, drawing together results from the experiments reported here may contribute to our understanding of the relationship between acquisition and generalisation. For instance, early-acquired items may subsequently form the basis of generalisation (Yang, 2016).

7.3.2 Key findings

Experiment 1 investigated participants' pronunciations of nonwords written in English orthography. To pronounce items in vowel winner conditions, participants used the vowel winner pronunciation (the most common pronunciation of the vowel grapheme which passed the tolerance test) the majority of the time. This result supports the TP's prediction that the most general spelling-sound correspondence which is sufficiently consistent to pass the tolerance test (i.e., at the level of the grapheme) should be used productively. The rule-based DRC performed as well as the TP in vowel winner conditions, as both accounts predict use of the most common pronunciation for these items. Meanwhile, the statistical models performed less well, as they overpredicted interference from the level of the word body for items which have a relatively consistent vowel grapheme.

However, use of the vowel winner pronunciation was notably lower for one group of vowel winner items. These were from the vowel winner, body winner, conflict condition, where items have an alternative, consistent pronunciation of the body that conflicts with the vowel winner pronunciation. In participants' responses for these items, the alternative body pronunciation seems to interfere with the vowel winner pronunciation to some extent; this interference is *not* predicted by the TP. In this way, the TP cannot precisely capture the way that readers resolve the conflict between alternative pronunciations for items with a relatively consistent vowel grapheme, and this remains to be explained by future research. Nevertheless, the vowel winner pronunciation remained the most common response for these items overall.

Items from the vowel fail, body winner condition have inconsistent vowel graphemes but a consistent pronunciation of the word body which passes the tolerance test. Here, the TP predicts that this body winner pronunciation should be used, according to the recursive application of the tolerance threshold. As predicted, participants used the body winner pronunciation more often in this condition than in the vowel winner, body winner, conflict condition, and further, the TP was a closer match to participants' responses for these items than any of the three computational models. In this way, the TP seems to capture something that previous research has not been able to: namely, the way in which readers use contextual information (or larger orthographic units) to inform their pronunciations when an individual grapheme is inconsistent. Therefore, Experiment 1 provided evidence that the TP is able to predict readers' use of familiar body-level correspondences using the recursive application of a categorical threshold of consistency.

Comparison of adults’ and children’s pronunciations in this experiment revealed that children’s use of the vowel winner pronunciation was higher than adults’ in the vowel winner, body winner, conflict condition. This finding is in line with previous research suggesting that younger readers rely more on GPCs than skilled readers (Marsh et al., 1981; Coltheart & Leahy 1992; Brown & Deavers, 1999), although it is notable that this difference was not observed in the other vowel winner conditions. It is possible that the explicit phonics training the children received at school boosts their use of these pronunciations in their responses compared to adults’, and specifically helps them to override a conflicting body pronunciation. Importantly, children’s behaviour in this condition is more closely aligned with the TP prediction than adults’, although it is unknown whether this could also be a result of their instruction or their reading experience. Alternatively, it could be the case that adults are less categorical in their reading behaviour, and more likely to diverge from TP predictions – perhaps due to their more varied text experience.

Adults’ and children’s nonword reading behaviour also diverged in the body winner conditions, in which a pronunciation of the word body is consistent enough to pass the tolerance test. Across these conditions, child participants used the body winner pronunciation less often than adult participants, which supports previous evidence that children make less use of the body unit than adults when reading aloud (Marsh et al., 1981; Treiman et al., 1990; Bruck & Treiman, 1992; Coltheart & Leahy, 1992; Brown & Deavers, 1999). This result is also consistent with the suggestion that extensive text experience is required to use these more complex orthography-phonology mappings (Treiman et al, 2003; Treiman and Kessler, 2019) particularly in the absence of explicit instruction of these correspondences.

Experiment 2 used an artificial orthography to investigate the generalisation of novel vowel symbols with inconsistent pronunciations. A rule-based approach would predict that the most frequent pronunciation of each vowel symbol should be used productively to pronounce untrained items (i.e. “regularisation”), whilst a statistical approach predicts that generalisation should be based on the frequency distribution of pronunciations in the input (e.g. matching the type or token frequency of alternative pronunciations during training). Meanwhile, the TP predicts that participants should regularise their pronunciations of the two vowel symbols that pass the tolerance test, but should not regularise the pronunciation of the third vowel symbol which does not. Results from the Generalisation task did not support a rule-based approach, as the rate of participants' regularisation was not high in all three conditions. Whilst token frequency did have an effect on regularisation as a statistical account would predict, passing the TP had an effect beyond this for both adults and children. Therefore, it seems that the TP is able to account for variance in participants' regularisation that token frequency cannot. Overall, these findings suggest there is a categorical effect of consistency on learners' generalisation of novel spelling-sound correspondences which is in line with the predictions of the Tolerance Principle, but not existing rule-based or statistical models of reading.

Similarly, results from Experiments 1 and 3 suggest that the categorical metric of consistency provided by the TP offers a novel contribution to our understanding of the relationship between the consistency and generalisation of a pattern. In Experiment 1, participants often regularised the pronunciation of vowel graphemes that had relatively high inconsistency in the corpus, yet were predicted to be regularised by the TP. This effect was additional to that of a continuous, type-based consistency measure, whilst a continuous, token-based consistency measure did not predict regularisation. This result is striking, as it suggests that the TP modulates the relationship between consistency and regularisation. When irregular items occurred with high token frequency in Experiment 3, the TP was better able to predict vowel regularisation after a revised threshold was calculated on the basis of the items each participant had successfully acquired. However, participants' behaviour across all experiments was not as categorical as the TP predicts.

This research also offered insights regarding participants' use of token frequencies from their input. In Experiment 1, participants' pronunciations of nonword items without a TP prediction offered an indication that token frequency information from the grapheme and body level can inform nonword pronunciations by adults and children. In Experiment 2, token frequency of trained regular items had an effect on adults' and children's regularisation, suggesting that this statistical property is used in generalisation. However, the TP had an effect above token frequency, which indicates that during generalisation, participants are imposing some additional structure on their input beyond the frequency distributions they were exposed to. Experiment 3 manipulated the token frequency of trained items, with irregular items assigned to the top of the frequency distribution during training. An initial analysis suggested that token frequency still had a significant effect on regularisation. However, analysis using a recalculated tolerance threshold for each participant revealed a different pattern of results: whilst this individual TP predicted participants’ regularisation, token frequency did not. This finding indicates that token frequency has an indirect role to play in the generalisation process; items that are highly frequent in the input may be learned most successfully; these items may subsequently form the basis of generalisation.

Experiment 4 investigated the recursive application of the tolerance threshold for novel vowel symbols. This experiment offered an opportunity to explore whether the TP could predict when learners use contextual information to inform the pronunciation of an inconsistent grapheme within an artificial orthography paradigm. Results from this Generalisation task suggested that adults had difficulty in extracting and generalising context-dependent spelling-sound correspondences; only four participants were able to extract the context-sensitive sub-regularity that was available for the vowel symbol which did not pass the tolerance test at the individual grapheme level. Those participants that did use the sub-rule categorically scored relatively well on accuracy of trained items, lending weight to previous findings which suggest that learning such context-sensitive correspondences is difficult and takes time to achieve. Moreover, participants regularised the vowel pronunciation less often for this symbol than for the symbol that did pass the tolerance test at the grapheme level, indicating that more broadly, participants are sensitive to the consistency of spelling-sound correspondences as the TP would predict.

There are parallels to be drawn between the difference in adults’ and children’s use of the body winner pronunciation in Experiment 1, and the behaviour of adult participants in Experiment 4 developing orthography-phonology knowledge within an artificial orthography paradigm. As only a few participants formed a productive, context-sensitive sub-rule, the adult readers in this learning paradigm could be mirroring the behaviour of younger readers of English who made less use of context-sensitive information than might be expected given the distributions of their input. These results highlight the challenge of extracting and generalising more complex spelling-sound mappings, even when they offer increased consistency compared to simpler grapheme-phoneme mappings.

The participants who did form the sub-rule were ranked in the top 30% of participants according to their accurate pronunciation of all trained items, and were at least 50% correct on trained subset items (which featured the critical word body). This result supports previous findings that more skilled readers are able to make increasing use of context-sensitive information (Treiman et al., 1990; Laxon et al., 1991; Coltheart & Leahy, 1992; Treiman et al., 2003). Further work is needed to specify the precise accumulation of this knowledge at different stages of reading development, but broadly speaking, these results hint that secure knowledge of items that involve more complex mappings may be necessary to use these patterns productively. The importance of acquiring secure knowledge to support generalisation was also highlighted in Experiment 3, where an individual tolerance threshold, based on the specific trained items each participant had accurately learned, successfully predicted generalisation behaviour.

Nevertheless, understanding an individual's text experience may provide some additional insight into the trajectory of their reading development. For instance, the TP was not able to account for variability across participants in Experiment 1, and contrary to my predictions based on the TP theory, children's nonword pronunciation responses were more variable than adults' despite their more similar reading experience and instruction. It is possible that differences between the orthography-phonology knowledge of developing readers are more pronounced due to their individual levels of progress, but that with cumulative years of reading experience and increased reading skill, readers' rule systems may begin to converge and thus produce more similar responses.

7.3.3 Implications for wider literature

7.3.3.1 Implications for word reading

Taken together, these findings provide a range of insights into skilled and developing reading behaviour, which can also be used to assess the TP in the context of extant models of word reading and address the outstanding questions laid out in *Section 7.3.1.1*. Indeed, they suggest that the TP offers valuable advances in our understanding of this area of research. In particular, there is a categorical effect of consistency on readers' generalisation that is not accounted for by models of reading, whereby readers are more likely to use a spelling-sound correspondence productively when the consistency of this correspondence falls beneath the tolerance threshold. In a familiar orthography, the TP is also able to predict more successfully than extant models the way in which readers use context-sensitive information to inform the pronunciation of an inconsistent grapheme. However, the TP was less successful in predicting learners' ability to extract and generalise context-sensitive information for inconsistent graphemes in a novel artificial orthography. Further research is needed to explore whether participants' behaviour in this setting would align more closely with the TP predictions given further training on the artificial orthography. Additionally, readers in Experiment 1 demonstrated some level of interference from the word body in their pronunciations of nonword items with a consistent vowel grapheme pronunciation and a consistent, conflicting body pronunciation. In this way, readers' behaviour was less categorical than the TP predicted; even when a vowel grapheme is sufficiently consistent to pass the tolerance test, readers will use a consistent conflicting body pronunciation some of the time. The current findings have highlighted that readers' behaviour is not categorical in this context, but future research should address why this is the case, and precisely how readers construct pronunciations in such instances.

Overall, I propose that the TP's novel role for consistency may lie behind the progress it has made in predicting readers' generalisation of spelling-sound correspondences more successfully than computational models of reading. Specifically, its use of type frequency counts in a quantitative assessment of consistency to provide a tipping point of productivity, along with its recursive application, enables precise predictions to be made on the basis of input statistics. These predictions include which spelling-sound correspondences should be generalised, when information from different orthographic grain sizes should be used, and how readers extract productive patterns from a quasi-regular input. This role for consistency allows the TP to bridge the gap between established rule-based and statistical models and goes some way towards addressing their shortcomings in accounting for readers' behaviour. Future approaches to word reading that seek to improve upon the TP should also be able to capture the categorical effect of consistency described above; to predict the use of smaller versus larger orthographic grain sizes; and additionally, to account for those specific instances in which readers use context-sensitive information more or less often than is warranted by the consistency of an individual grapheme.

7.3.3.2 Implications for reading development and instruction

The investigation of the TP presented here has a number of implications for the study of reading development and instruction. Firstly, it is striking that even through use of an adult corpus, the TP was able to predict 9-10 year-old children's nonword pronunciations to a good level (72%), and more successfully than extant models. The TP was also able to predict children's generalisation in an artificial orthography beyond frequency effects. These findings suggests that the TP does add to our understanding of reading development by demonstrating that younger readers are sensitive to consistency at different orthographic levels according to a categorical threshold. This effect is not predicted by extant models of word reading. Additionally, the findings open the possibility that use of the TP with corpora that reflect different stages of literacy development would be a valuable undertaking in order to understand this trajectory further.

It is clear that the acquisition and generalisation of context-sensitive spelling-sound patterns or sub-regularities is not a trivial task for developing readers. Consequently, the current findings lend support to previous research suggesting that explicit teaching might be required for some learners to use these patterns productively (Treiman & Kessler, 2019). However, the approach used here could be particularly valuable by helping to identify precisely which correspondences should be taught. Results from this investigation suggest that spelling-sound correspondences involving larger orthographic units are sometimes more reliable than those involving smaller orthographic units. This finding is in line with results from other quantitative approaches which also aimed to capture the most efficient correspondences between spelling and sound (Vousden, 2008; Vousden et al., 2011). However, the current work extends this previous research in a profitable way by offering an assessment of consistency for individual mappings at different levels and predicting generalisation accordingly. Armed with the knowledge that specific spelling-sound correspondences are sufficiently consistent to be efficient for generalisation, and yet that they may be difficult to acquire implicitly, we could perhaps inform reading instruction by explicitly targeting these correspondences.

Certainly, this is not necessarily to say that the earliest stages of reading instruction should involve the teaching of complex spelling-sound mappings. In a longitudinal study, Shapiro and Solity (2016) compared the progress of children learning to read through two different phonics programmes: Letters and Sounds, which teaches alternative letter-sound mappings, and Early Reading Research, which teaches the most consistent letter-sound mappings and a set of frequent words by sight. Whilst there was no overall difference between the efficacy of the programmes according to later reading attainments, children with poor phonological awareness at school entry gained higher reading scores at the end of the first year under Early Reading Research than Letters and Sounds. Therefore, the authors suggest that simplifying phonics programmes to include only the most consistent correspondences (plus a set of sight words) may be beneficial. However, this study does not address the possibility that introducing alternative, context-sensitive mappings at later stages of the instruction programme could be of additional benefit, and might not interact with early phonological awareness in the way that Letters and Sounds was found to do. After all, there were no long-term differences in levels of reading attainment under each programme, which suggests that a phonics programme which includes more complex mappings is effective overall.

In a similar vein, Bruck and Treiman (1992) found that explicit teaching of body-level analogies to early readers did not result in successful generalisation of these correspondences, although this method did convey early learning benefits above the teaching of word-initial CV- or grapheme-level analogies. Whilst this finding may indicate that teaching more complex correspondences involving larger orthographic units is not advantageous for the development of productive knowledge of spelling-sound correspondences, I would highlight that this study specifically involved beginning level readers. Therefore, I suggest that it may be worth investigating further whether targeted instruction of certain context-sensitive spelling-sound correspondences during later stages of reading instruction offers improved reading outcomes.

For children with reading disability, Steacy et al., (2016) investigated the effect of reading instruction programme on transfer of decoding skill. They compared Phonics for Reading (a synthetic phonics programme which teaches phonological awareness and GPCs) with Phonological and Strategy Training (a programme which teaches variable vowel pronunciations, sight words, and morphological strategies, in addition to phonological awareness and simple GPCs). Whilst there was no overall difference in effectiveness of the two instruction programmes, children under Phonological and Strategy Training performed better at decoding words with variable (i.e. non-GPC) vowel pronunciations, whilst children under Phonics for

Reading performed better at decoding words without variable vowels (i.e. those that use GPCs). Overall, the study indicates that the sub-lexical emphasis of an instruction programme can affect specific transfer gains for word reading. Additionally, the authors highlight the benefit of teaching less able readers flexibility with vowels through a variety of strategies, particularly considering the difficulties these readers may have with vowel representations (e.g., Ehri & Saltmarsh, 1995; Shankweiler & Liberman, 1972). This research demonstrates the value of comparing specific instruction dimensions that vary across teaching programmes. However, a more extensive comparison of a range of programmes with typically developing readers is still required.

This endeavour should include a comparison of instruction programmes which vary in their teaching of graphemes with multiple pronunciations and context-sensitive units at different stages of reading acquisition. For instance, Solity (2020) reports that the Letters and Sounds programme teaches 34 graphemes with multiple pronunciations; Read Write Inc. teaches 13; Jolly Phonics teaches 5; and Optima Reading teaches none. The specific benefits of these different instructional approaches are not known. Nor do we fully understand the ways in which readers' development might be affected by the inclusion of context-specific cues for alternative pronunciations, or when these more complex correspondences should be introduced. A systematic comparison of existing programmes would help us to determine the optimum number of spelling-sound correspondences and the order in which to deliver them, given the limited amount of research in this area. Furthermore, the TP offers a potential framework within which these questions could be explored, and an objective measure against which to compare alternative schemes. By determining whether a spelling-sound correspondence is sufficiently consistent to pass the tolerance test given the size of a learner's vocabulary and the properties of the items within that vocabulary, the TP can thereby indicate which correspondences should be targeted during instruction at different stages. An assessment of the correspondences taught by different phonics programmes could be carried out in this way, to reveal which programmes offer the optimum delivery of spelling-sound correspondences according to the TP.

Beyond a comparison of existing phonics schemes, this method could also inform the development of new approaches towards reading instruction. For instance, Compton et al.'s (2014) proposed connectionist approach involves the use of carefully constructed corpora and the instruction of spelling-sound correspondences at multiple levels alongside word-specific representations. As words are systematically added to the lexicon, probabilistic learning of constraints between orthographic and phonological units is expected to take place. In this context, the TP could again be used to identify which correspondences should be taught at which stage, according to the size and properties of the training corpus.

Given the limited amount of research in this area, Castles et al. (2018) called for a systematic investigation into the efficacy of teaching single-letter or multiple-letter grapheme-phoneme mappings and their consistency in certain contexts. Following the results presented here, I echo this call, and suggest that the TP’s quantitative approach offers a valuable tool with which to carry out this inquiry. The current research offers an initial advancement towards this aim by identifying a set of productive spelling-sound correspondences at different orthographic levels. The next step would be to assess whether teaching this range of mappings at different stages of reading instruction (as discussed above) offers an additional benefit to young readers compared to the use of an individual, regular grapheme-phoneme correspondence approach. At a broader level, Treiman (2018) also questioned whether phonics instruction could be improved, whilst highlighting the difficulty phonics advocates face in suggesting adjustments lest they are seen to weaken their stance. I suggest that the most effective way to approach this challenge is by building the body of empirical findings which highlight the potential benefits of an increasingly nuanced instruction programme.

Overall, the findings presented in this thesis suggest that an understanding of reading development requires close attention both to the text experience a child has received, and the items they have managed to acquire; particularly as their pattern of acquisition and generalisation of spelling-sound knowledge may be shaped to some extent by the distributions of their own specific input. Most pertinently, an individual leaner’s trajectory towards skilled reading will be precisely that – individual – and may additionally require explicit instruction in order to take advantage of more complex correspondences between spelling and sound.

7.3.3.3 Implications for the study of statistical learning

The findings described above offer some important insights for the statistical learning literature which must be accounted for in future work. Most strikingly, we now have a greater understanding of the process whereby learners generalise from the statistical distributions in their input. Specifically, if the consistency of a pattern falls within a critical threshold of tolerated exceptions, learners are able to use this pattern productively in a way that goes beyond the statistical distributions of the input. The observed shift in generalisation behaviour as the consistency crosses this threshold lends support to recent evidence of a limit to the level of inconsistency from which learners can base generalisation (Schuler, 2017; Schuler et al., 2021). However, the current results go further by demonstrating that this type-based categorical effect of consistency on generalisation may be additional to any effects of token frequency, and also to the effects of continuous type- or token-based consistency measures.

Whilst we have seen across these experiments that there may be a categorical effect of consistency on participants' generalisation, it is also clear that participants' generalisation was not categorical - in contrast to Yang's (2016) claim. However, it is not known whether learners' generalisation in the context of artificial orthography learning would become more categorical with improved knowledge of trained items. We also did not observe a distinction between children's categorical generalisation and adults' more probabilistic behaviour, as found in some previous research (Hudson Kam & Newport, 2005; Schuler et al., 2021).

Another significant contribution is the evidence from Experiment 3 that token frequency may play an underlying role in the path to generalisation: token frequency might be important for secure learning of individual items, but does not necessarily disrupt productivity directly. Rather, it may determine which items are used as the basis for generalisation. This role for token frequency is consistent with the TP account, whereby the earliest balance of productivity will be based on a small number of high frequency items, and may be adjusted as vocabulary knowledge grows. However, it is less easily accounted for by a statistical learning approach that predicts a direct effect of token frequency on generalisation (at least for adults).

More broadly, it seems that there is a close relationship between learning and generalisation. This conclusion was indicated by participants' generalisation in Experiment 3 which demonstrated an effect of a tolerance threshold based on individuals' successfully-acquired trained items beyond the effect of a threshold based on all input items. A similar suggestion was made in Experiment 4, where only learners with knowledge of relevant trained items demonstrated categorical use of a sub-rule. Accordingly, understanding which items are easier to acquire may help define the path towards productivity, as generalisation may be based on the forms a learner has acquired rather than those they are exposed to. However, it is also possible that a consistent pattern observed across items may support learning of these individual items; future work should aim to further uncover this potentially bi-directional relationship between the acquisition and generalisation of productive patterns.

Overall, this thesis offers a more fine-grained account of the way learners use input statistics to form general rules that can be applied to unfamiliar items. Namely, it is has provided evidence of a categorical effect of consistency on generalisation beyond other input statistics, whilst suggesting how such input statistics may be important for the acquisition of individual items, and subsequently the generalisation of patterns across these items.

7.4 Limitations of this thesis and future directions

A potential limitation of Experiment 1 is that nonword pronunciations predicted by the TP were generated using word frequencies from an adult corpus, against which pronunciations from both adults and children were assessed. If a suitable children’s corpus were available then a separate set of thresholds could be calculated to generate specific predicted pronunciations, and then compared with children’s responses. This approach may be more appropriate for the investigation of child reading behaviour. For instance, it may allow more specific insights to be gained into the applicability of the TP for the development of spelling-sound knowledge, particularly if the corpus reflected the growth of children’s text experience at different stages. For instance, it would allow a closer investigation into the relationship between accumulating knowledge of more complex spelling-sound correspondences and the generalisation of such patterns.

A further limitation of this thesis is that the recursive application of the TP to acquire and generalise context-specific spelling-sound correspondences was not assessed with children in an artificial orthography paradigm.[20] Conducting this research would not only allow a closer investigation of the recursive application of the TP in younger learners, but may also be informative more generally for our understanding of the effect of age-related differences and maturational constraints on extracting complex patterns from input distributions. Another age limitation involves the age group of children participating in Experiments 1 and 2. Whilst adult and child participants (aged 8-10) did not display vastly different patterns of behaviour in these experiments, it is possible that more distinctive developmental effects may be revealed by examining the generalisation behaviour of younger-aged children. In this way, an investigation of the TP in word reading with early-stage readers could be informative for both our understanding of the TP theory and the trajectory of reading acquisition.

Additionally, the context-specific spelling-sound correspondences used in Experiment 4 involved only contingencies between the word-medial vowel and the word-final consonant. This methodological decision allowed a comparison with parallel research involving the word body in English orthography. However, it is possible that adult participants' previous experience of body-rime correspondences in English could have interfered with their extraction of correspondences involving these orthographic units within a novel orthography. Future investigations which explore dependencies between word-initial consonants and word-medial vowels (which are less common in the English writing system (Treiman et al., 1995)) would allay such concerns.

Although the TP was found to successfully predict generalisation a number of times across the studies reported here, generalisation was also not as categorical as the TP account would expect. For instance, few participants demonstrated the recursive application of the TP in an artificial orthography learning context in Experiment 4. This current research is not able to address whether participants' behaviour would become increasingly categorical - and potentially more aligned with the TP - after additional training which would allow knowledge of trained items to become secure. In Experiment 4 specifically, participants' successful acquisition of the artificial language may have been negatively affected by carrying out the study online rather than in person. Therefore, the research reported here could be informatively expanded by using a

[20] Conduct of this experiment with child participants was cancelled due to the restrictions introduced during the Covid-19 pandemic.

more intensive and prolonged training phase (preferably carried out in person) which allows learners to develop more secure knowledge of the artificial language before generalisation is elicited. Certainly, future research should ensure that close attention is also paid to the specific trajectory of individual learners' patterns of acquisition and generalisation.

Further insights may also be gained by looking more specifically at the interaction between direct instruction and the effect of the TP on generalisation. The current research focused on the acquisition and generalisation of spelling-sound correspondences without direct instruction of these mappings. However, there is a range of possibilities for future research involving explicit instruction. For instance, research could address whether direct training of correspondences which pass the tolerance threshold is beneficial, particularly where these mappings are more complex. One possible outcome is that introducing targeted instruction of such mappings during later stages of literacy teaching enables learners to take advantage of the underlying statistical regularities of text. As described in *Section 7.3.3.2*, a detailed investigation of the optimal number, level and order of taught spelling-sound mappings would be an important contribution to reading instruction policy and practice. The TP's assessment of consistency could be a valuable tool in this process, and the current findings offer a starting foundation on which to build.

Relatedly, the effect of the TP on reading behaviour during an extended teaching programme that builds spelling-sound knowledge in a gradual, cumulative fashion could be explored. This approach could be valuable as it would be more akin to children's experience of reading instruction delivered over time in the classroom. Thus, it could inform our understanding of the way in which children build a detailed system of orthography-phonology knowledge in real time.

Finally – and importantly – future research should examine other possible thresholds that predict an interaction between consistency at different orthographic levels. Even if it transpires that the TP is not the optimal account of word reading, this initial investigation has opened a new door for research on reading whereby a categorical threshold of consistency can be used to predict the productive use of smaller or larger orthographic units. Indeed, the approach applied here offers a powerful investigative mechanism: experimentally examining theoretical predictions which are generated according to a learner's specific input. Future research may be fruitful by developing new frameworks on a similar basis.

7.5 Conclusions

This thesis has presented research on generalisation within a quasi-regular domain which adds to our understanding of word reading, reading development and statistical learning. Namely, readers' generalisation of spelling-sound correspondences demonstrates a categorical effect of consistency which is predicted by Yang's (2016) tolerance threshold. Support for the recursive application of this threshold was offered by skilled readers using a familiar orthography, but less so from learners within an artificial orthography paradigm. Overall, the investigation revealed that readers' patterns of generalisation extend beyond the statistical distributions they have been exposed to during text experience. Therefore, it seems that readers carry out an active process to extract certain patterns from their input and use them productively to read aloud novel items, thereby adding structure to the variation in the input they have received. Additionally, this research provides the first evidence that the Tolerance Principle (Yang, 2016) can be usefully applied to domains beyond spoken language for which it was proposed. A notable consequence of these findings is that the Tolerance Principle may offer important insights about how humans extract and generalise information from their input that are applicable across cognitive domains.

Appendix A

Derivation of the Tolerance Principle

This overview summarises the derivation of the Tolerance Principle as laid out by Yang (2016, p. 60 – 66). See Schuler (unpublished, 2017) for a similar summary, Yang (2016) for a full elucidation and Yang (2018 p. 8 (a User's Guide)) for further details.

The derivation of the Tolerance Principle assumes that word frequency follows a Zipfian distribution (Zipf, 1949). Specifically, in a sample of N individual word types $\{w_1, w_2, \ldots w_N\}$, the rank (r_i) of a word (w_i) is inversely proportional to its frequency (f_i). Therefore, it is the case that r_i and f_i multiply to a constant C. This can be used to approximate the probability (p_i) with which (w_i) will occur, and can be expressed as:

$$\begin{aligned} p_i &= f_i / \sum_{k=1}^{N} f_k \\ &= \left(\frac{C}{r_i}\right) / \sum_{k=1}^{N} \frac{C}{r_k} \\ &= \frac{1}{iH_N} \text{ where } H_N = \sum_{k=1}^{N} \frac{1}{k} \end{aligned}$$

(the *N*th harmonic number)

According to the serial search mechanism, accessing the *r*th-ranked word in a list of *N* items will take *r* search steps. Therefore, the expected time to access a word that has been stored in a frequency-ranked list, *T(N,N)*, can be captured as:

$$T(N,N) = \sum_{r=1}^{N} r \frac{1}{rH_N} = \frac{N}{H_N}$$

Meanwhile, *T (N,e)* is the expected time to access the productive rule following a search of *e* exceptions ranked by frequency. This is the weighted average of the time it would take to search for an exception, and the time it would take to apply the rule, over the probability of occurrence of these two types of items. Specifically, the expected time to access an exception is *T(e,e)* or *e/He* (which is determined by the rank of the exception word in the list). The expected time to apply the rule to other, non-exception (*N* – *e*) items is *e*, i.e., the number of exceptions (as they must all be evaluated and rejected before the application of the rule). The overall average, *T (N,e)*, is given as:

$$T(N,e) = \frac{e}{N}T(e,e) + (1 - \frac{e}{N})e$$

$$= \frac{e}{N}\frac{e}{H_e} + (1 - \frac{e}{N})e$$

Together with Sam Gutmann, Yang derives the closed-form solution to the equation $T(N,N) = T(N,e)$. They begin by approximating the *N*th harmonic number, H_N, (which is found in the Zipfian assumption of word frequencies, as above) with the natural log of *N* (log*N*):

$$H_N \approx \ln N$$

To find $x = e / N$:

$$x\frac{e}{ln_e} + (1 - x)e = \frac{N}{lnN}$$

Dividing both sides of the equation by *N*:

$$x^2\frac{1}{lnN + lnX} + (1 - x)x = \frac{1}{lnN}$$

To allow:

$$f(x) = x^2\frac{1}{lnN + lnx} + (1 - x)x - \frac{1}{lnN}$$

Observing:

$$\begin{aligned} f(\frac{1}{lnN}) &= \frac{(1/lnN)^2}{lnN + lnlnN} + (1 - \frac{1}{lnN})\frac{1}{lnN} - \frac{1}{lnN} \\ &= -(\frac{1}{lnN})^2 + (\frac{1}{lnN})^3 \frac{lnN}{lnN + lnlnN} \end{aligned}$$

$$\approx -(\frac{1}{lnN})^2$$

$$\approx 0 \text{ for large values of } N$$

Thus deriving the tolerance threshold of exceptions for a productive rule:

$$e \leq \theta_N = N/ln(N)$$

Appendix B

Nonword Stimuli Used in Experiment 1

Nonword Item	Condition
DRAVE	1. Vowel winner, body winner, no conflict
TAVE	1. Vowel winner, body winner, no conflict
SCAVE	1. Vowel winner, body winner, no conflict
SMOOT	1. Vowel winner, body winner, no conflict
PROOT	1. Vowel winner, body winner, no conflict
YOOT	1. Vowel winner, body winner, no conflict
CREIL	1. Vowel winner, body winner, no conflict
CHEIL	1. Vowel winner, body winner, no conflict
THEIL	1. Vowel winner, body winner, no conflict
SMEAT	1. Vowel winner, body winner, no conflict
THEAT	1. Vowel winner, body winner, no conflict
PREAT	1. Vowel winner, body winner, no conflict
PLINT	1. Vowel winner, body winner, no conflict
TRINT	1. Vowel winner, body winner, no conflict
CHINT	1. Vowel winner, body winner, no conflict
SLEAM	1. Vowel winner, body winner, no conflict
YEAM	1. Vowel winner, body winner, no conflict
FREAM	1. Vowel winner, body winner, no conflict
FOVE	1. Vowel winner, body winner, no conflict
BROVE	1. Vowel winner, body winner, no conflict
NOVE	1. Vowel winner, body winner, no conflict
BROOL	1. Vowel winner, body winner, no conflict
VOOL	1. Vowel winner, body winner, no conflict
MOOL	1. Vowel winner, body winner, no conflict

RULL	1. Vowel winner, body winner, no conflict
SULL	1. Vowel winner, body winner, no conflict
TRULL	1. Vowel winner, body winner, no conflict
DRUSH	1. Vowel winner, body winner, no conflict
GLUSH	1. Vowel winner, body winner, no conflict
NUSH	1. Vowel winner, body winner, no conflict
TRINK	1. Vowel winner, body winner, no conflict
DINK	1. Vowel winner, body winner, no conflict
HINK	1. Vowel winner, body winner, no conflict
YAUNCH	1. Vowel winner, body winner, no conflict
DRAUNCH	1. Vowel winner, body winner, no conflict
MAUNCH	1. Vowel winner, body winner, no conflict
KAID	1. Vowel winner, body winner, no conflict
VAID	1. Vowel winner, body winner, no conflict
THAID	1. Vowel winner, body winner, no conflict
LOAP	1. Vowel winner, body winner, no conflict
FROAP	1. Vowel winner, body winner, no conflict
BOAP	1. Vowel winner, body winner, no conflict
VORN	1. Vowel winner, body winner, no conflict
JORN	1. Vowel winner, body winner, no conflict
ZORN	1. Vowel winner, body winner, no conflict
SHORM	1. Vowel winner, body winner, no conflict
ZORM	1. Vowel winner, body winner, no conflict
BORM	1. Vowel winner, body winner, no conflict
YEIN	1. Vowel winner, body winner, no conflict
GLEIN	1. Vowel winner, body winner, no conflict
FLEIN	1. Vowel winner, body winner, no conflict
PEIGHT	1. Vowel winner, body winner, no conflict

VEIGHT	1. Vowel winner, body winner, no conflict
DREIGHT	1. Vowel winner, body winner, no conflict
VARN	1. Vowel winner, body winner, no conflict
PARN	1. Vowel winner, body winner, no conflict
BLARN	1. Vowel winner, body winner, no conflict
YAUT	1. Vowel winner, body winner, no conflict
JAUT	1. Vowel winner, body winner, no conflict
PRAUT	1. Vowel winner, body winner, no conflict
LIND	2. Vowel winner, body winner, conflict
YIND	2. Vowel winner, body winner, conflict
TRIND	2. Vowel winner, body winner, conflict
GLEAD	2. Vowel winner, body winner, conflict
SMEAD	2. Vowel winner, body winner, conflict
VEAD	2. Vowel winner, body winner, conflict
MOOK	2. Vowel winner, body winner, conflict
DOOK	2. Vowel winner, body winner, conflict
PLOOK	2. Vowel winner, body winner, conflict
BREALT	2. Vowel winner, body winner, conflict
CHEALT	2. Vowel winner, body winner, conflict
GREALT	2. Vowel winner, body winner, conflict
GREAMT	2. Vowel winner, body winner, conflict
BLEAMT	2. Vowel winner, body winner, conflict
PEAMT	2. Vowel winner, body winner, conflict
DEAPT	2. Vowel winner, body winner, conflict
VEAPT	2. Vowel winner, body winner, conflict
FREAPT	2. Vowel winner, body winner, conflict
PLUTH	2. Vowel winner, body winner, conflict
NUTH	2. Vowel winner, body winner, conflict

GUTH	2. Vowel winner, body winner, conflict
PLEANT	2. Vowel winner, body winner, conflict
GEANT	2. Vowel winner, body winner, conflict
HEANT	2. Vowel winner, body winner, conflict
FEALM	2. Vowel winner, body winner, conflict
PEALM	2. Vowel winner, body winner, conflict
TREALM	2. Vowel winner, body winner, conflict
SHEALTH	2. Vowel winner, body winner, conflict
PEALTH	2. Vowel winner, body winner, conflict
TREALTH	2. Vowel winner, body winner, conflict
NOUCH	3. Vowel all fail, body winner
SOUCH	3. Vowel all fail, body winner
FOUCH	3. Vowel all fail, body winner
LOUNT	3. Vowel all fail, body winner
BROUNT	3. Vowel all fail, body winner
PLOUNT	3. Vowel all fail, body winner
VOUST	3. Vowel all fail, body winner
NOUST	3. Vowel all fail, body winner
TROUST	3. Vowel all fail, body winner
NOWL	3. Vowel all fail, body winner
BROWL	3. Vowel all fail, body winner
CHOWL	3. Vowel all fail, body winner
MIEF	3. Vowel all fail, body winner
HIEF	3. Vowel all fail, body winner
PRIEF	3. Vowel all fail, body winner
FIEK	3. Vowel all fail, body winner
JIEK	3. Vowel all fail, body winner
DRIEK	3. Vowel all fail, body winner

BIELD	3. Vowel all fail, body winner
ZIELD	3. Vowel all fail, body winner
PRIELD	3. Vowel all fail, body winner
PLOUND	3. Vowel all fail, body winner
NOUND	3. Vowel all fail, body winner
VOUND	3. Vowel all fail, body winner
ZOWD	3. Vowel all fail, body winner
FOWD	3. Vowel all fail, body winner
TROWD	3. Vowel all fail, body winner
JOUT	3. Vowel all fail, body winner
PROUT	3. Vowel all fail, body winner
ZOUT	3. Vowel all fail, body winner
NUILD	3. Vowel all fail, body winner
ZUILD	3. Vowel all fail, body winner
TUILD	3. Vowel all fail, body winner
THOCK	3. Vowel all fail, body winner
PLOCK	3. Vowel all fail, body winner
GROCK	3. Vowel all fail, body winner
NOTH	3. Vowel all fail, body winner
JOTH	3. Vowel all fail, body winner
FLOTH	3. Vowel all fail, body winner
CROLD	3. Vowel all fail, body winner
VOLD	3. Vowel all fail, body winner
BROLD	3. Vowel all fail, body winner
DRON	3. Vowel all fail, body winner
PON	3. Vowel all fail, body winner
BLON	3. Vowel all fail, body winner
JEART	3. Vowel all fail, body winner

VEART	3. Vowel all fail, body winner
ZEART	3. Vowel all fail, body winner
BOUP	3. Vowel all fail, body winner
CHOUP	3. Vowel all fail, body winner
FROUP	3. Vowel all fail, body winner
TROW	4. Vowel all fail, body all fail
DOW	4. Vowel all fail, body all fail
FROW	4. Vowel all fail, body all fail
GLOWN	4. Vowel all fail, body all fail
KOWN	4. Vowel all fail, body all fail
YOWN	4. Vowel all fail, body all fail
JOUGH	4. Vowel all fail, body all fail
PROUGH	4. Vowel all fail, body all fail
DROUGH	4. Vowel all fail, body all fail
JOLL	4. Vowel all fail, body all fail
CHOLL	4. Vowel all fail, body all fail
GROLL	4. Vowel all fail, body all fail
TROOD	5. Vowel winner, body all fail
NOOD	5. Vowel winner, body all fail
GROOD	5. Vowel winner, body all fail
PLEARD	6. Vowel all fail, body all pass
MEARD	6. Vowel all fail, body all pass
ZEARD	6. Vowel all fail, body all pass
TIEND	6. Vowel all fail, body all pass
VIEND	6. Vowel all fail, body all pass
JIEND	6. Vowel all fail, body all pass
SHOUTH	6. Vowel all fail, body all pass
HOUTH	6. Vowel all fail, body all pass

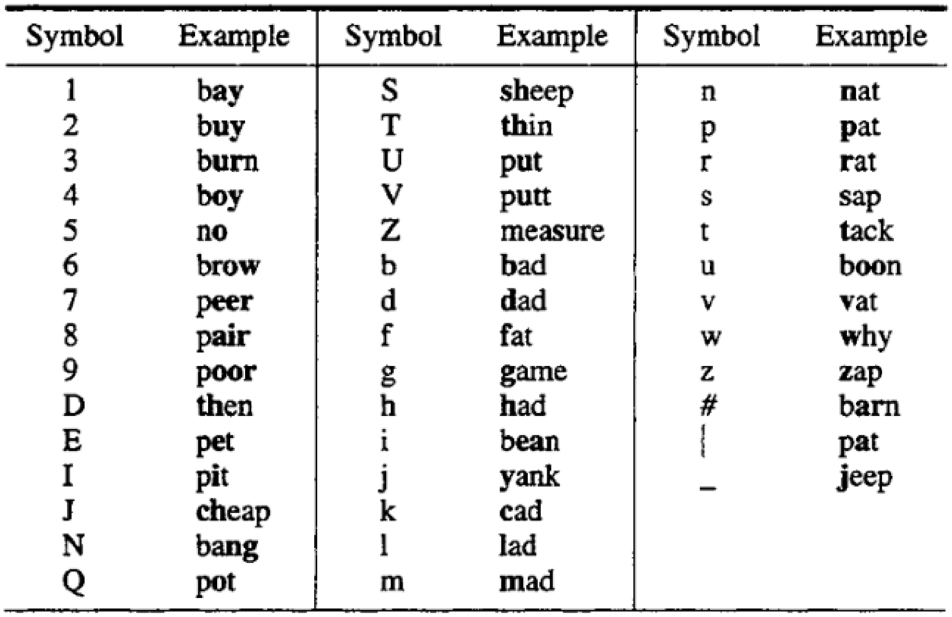
FOUTH	6. Vowel all fail, body all pass
NOULD	6. Vowel all fail, body all pass
VOULD	6. Vowel all fail, body all pass
JOULD	6. Vowel all fail, body all pass
BROUL	6. Vowel all fail, body all pass
CHOUL	6. Vowel all fail, body all pass
MOUL	6. Vowel all fail, body all pass
LEARTH	6. Vowel all fail, body all pass
KEARTH	6. Vowel all fail, body all pass
NEARTH	6. Vowel all fail, body all pass
SONT	6. Vowel all fail, body all pass
BONT	6. Vowel all fail, body all pass
RONT	6. Vowel all fail, body all pass
THOMB	6. Vowel all fail, body all pass
POMB	6. Vowel all fail, body all pass
CHOMB	6. Vowel all fail, body all pass
TROLF	6. Vowel all fail, body all pass
HOLF	6. Vowel all fail, body all pass
VOLF	6. Vowel all fail, body all pass
CLOST	6. Vowel all fail, body all pass
FOST	6. Vowel all fail, body all pass
SOST	6. Vowel all fail, body all pass
KIMB	7. Vowel winner, body all pass
NIMB	7. Vowel winner, body all pass
FRIMB	7. Vowel winner, body all pass
SARCE	7. Vowel winner, body all pass
FLARCE	7. Vowel winner, body all pass
DARCE	7. Vowel winner, body all pass

PEAF	7. Vowel winner, body all pass
SEAF	7. Vowel winner, body all pass
GLEAF	7. Vowel winner, body all pass
THILD	7. Vowel winner, body all pass
SILD	7. Vowel winner, body all pass
PRILD	7. Vowel winner, body all pass

Appendix C

The Phonemic Vocabulary of the Dual-Route Cascaded Model

(Coltheart & Rastle, 1999, p. 498)



Symbol	Example	Symbol	Example	Symbol	Example
1	bay	S	sheep	n	nat
2	buy	T	thin	p	pat
3	burn	U	put	r	rat
4	boy	V	putt	s	sap
5	no	Z	measure	t	tack
6	brow	b	bad	u	boon
7	peer	d	dad	v	vat
8	pair	f	fat	w	why
9	poor	g	game	z	zap
D	then	h	had	#	barn
E	pet	i	bean	{	pat
I	pit	j	yank	_	jeep
J	cheap	k	cad		
N	bang	l	lad		
Q	pot	m	mad		

Appendix D

R Studio Output for Mixed-Effects Models by Chapter

Chapter 3: Generalisation of orthography-phonology correspondences in nonword reading by adults and children

```
Vowel_winner$Group <- factor(Vowel_winner$Group, levels=c("adult", "child"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Vowel_Winner_score ~ Condition * Group + (1 | Participant) +      (1 | Item)
   Data: Vowel_winner

     AIC      BIC   logLik deviance df.resid
  4992.6   5058.8  -2486.3   4972.6     5517

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.7025 -0.3308  0.2849  0.4890  4.1974

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 1.9220   1.3864
 Participant (Intercept) 0.2796   0.5288
Number of obs: 5527, groups:  Item, 105; Participant, 53

Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)
 (Intercept)           1.77397    0.22688   7.819 5.33e-15 ***
 Condition2           -1.52688    0.33272  -4.589 4.45e-06 ***
 Condition5            1.57982    1.00872   1.566   0.1173
 Condition7            0.60199    0.48984   1.229   0.2191
 Groupchild           -0.01651    0.18036  -0.092   0.9271
 Condition2:Groupchild  0.66963    0.15533   4.311 1.62e-05 ***
 Condition5:Groupchild -1.44475    0.66037  -2.188   0.0287 *
 Condition7:Groupchild -0.39953    0.26258  -1.522   0.1281
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Vowel_Winner_score ~ Condition * Group + (1 | Participant) +      (1 | Item)
   Data: Vowel_winner

     AIC      BIC   logLik deviance df.resid
  4992.6   5058.8  -2486.3   4972.6     5517

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.7026 -0.3308  0.2849  0.4890  4.1975

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 1.9221   1.3864
 Participant (Intercept) 0.2796   0.5288
Number of obs: 5527, groups:  Item, 105; Participant, 53

  Fixed effects:
                        Estimate Std. Error z value Pr(>|z|)
  (Intercept)             0.2471     0.2874   0.860  0.38992
  Condition1              1.5269     0.3328   4.588 4.47e-06 ***
  Condition5              3.1067     1.0257   3.029  0.00246 **
  Condition7              2.1289     0.5215   4.083 4.45e-05 ***
  Groupchild              0.6531     0.1851   3.527  0.00042 ***
  Condition1:Groupchild  -0.6696     0.1553  -4.311 1.63e-05 ***
  Condition5:Groupchild  -2.1144     0.6622  -3.193  0.00141 **
  Condition7:Groupchild  -1.0692     0.2665  -4.013 6.01e-05 ***
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Vowel_Winner_score ~ Condition * Group + (1 | Participant) +      (1 | Item)
   Data: Vowel_winner

     AIC      BIC   logLik deviance df.resid
  4992.6   5058.8  -2486.3   4972.6     5517

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.7026 -0.3308  0.2849  0.4890  4.1975

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 1.9221   1.3864
 Participant (Intercept) 0.2796   0.5288
Number of obs: 5527, groups:  Item, 105; Participant, 53

 Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)
 (Intercept)             2.3760     0.4610   5.154 2.55e-07 ***
 Condition2             -2.1289     0.5213  -4.084 4.43e-05 ***
 Condition1             -0.6020     0.4897  -1.229    0.219
 Condition5              0.9778     1.0846   0.902    0.367
 Groupchild             -0.4161     0.2818  -1.477    0.140
 Condition2:Groupchild   1.0692     0.2664   4.013 5.99e-05 ***
 Condition1:Groupchild   0.3995     0.2625   1.522    0.128
 Condition5:Groupchild  -1.0452     0.6942  -1.506    0.132
```

```
Vowel_winner$Group <- factor(Vowel_winner$Group, levels=c("child", "adult"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Vowel_Winner_score ~ Condition * Group + (1 | Participant) +      (1 | Item)
   Data: Vowel_winner

     AIC      BIC   logLik deviance df.resid
  4992.6   5058.8  -2486.3   4972.6     5517

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.7026 -0.3308  0.2849  0.4890  4.1975

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 1.9220   1.3864
 Participant (Intercept) 0.2796   0.5288
Number of obs: 5527, groups:  Item, 105; Participant, 53

Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)
 (Intercept)            1.7574     0.2199   7.991 1.34e-15 ***
 Condition2            -0.8572     0.3302  -2.596  0.00943 **
 Condition5             0.1352     0.8800   0.154  0.87793
 Condition7             0.2025     0.4754   0.426  0.67007
 Groupadult             0.0165     0.1803   0.092  0.92708
 Condition2:Groupadult  -0.6696     0.1553  -4.312 1.62e-05 ***
 Condition5:Groupadult   1.4448     0.6598   2.190  0.02855 *
 Condition7:Groupadult   0.3996     0.2625   1.522  0.12803
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Vowel_Winner_score ~ Condition * Group + (1 | Participant) +      (1 | Item)
   Data: Vowel_winner

     AIC      BIC   logLik deviance df.resid
  4992.6   5058.8  -2486.3   4972.6     5517

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.7027 -0.3308  0.2849  0.4890  4.1975

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 1.9219   1.3863
 Participant (Intercept) 0.2796   0.5288
Number of obs: 5527, groups:  Item, 105; Participant, 53

  Fixed effects:
                        Estimate Std. Error z value Pr(>|z|)
  (Intercept)             0.9003     0.2833   3.178 0.001485 **
  Condition1              0.8573     0.3303   2.595 0.009452 **
  Condition5              0.9924     0.8986   1.104 0.269421
  Condition7              1.0599     0.5086   2.084 0.037151 *
  Groupadult             -0.6531     0.1851  -3.528 0.000419 ***
  Condition1:Groupadult   0.6695     0.1553   4.311 1.63e-05 ***
  Condition5:Groupadult   2.1145     0.6618   3.195 0.001398 **
  Condition7:Groupadult   1.0692     0.2664   4.013 6.00e-05 ***
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Vowel_Winner_score ~ Condition * Group + (1 | Participant) +      (1 | Item)
   Data: Vowel_winner

     AIC      BIC   logLik deviance df.resid
  4992.6   5058.8  -2486.3   4972.6     5517

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.7026 -0.3308  0.2849  0.4890  4.1975

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 1.9220   1.3864
 Participant (Intercept) 0.2796   0.5288
Number of obs: 5527, groups:  Item, 105; Participant, 53

  Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
  (Intercept)           1.95994    0.44475   4.407 1.05e-05 ***
  Condition2           -1.05968    0.50844  -2.084   0.0371 *
  Condition1           -0.20247    0.47556  -0.426   0.6703
  Condition5           -0.06727    0.96117  -0.070   0.9442
  Groupadult            0.41605    0.28183   1.476   0.1399
  Condition2:Groupadult -1.06916    0.26647  -4.012 6.01e-05 ***
  Condition1:Groupadult -0.39954    0.26259  -1.522   0.1281
  Condition5:Groupadult  1.04530    0.69492   1.504   0.1325
```

```
Body_winner$Group <- factor(Body_winner$Group, levels=c("adult", "child"))

Body_winner$Condition <- factor(Body_winner$Condition, levels=c("2", "3"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Body_Winner_score ~ Condition + Group + (1 + Condition | Participant) +      (1 |
Item)
   Data: Body_winner

     AIC      BIC   logLik deviance df.resid
  4380.1   4424.6  -2183.1   4366.1     4257

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2738 -0.5520  0.2081  0.5600  5.5109

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 Item        (Intercept) 1.0707   1.0348
 Participant (Intercept) 1.0233   1.0116
             Condition3  0.9798   0.9898   -0.73
Number of obs: 4264, groups:  Item, 81; Participant, 53

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -1.2111     0.2708  -4.472 7.73e-06 ***
  Condition3    2.5575     0.2897   8.828  < 2e-16 ***
  Groupchild   -0.4430     0.2080  -2.130   0.0332 *
```

```
Body_winner$Condition <- factor(Body_winner$Condition, levels=c("3", "2"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Body_Winner_score ~ Condition + Group + (1 + Condition | Participant) +      (1 |
Item)
   Data: Body_winner

     AIC      BIC   logLik deviance df.resid
  4380.1   4424.6  -2183.1   4366.1     4257

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2738 -0.5520  0.2081  0.5600  5.5109

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 Item        (Intercept) 1.0708   1.0348
 Participant (Intercept) 0.5432   0.7370
             Condition2  0.9798   0.9899   -0.34
Number of obs: 4264, groups:  Item, 81; Participant, 53

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)   1.3463     0.2181   6.172 6.73e-10 ***
  Condition2   -2.5575     0.2897  -8.828  < 2e-16 ***
  Groupchild   -0.4429     0.2080  -2.129   0.0332 *
```

```
Body_winner$Group <- factor(Body_winner$Group, levels=c("child", "adult"))
Body_winner$Condition <- factor(Body_winner$Condition, levels=c("2", "3"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Body_Winner_score ~ Condition + Group + (1 + Condition | Participant) +      (1 |
Item)
   Data: Body_winner

     AIC      BIC   logLik deviance df.resid
  4380.1   4424.6  -2183.1   4366.1     4257

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2738 -0.5520  0.2081  0.5600  5.5109

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 Item        (Intercept) 1.0707   1.0348
 Participant (Intercept) 1.0233   1.0116
             Condition3  0.9798   0.9898   -0.73
Number of obs: 4264, groups:  Item, 81; Participant, 53

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -1.6541     0.2641  -6.264 3.76e-10 ***
  Condition3    2.5575     0.2897   8.828  < 2e-16 ***
  Groupadult    0.4430     0.2080   2.130   0.0332 *
```

```
Body_winner$Condition <- factor(Body_winner$Condition, levels=c("3", "2"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Body_Winner_score ~ Condition + Group + (1 + Condition | Participant) +      (1 |
Item)
   Data: Body_winner

     AIC      BIC   logLik deviance df.resid
  4380.1   4424.6  -2183.1   4366.1     4257

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2738 -0.5520  0.2081  0.5600  5.5109

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 Item        (Intercept) 1.0707   1.0348
 Participant (Intercept) 0.5432   0.7370
             Condition2  0.9798   0.9899   -0.34
Number of obs: 4264, groups:  Item, 81; Participant, 53

 Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)   0.9034     0.2067   4.370 1.24e-05 ***
 Condition2   -2.5575     0.2897  -8.829  < 2e-16 ***
 Groupadult    0.4430     0.2080   2.130   0.0332 *
```

```
Ad_ch_models$Group <- factor(Ad_ch_models$Group, levels=c("Adult", "Child"))
Ad_ch_models$Model <- factor(Ad_ch_models$Model, levels=c("TP", "CDP", "DRC",
"Triangle"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Model * Group + (1 | Participant) + (1 | Item)
   Data: Ad_ch_models

     AIC      BIC   logLik deviance df.resid
 29033.9  29116.6 -14506.9  29013.9    29044

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.1219  -0.6759   0.2867   0.6041   5.9411

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 2.4500   1.5652
 Participant (Intercept) 0.1765   0.4201
Number of obs: 29054, groups:  Item, 138; Participant, 53

 Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
 (Intercept)                1.369011   0.165414   8.276  < 2e-16 ***
 ModelCDP                  -0.516212   0.061985  -8.328  < 2e-16 ***
 ModelDRC                  -0.452589   0.062074  -7.291 3.08e-13 ***
 ModelTriangle             -0.671401   0.061751 -10.873  < 2e-16 ***
 GroupChild                 0.086926   0.131160   0.663    0.507
 ModelCDP:GroupChild       -0.119097   0.084137  -1.416    0.157
 ModelDRC:GroupChild       -0.008476   0.084456  -0.100    0.920
 ModelTriangle:GroupChild  -0.248362   0.083681  -2.968    0.003 **
```

```
Ad_ch_models$Model <- factor(Ad_ch_models$Model, levels=c("CDP", "TP", "DRC",
"Triangle"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Model * Group + (1 | Participant) + (1 | Item)
   Data: Ad_ch_models

     AIC      BIC   logLik deviance df.resid
 29033.9  29116.6 -14506.9  29013.9    29044

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.1219  -0.6759   0.2867   0.6041   5.9411

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 2.4499   1.5652
 Participant (Intercept) 0.1765   0.4201
Number of obs: 29054, groups:  Item, 138; Participant, 53

  Fixed effects:
                             Estimate Std. Error z value Pr(>|z|)
  (Intercept)                 0.85279    0.16460   5.181 2.21e-07 ***
  ModelTP                     0.51621    0.06198   8.329  < 2e-16 ***
  ModelDRC                    0.06363    0.06034   1.054  0.29169
  ModelTriangle              -0.15518    0.05994  -2.589  0.00963 **
  GroupChild                 -0.03217    0.12948  -0.248  0.80379
  ModelTP:GroupChild          0.11910    0.08413   1.416  0.15686
  ModelDRC:GroupChild         0.11062    0.08192   1.350  0.17695
  ModelTriangle:GroupChild   -0.12928    0.08110  -1.594  0.11092
```

```
Ad_ch_models$Model <- factor(Ad_ch_models$Model, levels=c("DRC", "CDP", "TP",
"Triangle"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Model * Group + (1 | Participant) + (1 | Item)
   Data: Ad_ch_models

     AIC      BIC   logLik deviance df.resid
 29033.9  29116.6 -14506.9  29013.9    29044

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.1220  -0.6759   0.2867   0.6041   5.9411

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 2.4500   1.5652
 Participant (Intercept) 0.1765   0.4201
Number of obs: 29054, groups:  Item, 138; Participant, 53

 Fixed effects:
                           Estimate Std. Error z value Pr(>|z|)
 (Intercept)               0.916424   0.164799   5.561 2.68e-08 ***
 ModelCDP                 -0.063620   0.060356  -1.054  0.29185
 ModelTP                   0.452585   0.062078   7.291 3.09e-13 ***
 ModelTriangle            -0.218812   0.060074  -3.642  0.00027 ***
 GroupChild                0.078439   0.129740   0.605  0.54546
 ModelCDP:GroupChild      -0.110621   0.081938  -1.350  0.17700
 ModelTP:GroupChild        0.008482   0.084458   0.100  0.92000
 ModelTriangle:GroupChild -0.239887   0.081465  -2.945  0.00323 **
```

```
Ad_ch_models$Model <- factor(Ad_ch_models$Model, levels=c("Triangle", "CDP", "DRC",
"TP"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Model * Group + (1 | Participant) + (1 | Item)
   Data: Ad_ch_models

     AIC      BIC   logLik deviance df.resid
 29033.9  29116.6 -14506.9  29013.9    29044

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.1219  -0.6759   0.2867   0.6041   5.9412

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 2.4500   1.5652
 Participant (Intercept) 0.1765   0.4201
Number of obs: 29054, groups:  Item, 138; Participant, 53

  Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
  (Intercept)          0.69762    0.16455   4.240 2.24e-05 ***
  ModelCDP             0.15519    0.05994   2.589 0.009624 **
  ModelDRC             0.21880    0.06006   3.643 0.000269 ***
  ModelTP              0.67140    0.06175  10.873  < 2e-16 ***
  GroupChild          -0.16145    0.12919  -1.250 0.211417
  ModelCDP:GroupChild  0.12927    0.08109   1.594 0.110934
  ModelDRC:GroupChild  0.23991    0.08144   2.946 0.003223 **
  ModelTP:GroupChild   0.24837    0.08367   2.968 0.002995 **
```

```
Ad_ch_models$Group <- factor(Ad_ch_models$Group, levels=c("Child", "Adult"))
Ad_ch_models$Model <- factor(Ad_ch_models$Model, levels=c("TP", "CDP", "DRC",
"Triangle"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Model * Group + (1 | Participant) + (1 | Item)
   Data: Ad_ch_models

     AIC      BIC   logLik deviance df.resid
 29033.9  29116.6 -14506.9  29013.9    29044

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.1220  -0.6759   0.2867   0.6041   5.9411

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 2.4500   1.5652
 Participant (Intercept) 0.1765   0.4201
Number of obs: 29054, groups:  Item, 138; Participant, 53

  Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
  (Intercept)               1.455926   0.160728   9.058  < 2e-16 ***
  ModelCDP                 -0.635305   0.057087 -11.129  < 2e-16 ***
  ModelDRC                 -0.461066   0.057387  -8.034 9.41e-16 ***
  ModelTriangle            -0.919765   0.056845 -16.180  < 2e-16 ***
  GroupAdult               -0.086940   0.131154  -0.663    0.507
  ModelCDP:GroupAdult       0.119101   0.084143   1.415    0.157
  ModelDRC:GroupAdult       0.008479   0.084456   0.100    0.920
  ModelTriangle:GroupAdult  0.248369   0.083684   2.968    0.003 **
```

```
Ad_ch_models$Model <- factor(Ad_ch_models$Model, levels=c("CDP", "TP", "DRC",
"Triangle"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Model * Group + (1 | Participant) + (1 | Item)
   Data: Ad_ch_models

     AIC      BIC   logLik deviance df.resid
 29033.9  29116.6 -14506.9  29013.9    29044

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.1220  -0.6759   0.2867   0.6041   5.9412

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 2.4500   1.5652
 Participant (Intercept) 0.1765   0.4201
Number of obs: 29054, groups:  Item, 138; Participant, 53

 Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
 (Intercept)                 0.82056    0.15984   5.134 2.84e-07 ***
 ModelTP                     0.63530    0.05708  11.130  < 2e-16 ***
 ModelDRC                    0.17424    0.05542   3.144  0.00167 **
 ModelTriangle              -0.28446    0.05465  -5.205 1.94e-07 ***
 GroupAdult                  0.03227    0.12950   0.249  0.80319
 ModelTP:GroupAdult         -0.11911    0.08412  -1.416  0.15680
 ModelDRC:GroupAdult        -0.11065    0.08193  -1.351  0.17683
 ModelTriangle:GroupAdult    0.12923    0.08109   1.594  0.11103
```

```
Ad_ch_models$Model <- factor(Ad_ch_models$Model, levels=c("DRC", "CDP", "TP",
"Triangle"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Model * Group + (1 | Participant) + (1 | Item)
   Data: Ad_ch_models

     AIC      BIC   logLik deviance df.resid
 29033.9  29116.6 -14506.9  29013.9    29044

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.1219  -0.6759   0.2867   0.6041   5.9411

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 2.4500   1.5652
 Participant (Intercept) 0.1765   0.4201
Number of obs: 29054, groups:  Item, 138; Participant, 53

 Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
 (Intercept)                0.994870   0.159930   6.221 4.95e-10 ***
 ModelCDP                  -0.174240   0.055419  -3.144  0.00167 **
 ModelTP                    0.461066   0.057382   8.035 9.36e-16 ***
 ModelTriangle             -0.458698   0.055079  -8.328  < 2e-16 ***
 GroupAdult                -0.078457   0.129710  -0.605  0.54527
 ModelCDP:GroupAdult        0.110625   0.081931   1.350  0.17694
 ModelTP:GroupAdult        -0.008471   0.084449  -0.100  0.92010
 ModelTriangle:GroupAdult   0.239892   0.081451   2.945  0.00323 **
```

```
Ad_ch_models$Model <- factor(Ad_ch_models$Model, levels=c("Triangle", "CDP", "DRC",
"TP"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Model * Group + (1 | Participant) + (1 | Item)
   Data: Ad_ch_models

     AIC      BIC   logLik deviance df.resid
 29033.9  29116.6 -14506.9  29013.9    29044

Scaled residuals:
     Min       1Q   Median       3Q      Max
-10.1220  -0.6759   0.2867   0.6041   5.9412

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 2.4500   1.5652
 Participant (Intercept) 0.1765   0.4201
Number of obs: 29054, groups:  Item, 138; Participant, 53

  Fixed effects:
                       Estimate Std. Error z value Pr(>|z|)
  (Intercept)           0.53616    0.15950   3.361 0.000775 ***
  ModelCDP              0.28446    0.05464   5.206 1.93e-07 ***
  ModelDRC              0.45870    0.05507   8.329  < 2e-16 ***
  ModelTP               0.91977    0.05683  16.185  < 2e-16 ***
  GroupAdult            0.16144    0.12913   1.250 0.211242
  ModelCDP:GroupAdult -0.12927    0.08108  -1.594 0.110840
  ModelDRC:GroupAdult -0.23989    0.08143  -2.946 0.003219 **
  ModelTP:GroupAdult  -0.24837    0.08365  -2.969 0.002986 **
```

```
Models_vpass$Group <- factor(Models_vpass$Group, levels=c("Adult", "Child"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Model * Group + (1 | Participant) + (1 + Group | Item)
   Data: Models_vpass

     AIC      BIC   logLik deviance df.resid
 18017.5  18111.7  -8996.8  17993.5    18947

Scaled residuals:
    Min      1Q  Median      3Q     Max
-8.6906 -0.6278  0.2343  0.5611  6.8578

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 Item        (Intercept) 3.4940   1.8692
             GroupChild  0.6477   0.8048   -0.38
 Participant (Intercept) 0.1705   0.4129
Number of obs: 18959, groups:  Item, 90; Participant, 53

 Fixed effects:
                     Estimate Std. Error z value Pr(>|z|)
 (Intercept)          0.57611    0.22274   2.587 0.009695 **
 ModelCDP             0.31873    0.07660   4.161 3.17e-05 ***
 ModelDRC             0.73078    0.07793   9.378  < 2e-16 ***
 ModelTP              0.80759    0.07833  10.310  < 2e-16 ***
 GroupChild          -0.19732    0.16213  -1.217 0.223590
 ModelCDP:GroupChild  0.17605    0.10323   1.705 0.088117 .
 ModelDRC:GroupChild  0.36469    0.10617   3.435 0.000592 ***
 ModelTP:GroupChild   0.42540    0.10712   3.971 7.15e-05 ***
```

```
Models_vpass$Group <- factor(Models_vpass$Group, levels=c("Child", "Adult"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Model * Group + (1 | Participant) + (1 + Group | Item)
   Data: Models_vpass

     AIC      BIC   logLik deviance df.resid
 18017.5  18111.7  -8996.8  17993.5    18947

Scaled residuals:
    Min      1Q  Median      3Q     Max
-8.6898 -0.6278  0.2343  0.5612  6.8577

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 Item        (Intercept) 2.9997   1.7320
             GroupAdult  0.6478   0.8049   -0.06
 Participant (Intercept) 0.1705   0.4130
Number of obs: 18959, groups:  Item, 90; Participant, 53

 Fixed effects:
                     Estimate Std. Error z value Pr(>|z|)
 (Intercept)          0.37915    0.20479   1.851  0.06412 .
 ModelCDP             0.49474    0.06920   7.149 8.71e-13 ***
 ModelDRC             1.09553    0.07212  15.189  < 2e-16 ***
 ModelTP              1.23279    0.07309  16.866  < 2e-16 ***
 GroupAdult           0.19707    0.16214   1.215  0.22420
 ModelCDP:GroupAdult -0.17608    0.10323  -1.706  0.08805 .
 ModelDRC:GroupAdult -0.36480    0.10617  -3.436  0.00059 ***
 ModelTP:GroupAdult  -0.42536    0.10712  -3.971 7.17e-05 ***
```

```
Models_vfail$Group <- factor(Models_vfail$Group, levels=c("Adult", "Child"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Model + Group + (1 | Participant) + (1 + Group | Item)
   Data: Models_vfail

     AIC      BIC   logLik deviance df.resid
 10129.4  10194.4  -5055.7  10111.4    10086

Scaled residuals:
    Min      1Q  Median      3Q     Max
-8.0087 -0.6221  0.2918  0.5844  5.3573

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 Participant (Intercept) 0.4926   0.7019
 Item        (Intercept) 2.6254   1.6203
             GroupChild  0.4510   0.6715   -0.48
Number of obs: 10095, groups:  Participant, 53; Item, 48

 Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)  1.18055    0.28240   4.180 2.91e-05 ***
 ModelCDP    -0.09633    0.06998  -1.377    0.169
 ModelDRC    -0.61778    0.06926  -8.920  < 2e-16 ***
 ModelTP      0.48038    0.07254   6.622 3.53e-11 ***
 GroupChild  -0.36388    0.22520  -1.616    0.106
```

```
Models_vfail$Group <- factor(Models_vfail$Group, levels=c("Child", "Adult"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Model + Group + (1 | Participant) + (1 + Group | Item)
   Data: Models_vfail

     AIC      BIC   logLik deviance df.resid
 10129.4  10194.4  -5055.7  10111.4    10086

Scaled residuals:
    Min      1Q  Median      3Q     Max
-8.0086 -0.6221  0.2918  0.5844  5.3573

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 Participant (Intercept) 0.4927   0.7019
 Item        (Intercept) 2.0415   1.4288
             GroupAdult  0.4510   0.6715   0.07
Number of obs: 10095, groups:  Participant, 53; Item, 48

 Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)  0.81667    0.25090   3.255  0.00113 **
 ModelCDP    -0.09633    0.06998  -1.376  0.16869
 ModelDRC    -0.61778    0.06926  -8.920  < 2e-16 ***
 ModelTP      0.48038    0.07254   6.622 3.54e-11 ***
 GroupAdult   0.36385    0.22528   1.615  0.10628
```

```
Ad_ch_consistency$Group <- factor(Ad_ch_consistency$Group, levels=c("Adult", "Child"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Vowel_reg_score ~ Vowel_H_type + Vowel_H_token + Group + (1 |      Participant) +
(1 | Item)
   Data: Ad_ch_consistency

     AIC      BIC   logLik deviance df.resid
  7668.0   7710.1  -3828.0   7656.0     8200

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.8264 -0.4255  0.2956  0.5132  4.4454

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 2.0679   1.4380
 Participant (Intercept) 0.2597   0.5096
Number of obs: 8206, groups:  Item, 156; Participant, 53

  Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
  (Intercept)    2.53798    0.30513   8.318  < 2e-16 ***
  Vowel_H_type  -1.18527    0.43045  -2.754  0.00589 **
  Vowel_H_token -0.38625    0.36393  -1.061  0.28854
  GroupChild     0.07855    0.15249   0.515  0.60647
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Vowel_reg_score ~ Vowel_H_type + Vowel_H_token + Group + TP_vowel_reg +
    (1 | Participant) + (1 | Item)
   Data: Ad_ch_consistency

     AIC      BIC   logLik deviance df.resid
  7633.7   7682.8  -3809.9   7619.7     8199

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.3561 -0.4114  0.2988  0.5093  4.8129

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 1.5398   1.2409
 Participant (Intercept) 0.2598   0.5097
Number of obs: 8206, groups:  Item, 156; Participant, 53

 Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
 (Intercept)      0.72955    0.38471   1.896   0.0579 .
 Vowel_H_type    -0.07893    0.41123  -0.192   0.8478
 Vowel_H_token   -0.66709    0.32097  -2.078   0.0377 *
 GroupChild       0.07909    0.15253   0.519   0.6041
 TP_vowel_reg2-1  2.84272    0.44123   6.443 1.17e-10 ***
```

```
Ad_ch_consistency$Group <- factor(Ad_ch_consistency$Group, levels=c("Child", "Adult"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Vowel_reg_score ~ Vowel_H_type + Vowel_H_token + Group + (1 |      Participant) +
(1 | Item)
   Data: Ad_ch_consistency

     AIC      BIC   logLik deviance df.resid
  7668.0   7710.1  -3828.0   7656.0     8200

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.8264 -0.4255  0.2956  0.5132  4.4454

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 2.0680   1.4380
 Participant (Intercept) 0.2597   0.5096
Number of obs: 8206, groups:  Item, 156; Participant, 53

  Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
  (Intercept)    2.61652    0.30172   8.672  < 2e-16 ***
  Vowel_H_type  -1.18523    0.43034  -2.754  0.00588 **
  Vowel_H_token -0.38628    0.36393  -1.061  0.28850
  GroupAdult    -0.07854    0.15249  -0.515  0.60651
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Vowel_reg_score ~ Vowel_H_type + Vowel_H_token + Group + TP_vowel_reg +
    (1 | Participant) + (1 | Item)
   Data: Ad_ch_consistency

     AIC      BIC   logLik deviance df.resid
  7633.7   7682.8  -3809.9   7619.7     8199

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.3561 -0.4114  0.2988  0.5093  4.8129

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 1.5398   1.2409
 Participant (Intercept) 0.2597   0.5097
Number of obs: 8206, groups:  Item, 156; Participant, 53

  Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
  (Intercept)      0.80865    0.38166   2.119   0.0341 *
  Vowel_H_type    -0.07894    0.41121  -0.192   0.8478
  Vowel_H_token   -0.66709    0.32099  -2.078   0.0377 *
  GroupAdult      -0.07907    0.15252  -0.518   0.6042
  TP_vowel_reg2-1  2.84267    0.44099   6.446 1.15e-10 ***
```

```
Call:
lm(formula = Part_Vowel_H ~ Vowel_H_type + Vowel_H_token + Group,
   data = Adult_child_H_data)

Residuals:
   Min      1Q  Median      3Q     Max
-1.6871 -0.5014 -0.0525  0.4283  3.1844

 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)    0.36024    0.09099   3.959 8.94e-05 ***
 Vowel_H_type   0.90345    0.10599   8.524 3.37e-16 ***
 Vowel_H_token -0.12619    0.09546  -1.322  0.18698
 GroupChild     0.20059    0.06614   3.033  0.00258 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6581 on 392 degrees of freedom
Multiple R-squared:  0.2317,  Adjusted R-squared:  0.2259
F-statistic: 39.41 on 3 and 392 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Part_Vowel_H ~ Vowel_H_type + Vowel_H_token + Group +
    TP, data = Adult_child_H_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7078 -0.5264 -0.0659  0.4272  3.1895

 Coefficients:
              Estimate Std. Error t value Pr(>|t|)
 (Intercept)    0.40785    0.13660   2.986  0.00301 **
 Vowel_H_type   0.88735    0.11154   7.955 1.95e-14 ***
 Vowel_H_token -0.12698    0.09557  -1.329  0.18475
 GroupChild     0.20059    0.06621   3.030  0.00261 **
 TP            -0.04163    0.08903  -0.468  0.64031
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6587 on 391 degrees of freedom
Multiple R-squared:  0.2322,  Adjusted R-squared:  0.2243
F-statistic: 29.56 on 4 and 391 DF,  p-value: < 2.2e-16
```

Chapter 4: Testing the Tolerance Principle in adults and children learning an artificial orthography

```
Adult_child_reg$Age <- factor(Adult_child_reg$Age, levels=c("Adult", "Child"))
Adult_child_reg$Condition <- factor(Adult_child_reg$Condition, levels=c("82","64", "46"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Condition + (1 | Participant)
   Data: Adult_AO1_reg

     AIC      BIC   logLik deviance df.resid
   829.1    847.4   -410.5    821.1      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3077 -0.7213  0.3023  0.6812  3.0010

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 1.117    1.057
Number of obs: 719, groups:  Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)   0.9251     0.2663   3.474 0.000513 ***
  Condition64  -0.4316     0.2113  -2.042 0.041103 *
  Condition46  -2.0873     0.2297  -9.087  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt64
Condition64 -0.420
Condition46 -0.415  0.498
>
```

```
Adult_child_reg$Condition <- factor(Adult_child_reg$Condition, levels=c("64","82", "46"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Condition + (1 | Participant)
   Data: Adult_AO1_reg

     AIC      BIC   logLik deviance df.resid
   829.1    847.4   -410.5    821.1      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3077 -0.7213  0.3023  0.6812  3.0010

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 1.117    1.057
Number of obs: 719, groups:  Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)   0.4935     0.2614   1.888   0.0590 .
  Condition82   0.4316     0.2113   2.042   0.0411 *
  Condition46  -1.6557     0.2214  -7.477 7.59e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) Cndt82
Condition82 -0.381
Condition46 -0.384  0.437
```

```
Adult_child_reg$Condition <- factor(Adult_child_reg$Condition, levels=c("46","64", "82"))
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Condition + (1 | Participant)
   Data: Adult_AO1_reg

     AIC      BIC   logLik deviance df.resid
   829.1    847.4   -410.5    821.1      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3077 -0.7213  0.3023  0.6812  3.0010

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 1.117    1.057
Number of obs: 719, groups:  Participant, 24

 Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)  -1.1622     0.2700  -4.304 1.68e-05 ***
 Condition64   1.6557     0.2214   7.477 7.60e-14 ***
 Condition82   2.0873     0.2297   9.086  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt64
Condition64 -0.448
Condition82 -0.441  0.562
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Token + (1 | Participant)
   Data: Adult_AO1_reg

     AIC      BIC   logLik deviance df.resid
   846.5    860.3   -420.3    840.5      716

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.0433 -0.6984  0.2473  0.6901  2.8952

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.95     0.9747
Number of obs: 719, groups:  Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)   -2.390      0.360  -6.641 3.13e-11 ***
  Token          4.141      0.480   8.627  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
      (Intr)
Token -0.797
>
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Token + TP + (1 | Participant)
   Data: Adult_AO1_reg

     AIC      BIC   logLik deviance df.resid
   828.8    847.1   -410.4    820.8      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4105 -0.6874  0.2932  0.6961  3.1170

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 1.042    1.021
Number of obs: 719, groups:  Participant, 24

  Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -1.0969     0.4690  -2.339   0.0194 *
  Token         1.5891     0.7483   2.123   0.0337 *
  TP2-1         1.3423     0.3099   4.332 1.48e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
      (Intr) Token
Token -0.872
TP2-1  0.618 -0.766
```

```
AO1_Pron_ad$Condition <- factor(AO1_Pron_ad$Condition, levels=c("82","64", "46"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Condition + (1 + Condition | Participant) + (1 | Item)
   Data: AO1_Pron_ad

     AIC      BIC   logLik deviance df.resid
   868.8    914.6   -424.4    848.8      709

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0670 -0.6490 -0.2331  0.6159  2.9205

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 Item        (Intercept) 0.8498   0.9218
 Participant (Intercept) 1.0410   1.0203
             Condition64 1.3546   1.1639   -0.55
             Condition46 1.1041   1.0508   -0.87  0.44
Number of obs: 719, groups:  Item, 30; Participant, 24

 Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)   0.9649     0.3959   2.437 0.014804 *
 Condition64  -0.9149     0.5280  -1.733 0.083160 .
 Condition46  -1.9318     0.5187  -3.724 0.000196 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt64
Condition64 -0.672
Condition46 -0.741  0.496
>
```

```
AO1_Pron_ad$Condition <- factor(AO1_Pron_ad$Condition, levels=c("64","82", "46"))


Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Condition + (1 + Condition | Participant) + (1 | Item)
   Data: AO1_Pron_ad

     AIC      BIC   logLik deviance df.resid
   868.8    914.6   -424.4    848.8      709

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0671 -0.6490 -0.2330  0.6159  2.9205

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 Item        (Intercept) 0.8499   0.9219
 Participant (Intercept) 1.0978   1.0478
             Condition82 1.3548   1.1640   -0.58
             Condition46 1.3918   1.1797   -0.89  0.60
Number of obs: 719, groups:  Item, 30; Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  0.04997    0.39325   0.127   0.8989
  Condition82  0.91492    0.52805   1.733   0.0832 .
  Condition46 -1.01680    0.52560  -1.935   0.0530 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt82
Condition82 -0.666
Condition46 -0.748  0.515
>
```

```
AO1_Pron_ad$Condition <- factor(AO1_Pron_ad$Condition, levels=c("46","64", "82"))


Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Condition + (1 + Condition | Participant) + (1 | Item)
   Data: AO1_Pron_ad

     AIC      BIC   logLik deviance df.resid
   868.8    914.6   -424.4    848.8      709

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0669 -0.6490 -0.2331  0.6159  2.9209

Random effects:
 Groups      Name        Variance Std.Dev. Corr
 Item        (Intercept) 0.8498   0.9218
 Participant (Intercept) 0.2894   0.5380
             Condition64 1.3925   1.1800   -0.46
             Condition82 1.1038   1.0506   -0.31  0.46
Number of obs: 719, groups:  Item, 30; Participant, 24

 Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)  -0.9667     0.3485  -2.774 0.005544 **
 Condition64   1.0168     0.5256   1.934 0.053068 .
 Condition82   1.9313     0.5187   3.723 0.000197 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt64
Condition64 -0.663
Condition82 -0.647  0.489
>
```

```
Child_AO1_reg <- Adult_child_reg[ which(Adult_child_reg$Age=='Child'), ]

Child_AO1_reg$Condition <- factor(Child_AO1_reg$Condition, levels=c("82","64", "46"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Condition + (1 | Participant)
   Data: Child_AO1_reg

     AIC      BIC   logLik deviance df.resid
   789.3    807.6   -390.7    781.3      713

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5777 -0.6875 -0.2697  0.6685  4.0050

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.6054   0.7781
Number of obs: 717, groups:  Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)   0.9612     0.2195   4.379 1.19e-05 ***
  Condition64  -0.9403     0.2046  -4.596 4.31e-06 ***
  Condition46  -2.9660     0.2577 -11.512  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) Cndt64
Condition64 -0.508
Condition46 -0.435  0.460
>
```

```
Child_AO1_reg$Condition <- factor(Child_AO1_reg$Condition, levels=c("64","82", "46"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Condition + (1 | Participant)
   Data: Child_AO1_reg

     AIC      BIC   logLik deviance df.resid
   789.3    807.6   -390.7    781.3      713

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5777 -0.6875 -0.2697  0.6685  4.0051

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.6055   0.7781
Number of obs: 717, groups:  Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  0.02087    0.21070   0.099    0.921
  Condition82  0.94030    0.20459   4.596 4.31e-06 ***
  Condition46 -2.02572    0.24443  -8.288  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt82
Condition82 -0.442
Condition46 -0.376  0.352
>
```

```
Child_AO1_reg$Condition <- factor(Child_AO1_reg$Condition, levels=c("46","64", "82"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Condition + (1 | Participant)
   Data: Child_AO1_reg

     AIC      BIC   logLik deviance df.resid
   789.3    807.6   -390.7    781.3      713

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5777 -0.6875 -0.2697  0.6685  4.0050

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.6054   0.7781
Number of obs: 717, groups:  Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -2.0048     0.2558  -7.838 4.56e-15 ***
  Condition64   2.0257     0.2444   8.288  < 2e-16 ***
  Condition82   2.9660     0.2577  11.512  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt64
Condition64 -0.646
Condition82 -0.634  0.669
>
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Token + (1 | Participant)
   Data: Child_AO1_reg

     AIC      BIC   logLik deviance df.resid
   808.8    822.6   -401.4    802.8      714

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.3268 -0.6740 -0.2962  0.6480  3.5491

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.8955   0.9463
Number of obs: 717, groups:  Participant, 24

 Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
 (Intercept)  -3.7822     0.3917  -9.657   <2e-16 ***
 Token         5.7916     0.5228  11.078   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
      (Intr)
Token -0.839
>
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Token + TP + (1 | Participant)
   Data: Child_AO1_reg

     AIC      BIC   logLik deviance df.resid
   789.1    807.4   -390.5    781.1      713

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1254 -0.6041 -0.2460  0.6918  4.3579

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.7618   0.8728
Number of obs: 717, groups:  Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -2.5800     0.4544  -5.678 1.36e-08 ***
  Token         3.3042     0.7241   4.563 5.03e-06 ***
  TP2-1         1.4414     0.3129   4.606 4.11e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
      (Intr) Token
Token -0.887
TP2-1  0.491 -0.675
>
```

```
Adult_child_reg$TP <- factor(Adult_child_reg$TP, levels=c("0","1"))

Adult_child_reg$Age <- factor(Adult_child_reg$Age, levels=c("Adult","Child"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Age * TP + (1 | Participant) + (1 | Item)
   Data: Adult_child_reg

     AIC      BIC   logLik deviance df.resid
  1701.8   1733.4   -844.9   1689.8     1430

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5994 -0.8414 -0.2779  0.8023  2.9675

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.0429   0.2071
 Participant (Intercept) 0.4560   0.6753
Number of obs: 1436, groups:  Item, 30; Participant, 27

  Fixed effects:
               Estimate Std. Error z value Pr(>|z|)
  (Intercept)   -0.9602     0.2111  -4.548 5.42e-06 ***
  AgeChild      -0.8869     0.2498  -3.551 0.000384 ***
  TP1            1.6747     0.1996   8.390  < 2e-16 ***
  AgeChild:TP1   0.7394     0.2822   2.620 0.008798 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) AgChld TP1
AgeChild    -0.451
TP1         -0.646  0.455
AgeChld:TP1  0.387 -0.864 -0.573
>
```

```
Adult_child_reg$TP <- factor(Adult_child_reg$TP, levels=c("1","0"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Age * TP + (1 | Participant) + (1 | Item)
   Data: Adult_child_reg

     AIC      BIC   logLik deviance df.resid
  1701.8   1733.4   -844.9   1689.8     1430

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5994 -0.8414 -0.2779  0.8023  2.9675

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.0429   0.2071
 Participant (Intercept) 0.4560   0.6753
Number of obs: 1436, groups:  Item, 30; Participant, 27

  Fixed effects:
               Estimate Std. Error z value Pr(>|z|)
  (Intercept)    0.7145     0.1730   4.130 3.63e-05 ***
  AgeChild      -0.1475     0.1423  -1.037   0.2998
  TP0           -1.6747     0.1996  -8.390  < 2e-16 ***
  AgeChild:TP0  -0.7394     0.2822  -2.620   0.0088 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) AgChld TP0
AgeChild    -0.418
TP0         -0.365  0.337
AgeChld:TP0  0.189 -0.467 -0.573
>
```

```
Adult_child_reg$Age <- factor(Adult_child_reg$Age, levels=c("Child","Adult"))

Adult_child_reg$TP <- factor(Adult_child_reg$TP, levels=c("0","1"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Age * TP + (1 | Participant) + (1 | Item)
   Data: Adult_child_reg

     AIC      BIC   logLik deviance df.resid
  1701.8   1733.4   -844.9   1689.8     1430

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5994 -0.8414 -0.2779  0.8023  2.9676

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.0429   0.2071
 Participant (Intercept) 0.4560   0.6753
Number of obs: 1436, groups:  Item, 30; Participant, 27

  Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
  (Intercept)   -1.8471     0.2437  -7.580 3.47e-14 ***
  AgeAdult       0.8870     0.2498   3.551 0.000383 ***
  TP1            2.4141     0.2344  10.300  < 2e-16 ***
  AgeAdult:TP1  -0.7394     0.2822  -2.620 0.008795 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) AgAdlt TP1
AgeAdult    -0.634
TP1         -0.742  0.652
AgeAdlt:TP1  0.550 -0.864 -0.716
>
```

```
Adult_child_reg$TP <- factor(Adult_child_reg$TP, levels=c("1","0"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Age * TP + (1 | Participant) + (1 | Item)
   Data: Adult_child_reg

     AIC      BIC   logLik deviance df.resid 
  1701.8   1733.4   -844.9   1689.8     1430 

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-2.5994 -0.8414 -0.2779  0.8022  2.9675 

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.0429   0.2071  
 Participant (Intercept) 0.4560   0.6753  
Number of obs: 1436, groups:  Item, 30; Participant, 27

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)    
  (Intercept)    0.5670     0.1721   3.295 0.000984 ***
  AgeAdult       0.1475     0.1423   1.037 0.299787    
  TP0           -2.4141     0.2344 -10.299  < 2e-16 ***
  AgeAdult:TP0   0.7394     0.2822   2.620 0.008799 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) AgAdlt TP0   
AgeAdult    -0.407              
TP0         -0.312  0.275       
AgeAdlt:TP0  0.197 -0.467 -0.716
>
```

```
AO1_Pron_ch$Condition <- factor(AO1_Pron_ch$Condition, levels=c("82","64", "46"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Acc ~ Condition + (1 | Participant) + (1 | Item)
   Data: AO1_Pron_ch

     AIC      BIC   logLik deviance df.resid
   789.8    812.7   -389.9    779.8      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1720 -0.5061 -0.3561  0.5475  3.1321

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 1.0879   1.0430
 Participant (Intercept) 0.1642   0.4052
Number of obs: 720, groups:  Item, 30; Participant, 24

 Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
 (Intercept)   0.7732     0.3779   2.046   0.0408 *
 Condition64  -1.2975     0.5176  -2.507   0.0122 *
 Condition46  -2.4467     0.5269  -4.643 3.43e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt64
Condition64 -0.695
Condition46 -0.684  0.502
>
```

```
AO1_Pron_ch$Condition <- factor(AO1_Pron_ch$Condition, levels=c("64","82", "46"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Acc ~ Condition + (1 | Participant) + (1 | Item)
   Data: AO1_Pron_ch

     AIC      BIC   logLik deviance df.resid
   789.8    812.7   -389.9    779.8      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1720 -0.5061 -0.3561  0.5475  3.1321

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 1.0879   1.0430
 Participant (Intercept) 0.1642   0.4052
Number of obs: 720, groups:  Item, 30; Participant, 24

 Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
 (Intercept)  -0.5243     0.3725  -1.408   0.1592
 Condition82   1.2975     0.5176   2.507   0.0122 *
 Condition46  -1.1492     0.5214  -2.204   0.0275 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt82
Condition82 -0.684
Condition46 -0.676  0.486
>
```

```
AO1_Pron_ch$Condition <- factor(AO1_Pron_ch$Condition, levels=c("46","64", "82"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Acc ~ Condition + (1 | Participant) + (1 | Item)
   Data: AO1_Pron_ch

     AIC      BIC   logLik deviance df.resid
   789.8    812.7   -389.9    779.8      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1720 -0.5061 -0.3561  0.5475  3.1321

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 1.0879   1.0430
 Participant (Intercept) 0.1642   0.4052
Number of obs: 720, groups:  Item, 30; Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -1.6735     0.3845  -4.352 1.35e-05 ***
  Condition64   1.1492     0.5214   2.204   0.0275 *
  Condition82   2.4467     0.5269   4.643 3.43e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt64
Condition64 -0.701
Condition82 -0.698  0.512
>
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Acc ~ Reg + (1 | Participant) + (1 | Item)
   Data: AO1_Pron_ch

     AIC      BIC   logLik deviance df.resid
   785.1    803.5   -388.6    777.1      716

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.0738 -0.5488 -0.3260  0.5684  3.7297

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 1.0801   1.0393
 Participant (Intercept) 0.1646   0.4057
Number of obs: 720, groups:  Item, 30; Participant, 24

  Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)  -1.8085     0.3585  -5.044 4.55e-07 ***
 RegR          2.1684     0.4415   4.912 9.03e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
     (Intr)
RegR -0.770
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Age + (1 | Participant) + (1 | Item)
   Data: AO1_ad_ch_pron

     AIC      BIC   logLik deviance df.resid
  1645.9   1667.0   -819.0   1637.9     1435

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0943 -0.5945 -0.3432  0.6415  3.9897

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.2424   0.4923
 Item        (Intercept) 1.6267   1.2754
Number of obs: 1439, groups:  Participant, 48; Item, 30

  Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
  (Intercept) -0.01784    0.26876  -0.066   0.9471
  AgeChild    -0.45132    0.18947  -2.382   0.0172 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
         (Intr)
AgeChild -0.350
```

Chapter 5: Testing the Tolerance Principle in adults learning an artificial orthography with high frequency irregulars

```
AO1_reg <- AO1_AO2_reg[ which(AO1_AO2_reg$Study=='1'), ]
AO2_reg <- AO1_AO2_reg[ which(AO1_AO2_reg$Study=='2'), ]

AO2_reg$Condition <- factor(AO2_reg$Condition, levels=c("82","64","46"))

Model1 <- glmer(Reg ~ Condition + (1|Participant), data = AO2_reg, family =binomial)
summary(Model1)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Condition + (1 | Participant)
   Data: AO2_reg

     AIC      BIC   logLik deviance df.resid
   900.2    918.5   -446.1    892.2      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9697 -0.7494 -0.2873  0.8163  3.4810

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.6553   0.8095
Number of obs: 719, groups:  Participant, 24

 Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
 (Intercept)   0.7774     0.2214   3.511 0.000446 ***
 Condition64  -1.3640     0.2068  -6.596 4.22e-11 ***
 Condition46  -1.3508     0.2064  -6.546 5.92e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt64
Condition64 -0.480
Condition46 -0.480  0.532
>
```

```
AO2_reg$Condition <- factor(AO2_reg$Condition, levels=c("64","82","46"))


Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Condition + (1 | Participant)
   Data: AO2_reg

     AIC      BIC   logLik deviance df.resid
   900.2    918.5   -446.1    892.2      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9697 -0.7494 -0.2873  0.8163  3.4810

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.6553   0.8095
Number of obs: 719, groups:  Participant, 24

  Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
  (Intercept) -0.58657    0.21878  -2.681  0.00734 **
  Condition82  1.36400    0.20679   6.596 4.22e-11 ***
  Condition46  0.01315    0.19980   0.066  0.94754
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt82
Condition82 -0.460
Condition46 -0.458  0.485
>
```

```
AO2_reg$Condition <- factor(AO2_reg$Condition, levels=c("46","64","82"))


Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Condition + (1 | Participant)
   Data: AO2_reg

     AIC      BIC   logLik deviance df.resid
   900.2    918.5   -446.1    892.2      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9697 -0.7494 -0.2873  0.8163  3.4810

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.6553   0.8095
Number of obs: 719, groups:  Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept) -0.57343    0.21843  -2.625  0.00866 **
  Condition64 -0.01314    0.19980  -0.066  0.94755
  Condition82  1.35085    0.20637   6.546 5.92e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt64
Condition64 -0.456
Condition82 -0.458  0.482
>
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Token + (1 | Participant) + (1 | Item)
   Data: AO2_reg

     AIC      BIC   logLik deviance df.resid
   906.2    924.6   -449.1    898.2      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5600 -0.8107 -0.3725  0.8409  4.6705

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.06963  0.2639
 Participant (Intercept) 0.73795  0.8590
Number of obs: 719, groups:  Item, 30; Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -1.6123     0.3125  -5.159 2.48e-07 ***
  Token         3.3103     0.5350   6.188 6.11e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
      (Intr)
Token -0.768
>
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Token + TP + (1 | Participant) + (1 | Item)
   Data: AO2_reg

     AIC      BIC   logLik deviance df.resid
   905.8    928.7   -447.9    895.8      714

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.8143 -0.8372 -0.3715  0.8545  4.7743

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.02248  0.1499
 Participant (Intercept) 0.75031  0.8662
Number of obs: 719, groups:  Item, 30; Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -1.9124     0.3610  -5.298 1.17e-07 ***
  Token         4.1503     0.7453   5.569 2.56e-08 ***
  TP2-1        -0.4394     0.2675  -1.643      0.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
      (Intr) Token
Token -0.832
TP2-1  0.536 -0.718
>
```

```
AO1_AO2_reg$Study <- as.factor(AO1_AO2_reg$Study)

AO1_AO2_reg$Study <- factor(AO1_AO2_reg$Study, levels=c("1","2"))

AO1_AO2_reg$TP <- factor(AO1_AO2_reg$TP, levels=c("0","1"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Study * TP + (1 | Participant) + (1 | Item)
   Data: AO1_AO2_reg

     AIC      BIC   logLik deviance df.resid
  1763.5   1795.1   -875.7   1751.5     1432

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4867 -0.7657 -0.1914  0.7730  3.4090

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.8766   0.9363
 Item        (Intercept) 0.1405   0.3749
Number of obs: 1438, groups:  Participant, 48; Item, 30

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -1.1644     0.2767  -4.209 2.57e-05 ***
  Study2        0.5780     0.3453   1.674   0.0942 .
  TP1           1.8802     0.2451   7.670 1.72e-14 ***
  Study2:TP1   -1.2068     0.2621  -4.604 4.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
           (Intr) Study2 TP1
Study2     -0.652
TP1        -0.599  0.310
Study2:TP1  0.364 -0.514 -0.603
>
```

```
AO1_AO2_reg$TP <- factor(AO1_AO2_reg$TP, levels=c("1","0"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Study * TP + (1 | Participant) + (1 | Item)
   Data: AO1_AO2_reg

     AIC      BIC   logLik deviance df.resid
  1763.5   1795.1   -875.7   1751.5     1432

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4867 -0.7657 -0.1914  0.7730  3.4090

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.8766   0.9363
 Item        (Intercept) 0.1405   0.3749
Number of obs: 1438, groups:  Participant, 48; Item, 30

 Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)   0.7158     0.2353   3.041  0.00235 **
 Study2       -0.6288     0.3081  -2.041  0.04123 *
 TP0          -1.8802     0.2451  -7.670 1.72e-14 ***
 Study2:TP0    1.2068     0.2621   4.604 4.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
           (Intr) Study2 TP0
Study2     -0.667
TP0        -0.337  0.165
Study2:TP0  0.200 -0.275 -0.603
>
```

```
AO1_AO2_reg$Study <- factor(AO1_AO2_reg$Study, levels=c("2","1"))

AO1_AO2_reg$TP <- factor(AO1_AO2_reg$TP, levels=c("0","1"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Study * TP + (1 | Participant) + (1 | Item)
   Data: AO1_AO2_reg

     AIC      BIC   logLik deviance df.resid
  1763.5   1795.1   -875.7   1751.5     1432

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4867 -0.7657 -0.1914  0.7730  3.4090

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.8766   0.9363
 Item        (Intercept) 0.1405   0.3748
Number of obs: 1438, groups:  Participant, 48; Item, 30

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -0.5864     0.2668  -2.198  0.02793 *
  Study1       -0.5779     0.3453  -1.674  0.09417 .
  TP1           0.6734     0.2265   2.973  0.00295 **
  Study1:TP1    1.2068     0.2621   4.604 4.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
           (Intr) Study1 TP1
Study1     -0.618
TP1        -0.570  0.259
Study1:TP1  0.287 -0.514 -0.505
>
```

```
AO1_AO2_reg$TP <- factor(AO1_AO2_reg$TP, levels=c("1","0"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Study * TP + (1 | Participant) + (1 | Item)
   Data: AO1_AO2_reg

     AIC      BIC   logLik deviance df.resid
  1763.5   1795.1   -875.7   1751.5     1432

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.4867 -0.7657 -0.1914  0.7730  3.4090

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.8766   0.9363
 Item        (Intercept) 0.1405   0.3749
Number of obs: 1438, groups:  Participant, 48; Item, 30

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)   0.0869     0.2314   0.375  0.70730
  Study1        0.6289     0.3081   2.041  0.04122 *
  TP0          -0.6733     0.2265  -2.973  0.00295 **
  Study1:TP0   -1.2068     0.2621  -4.604 4.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
           (Intr) Study1 TP0
Study1     -0.653
TP0        -0.321  0.139
Study1:TP0  0.163 -0.275 -0.505
>
```

```
AO2_reg_excl$Ind <- factor(AO2_reg_excl$Ind, levels=c("0","1"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ Token + Ind + (1 | Participant) + (1 | Item)
   Data: AO2_reg_excl

     AIC      BIC   logLik deviance df.resid
   799.5    822.2   -394.8    789.5      684

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.3578 -0.6317 -0.2430  0.6429  3.5086

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.2204   0.4695
 Participant (Intercept) 0.6419   0.8012
Number of obs: 689, groups:  Item, 30; Participant, 23

  Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)  -1.7787     0.3520  -5.054 4.34e-07 ***
 Token         0.2800     0.7294   0.384    0.701
 Ind1          2.3800     0.2938   8.101 5.43e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
      (Intr) Token
Token -0.627
Ind1  -0.115 -0.511
```

```
AO2_reg_excl$TP <- factor(AO2_reg_excl$TP, levels=c("0","1"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ TP + (1 | Participant) + (1 | Item)
   Data: AO2_reg_excl

     AIC      BIC   logLik deviance df.resid
   909.4    927.6   -450.7    901.4      685

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1628 -0.8321 -0.4679  0.8406  2.3228

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.2948   0.5429
 Participant (Intercept) 0.3964   0.6296
Number of obs: 689, groups:  Item, 30; Participant, 23

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -0.4955     0.2592  -1.911   0.0560 .
  TP1           0.7050     0.2740   2.573   0.0101 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
    (Intr)
TP1 -0.704
>
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Reg ~ TP + Ind + (1 | Participant) + (1 | Item)
   Data: AO2_reg_excl

     AIC      BIC   logLik deviance df.resid
   798.2    820.9   -394.1    788.2      684

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.5815 -0.6424 -0.2218  0.6538  3.6820

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.2112   0.4596
 Participant (Intercept) 0.6761   0.8223
Number of obs: 689, groups:  Item, 30; Participant, 23

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)   1.0085     0.3169   3.182  0.00146 **
  TP1          -0.3435     0.2873  -1.196  0.23186
  Ind0         -2.5551     0.2739  -9.329  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
     (Intr) TP1
TP1  -0.708
Ind0 -0.494  0.397
>
```

```
AO1_AO2_ad_pron$Condition <- factor(AO1_AO2_ad_pron$Condition, levels=c("82","64",
"46"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Condition + (1 | Participant) + (1 | Item)
   Data: AO2_Accuracy

     AIC      BIC   logLik deviance df.resid
   956.9    979.8   -473.5    946.9      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.7682 -0.8681 -0.4042  0.8383  2.1327

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.2469   0.4969
 Participant (Intercept) 0.2732   0.5227
Number of obs: 720, groups:  Item, 30; Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)   0.5087     0.2363   2.152 0.031368 *
  Condition64  -0.5838     0.2962  -1.971 0.048747 *
  Condition46  -0.9994     0.2983  -3.350 0.000807 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt64
Condition64 -0.635
Condition46 -0.633  0.505
>
```

```
AO1_AO2_ad_pron$Condition <- factor(AO1_AO2_ad_pron$Condition, levels=c("64","82",
"46"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Condition + (1 | Participant) + (1 | Item)
   Data: AO2_Accuracy

     AIC      BIC   logLik deviance df.resid
   956.9    979.8   -473.5    946.9      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.7682 -0.8681 -0.4042  0.8383  2.1327

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.2469   0.4969
 Participant (Intercept) 0.2732   0.5227
Number of obs: 720, groups:  Item, 30; Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
 (Intercept) -0.07509    0.23375  -0.321   0.7480
 Condition82  0.58377    0.29621   1.971   0.0487 *
 Condition46 -0.41566    0.29567  -1.406   0.1598
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt82
Condition82 -0.625
Condition46 -0.625  0.492
>
```

```
AO1_AO2_ad_pron$Condition <- factor(AO1_AO2_ad_pron$Condition, levels=c("46","64",
"82"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Condition + (1 | Participant) + (1 | Item)
   Data: AO2_Accuracy

     AIC      BIC   logLik deviance df.resid
   956.9    979.8   -473.5    946.9      715

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.7682 -0.8681 -0.4042  0.8383  2.1327

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.2469   0.4969
 Participant (Intercept) 0.2732   0.5227
Number of obs: 720, groups:  Item, 30; Participant, 24

  Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -0.4907     0.2359  -2.080 0.037485 *
  Condition64   0.4156     0.2957   1.406 0.159788
  Condition82   0.9994     0.2983   3.350 0.000807 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Cndt64
Condition64 -0.634
Condition82 -0.631  0.503
>
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Reg + (1 | Participant) + (1 | Item)
   Data: AO2_Accuracy

     AIC      BIC   logLik deviance df.resid
   942.9    961.2   -467.4    934.9      716

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.6633 -0.8332 -0.3902  0.8212  2.4370

Random effects:
 Groups      Name        Variance Std.Dev.
 Item        (Intercept) 0.1021   0.3196
 Participant (Intercept) 0.2717   0.5212
Number of obs: 720, groups:  Item, 30; Participant, 24

 Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)  -0.7010     0.1916  -3.658 0.000254 ***
 RegR          1.1276     0.2048   5.507 3.65e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
     (Intr)
RegR -0.652
>
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Study + (1 | Participant) + (1 | Item)
   Data: AO1_AO2_ad_pron

     AIC      BIC   logLik deviance df.resid
  1819.5   1840.6   -905.8   1811.5     1435

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.3579 -0.7829 -0.3281  0.7460  2.8878

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.2888   0.5374
 Item        (Intercept) 0.7952   0.8917
Number of obs: 1439, groups:  Participant, 48; Item, 30

 Fixed effects:
               Estimate Std. Error z value Pr(>|z|)
 (Intercept) -0.017716   0.213065  -0.083    0.934
 Study2      -0.003929   0.194175  -0.020    0.984

Correlation of Fixed Effects:
       (Intr)
Study2 -0.455
>
```

```
AO1_AO2_ad_pron$Study <- factor(AO1_AO2_ad_pron$Study, levels=c("1","2"))

AO1_AO2_ad_pron$Reg <- factor(AO1_AO2_ad_pron$Reg, levels=c("R","I"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Study * Reg + (1 | Participant) + (1 | Item)
   Data: AO1_AO2_ad_pron

     AIC      BIC   logLik deviance df.resid
  1797.6   1829.2   -892.8   1785.6     1433

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4293 -0.7905 -0.3418  0.7487  3.1960

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.2886   0.5372
 Item        (Intercept) 0.3269   0.5717
Number of obs: 1439, groups:  Participant, 48; Item, 30

 Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept)   0.6243     0.2041   3.059  0.00222 **
 Study2       -0.1782     0.2152  -0.828  0.40766
 RegI         -1.6270     0.2764  -5.886 3.96e-09 ***
 Study2:RegI   0.4567     0.2419   1.888  0.05905 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Study2 RegI
Study2      -0.532
RegI        -0.529  0.193
Study2:RegI  0.232 -0.430 -0.456
>
```

```
AO1_AO2_ad_pron$Reg <- factor(AO1_AO2_ad_pron$Reg, levels=c("I","R"))

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
['glmerMod']
 Family: binomial  ( logit )
Formula: Score ~ Study * Reg + (1 | Participant) + (1 | Item)
   Data: AO1_AO2_ad_pron

     AIC      BIC   logLik deviance df.resid
  1797.6   1829.2   -892.8   1785.6     1433

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4293 -0.7905 -0.3418  0.7487  3.1960

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 0.2886   0.5372
 Item        (Intercept) 0.3269   0.5718
Number of obs: 1439, groups:  Participant, 48; Item, 30

  Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
  (Intercept)  -0.7242     0.2380  -3.042  0.00235 **
  Study1       -0.2785     0.2451  -1.136  0.25589
  RegR          1.1703     0.2720   4.302 1.69e-05 ***
  Study1:RegR   0.4567     0.2419   1.888  0.05905 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Study1 RegR
Study1      -0.500
RegR        -0.691  0.257
Study1:RegR  0.298 -0.610 -0.426
>
```

References

Adelman, J. S., & Brown, G. D. A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, *115*(1), 214–227. https://doi.org/10.1037/0033-295X.115.1.214

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition, 90*(2), 119-161. https://doi.org/10.1016/S0010-0277(03)00146-X

Allen, S. (1996). *Aspects of argument structure acquisition in Inuktitut* (Vol. 13). John Benjamins Publishing. https://doi.org/10.1075/lald.13

Ambridge, B. (2010). Children's judgments of regular and irregular novel past-tense forms: New data on the English past-tense debate. *Developmental Psychology, 46*(6), 1497–1504. https://doi.org/10.1037/a0020668

Anderson, J. R. (1990). *The Adaptive Character of Thought* (1st ed.). Psychology Press. https://doi.org/10.4324/9780203771730

Anderson, S. (1969). *West Scandinavian Vowel Systems and the Ordering of Phonological Rules*. [Doctoral dissertation] Massachusetts Institute of Technology.

Andrews, S., & Scarratt, D. R. (1998). Rule and analogy mechanisms in reading nonwords: Hough dou peapel rede gnew wirds? *Journal of Experimental Psychology: Human Perception and Performance, 24*(4), 1052–1086. https://doi.org/10.1037/0096-1523.24.4.1052

Apfelbaum, K. S., Hazeltine, E., & McMurray, B. (2013). Statistical learning in reading: Variability in irrelevant letters helps children learn phonics skills. *Developmental Psychology, 49*(7), 1348–1365. https://doi.org/10.1037/a0029839

Arciuli, J., & Simpson, I. (2012). Statistical Learning Is Related to Reading Ability in Children and Adults. *Cognitive science, 36*, 286-304. 10.1111/j.1551-6709.2011.01200.x.

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

Baayen, R. H. (1989). *A Corpus-Based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation*. [Dissertation] Vrije Universiteit, Amsterdam.

Baayen, H., & Lieber, R. (1991). *Productivity and English derivation: A corpus-based study. Linguistics, 29*(5), 801-843. doi:10.1515/ling.1991.29.5.801.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1995). The CELEX lexical database (CD-ROM). Philadelphia: Linguistic Data Consortium.

Balota, D. A., & Duchek, J. M. (1988). Age-related differences in lexical access, spreading activation, and simple pronunciation. *Psychology and Aging, 3*(1), 84–93. https://doi.org/10.1037/0882-7974.3.1.84

Balota, D.A., Yap, M.J., Hutchison, K.A. et al. The English Lexicon Project. Behavior Research Methods 39, 445–459 (2007). https://doi.org/10.3758/BF03193014

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language, 68*(3), https://doi.org/10.1016/j.jml.2012.11.001

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

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

Besner, D., Twilley, L., McCann, R. S., & Seergobin, K. (1990). On the association between connectionism and data: Are a few words necessary? *Psychological Review, 97*(3), 432–446. https://doi.org/10.1037/0033-295X.97.3.432

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

Brown, G. D. A., & Deavers, R. P. (1999). Units of analysis in nonword reading: Evidence from children and adults. *Journal of Experimental Child Psychology, 73*(3), 208–242. https://doi.org/10.1006/jecp.1999.2502

Bruck, M., & Treiman, R. (1992). Learning to pronounce words: The limitations of analogies. *Reading Research Quarterly, 27*(4), 374–388. https://doi.org/10.2307/747676

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*(5), 425–455. https://doi.org/10.1080/01690969508407111

Bybee, J.L. & Moder, C.L. (1983). Rules and schemes in the development and use of the English past tense. *Language, 59*, 251-270.

Bybee, J. L., & Slobin, D. I. (1982). Rules and Schemas in the Development and Use of the English past Tense. *Language, 58*(2), 265-289.

Byrne, B. (1984). On teaching articulatory phonetics via an orthography. *Memory & Cognition, 12*(2), 181–189. https://doi.org/10.3758/BF03198432

Byrne, B.,& Carroll, M. (1989). Learning artificial orthographies: Further evidence of a nonanalytic acquisition procedure. *Memory & Cognition 17*, 311–317. https://doi.org/10.3758/BF03198469

Caprin, C., & Guasti, M. T. (2009). The acquisition of morphosyntax in Italian: A cross-sectional study. *Applied Psycholinguistics*, *30*(01), 23–32. https://doi/org/10.1017/S0142716408090024

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Cazden, C. B. (1968). The acquisition of noun and verb inflections. *Child Development*, *39*(2), 433–448. https://doi.org/10.2307/1126956

Cerella, J., & Fozard, J. L. (1984). Lexical access and age. *Developmental Psychology, 20*(2), 235–243. https://doi.org/10.1037/0012-1649.20.2.235

Chang, Y.-N., Monaghan, P., & Welbourne, S. (2019). A computational model of reading across development: Effects of literacy onset on language processing. *Journal of Memory and Language, 108*, Article 104025. https://doi.org/10.1016/j.jml.2019.05.003

Chang, Y.-N., & Monaghan, P. (2019). Quantity and diversity of preliteracy language exposure both affect literacy development: Evidence from a computational model of reading. *Scientific Studies of Reading, 23*(3), 235-253, 10.1080/10888438.2018.1529177

Chang, Y.-N., Taylor, J. S. H., Rastle, K., & Monaghan, P. (2020). The relationships between oral language and reading instruction: Evidence from a computational model of reading. *Cognitive Psychology, 123*, Article 101336. https://doi.org/10.1016/j.cogpsych.2020.101336

Chater, N., Clark, A., Perfors, A., & Goldsmith, J. A. (2015). *Empiricism and Language Learnability*. Oxford University Press. https://doi:10.1093/acprof:oso/9780198734260.001.0001

Chater N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends Cognitive Science, 7*(1), 19-22. https://doi: 10.1016/s1364-6613(02)00005-0.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry, 36*(1), 1–22. https://doi.org/10.1162/0024389052993655

Chomsky, N., & Morris, H. (1968). *The sound pattern of English*. Harper & Row.

Christiansen, M. H., & Chater, N. (2008) Language as shaped by the brain. *Behavioural Brain Science. 31*(5), 489-558. https://doi: 10.1017/S0140525X08004998.

Clahsen, H., & Penke, M. (1992). The acquisition of agreement morphology and its syntactic consequences: New evidence on German child language from the Simone-Corpus. In Meisel, J.M. (Eds.), *The acquisition of verb placement* (pp. 181-223). Springer. https://doi.org/10.1007/978-94-011-2803-2_7

Clark, R. (2001). Information theory, complexity, and linguistic descriptions. In Bertolo, S. (Ed.), *Parametric Linguistics and Learnability* (pp. 126-171). Cambridge University Press.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100*(4), 589–608. https://doi.org/10.1037/0033-295X.100.4.589

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon in Dornic, S. (Ed.), *Attention and performance VI*, (pp. 535-555). Lawrence Erlbaum Associates.

Coltheart, M., & Rastle, K. (1994). Serial processing in reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance, 20*(6), 1197–1211. https://doi.org/10.1037/0096-1523.20.6.1197

Coltheart, M. & Rastle, K. (1999). Serial and Strategic Effects in Reading Aloud. *Journal of Experimental Psychology:Human Perception and Performance, 25*(2), 482-503. https://doi.org/10.1037//0096-1523.25.2.482

Coltheart M., Rastle K., Perry C., Langdon R., & Ziegler J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review 108*(1), 204-56. https://doi: 10.1037/0033-295x.108.1.204

Coltheart, V., & Leahy, J. (1992). Children's and adults' reading of nonwords: Effects of regularity and consistency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(4), 718–729. https://doi.org/10.1037/0278-7393.18.4.718

Culbertson, J. & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in Psychology, 6.* 10.3389/fpsyg.2015.01964.

Cunningham, A.E., & Stanovich, K.E. (1993) Children's literacy environments and early word recognition subskills. *Reading and Writing: An Interdisciplinary Journal, 5*(2), 193–204. https://doi.org/10.1007/BF01027484

De Cat, C. (2018). Evaluating Yang's algorithms: An outline. *Linguistic Approaches to Bilingualism, 8*(6), 712–716. https://doi.org/10.1075/lab.18066.de

Ehri, L. C., & Saltmarsh, J. (1995). Beginning readers outperform older disabled readers in learning to read words by sight. *Reading and Writing, 7*(3), 295–326. https://doi.org/BF03162082

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition, 48*(1), 71-99.

Endress, A. D., & Hauser, M. D. (2011). The influence of type and token frequency on the acquisition of affixation patterns: Implications for language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(1), 77–95. https://doi.org/10.1037/a0020210

Ervin, S. M., & Miller, W. R. (1963). Language development. In H. W. Stevenson (Ed.) & J. Kagan, C. Spiker (Collaborators) & N. B. Henry, H. G. Richey (Eds.), *Child psychology: The sixty-second yearbook of the National Society for the Study of Education, Part 1* (pp. 108–143). National Society for the Study of Education; University of Chicago Press. https://doi.org/10.1037/13101-004

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 109*(44), 17897-17902. https://doi.org/10.1073/pnas.1215776109

Ferdinand, V. Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition, 184*, 53-68. https://:doi.org/10.1016/j.cognition.2018.12.002.

Forster, K. I., & Bednall, E. S. (1976). Terminating and exhaustive search in lexical access. *Memory & Cognition, 4*(1), 53–61. https://doi.org/10.3758/BF03213255

Forster, K.I., Forster, J.C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers 35*, 116–124. https://doi.org/10.3758/BF03195503

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage*. Houghton Mifflin.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences, 23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 5*(4), 674–691. https://doi.org/10.1037/0096-1523.5.4.674

Gopnik, M. (1990). Feature blindness: A case study. *Language Acquisition, 1*(2), 139–164. https://doi.org/10.1207/s15327817la0102_1

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In Hale, K., & Keyser S. J. (Eds.), *The view from building 20* (pp. 111-176). The MIT Press.

Harm, M. W., & Seidenberg, M. S. (2002). Division of labor in a multicomponent connectionist model of reading: Computing the meanings of words [Version submitted for publication].

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

Hooper, J. B. (1976). Word frequency in lexical diffusion and the source of morphophonological change. In Christie, W. (Ed.), *Current Progress in Historical Linguistics* (pp. 96–105). North-Holland.

Hudson Kam, C. L. & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1*(2), 151–195. https://doi.org/10.1080/15475441.2005.9684215

Hudson Kam, C. L. & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive psychology, 59*(1), 30–66. https://doi.org/10.1016/j.cogpsych.2009.01.001

Jaeger T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language, 59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Jaeger, T. Florian & Tily, Harry. (2011). On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(3), 323 - 335. https://doi.org/10.1002/wcs.126.

Jared, D. (1997). Spelling-sound consistency affects the naming of high-frequency words. *Journal of Memory and Language, 36*(4), 505-529. https://doi.org/10.1006/jmla.1997.2496

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. *Journal of Memory and Language, 46*(4), 723–750. https://doi.org/10.1006/jmla.2001.2827

Jared, D., McRae, K., & Seidenberg, M. S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language, 29*(6), 687–715. https://doi.org/10.1016/0749-596X(90)90044-Z

Joanisse, M. F., & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation and processing. *WIREs Cognitive Science, 6*(3), 235–247. https://doi.org/10.1002/wcs.1340

Kapatsinski, V. (2018). On the intolerance of the Tolerance Principle. *Linguistic Approaches to Bilingualism, 8*(6), 738–742. https://doi.org/10.1075/lab.18052.kap

Kay, J., & Bishop, D. (1987). Anatomical differences between nose, palm, and foot, or, the body in question: Further dissection of the processes of sub-lexical spelling-sound translation. In Coltheart, M. (Ed.), *Attention and performance XII: The psychology of reading* (pp. 449–469). Erlbaum.

Kessler, B., & Treiman, R. (2001). Relationship between sounds and letters in English monosyllables. *Journal of Memory and Language, 44*(4), 592–617. https://doi.org/10.1006/jmla.2000.2745

Kessler, B. (2009). Statistical learning of conditional orthographic correspondences. *Writing Systems Research, 1,* 19-34. https://doi.org/10.1093/wsr/wsp004

Kiparsky, P. (1982). Lexical morphology and phonology. In Yang, L.-S. (Ed.), *Linguistics in the morning calm* (pp. 3-91). Hanshin.

Koulaguina, E. & Shi, R. (2019) Rule generalization from inconsistent input in early infancy. *Language Acquisition, 26*(4), 416-435. https://doi.org/10.1080/10489223.2019.1572148

Kurumada C., Meylan S.C., & Frank M.C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition, 127*(3), 439-53. https://doi.org/10.1016/j.cognition.2013.02.002.

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)*1393. https://doi.org/10.3389/fpsyg.2018.01393

Laxon, V., Masterson, J., & Coltheart, V. (1991). Some bodies are easier to read: The effect of consistency and regularity on children's reading. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *43A*(4), 793–824. https://doi.org/10.1080/14640749108400958

Lazaridou-Chatzigoga, D., Katsos, N., & Stockall, L. (2019). Generalizing about striking properties: Do glippets love to play with fire? *Frontiers in Psychology (10)*1971. https://doi.org/10.3389/fpsyg.2019.01971

Lieberman, E., Michel, J. B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature, 449*(7163), 713–716. https://doi.org/10.1038/nature06137

Mach, E. (1959). *The analysis of sensations and the relation of the physical to the psychical*. Dover Publications. [Original work published 1886]

Marcus, G. (1995). Children's overregularization of English plurals: A quantitative analysis. *Journal of Child Language*, *22*(2), 447-59. https://doi.org/10.1017/S0305000900009879

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the society for research in child development, 57*(4), i–178. https://doi.org/10.2307/1166115

Marsh, G., Friedman, M., Welch, V., & Desberg, P. (1981). A cognitive-developmental theory of reading acquisition. In G. E. MacKinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 3, pp. 199–221). Academic Press.

Maye J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), 101-11. https://doi/org/10.1016/s0010-0277(01)00157-3

Mcclelland, J. & Bybee, J. (2007). Gradience of Gradience: A reply to Jackendoff. *Linguistic Review, 24*(4) 437-455. https://doi.org/10.1515/TLR.2007.019

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: An account of basic findings. *Psychological Review, 88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

Merkx, M., Rastle, K., & Davis, M. H. (2011). The acquisition of morphological knowledge investigated through artificial language learning. *The Quarterly Journal of Experimental Psychology, 64*(6), 1200–1220. https://doi.org/10.1080/17470218.2010.538211

Mousikou, P., Sadat, J., Lucas, R., & Rastle, K. (2017) Moving beyond the monosyllable in models of skilled reading: Mega-study of disyllabic nonword reading. *Journal of Memory and Language, 93*, 169-192. https://doi.org/10.1016/j.jml.2016.09.003.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111*(3), 721–756. https://doi.org/10.1037/0033-295X.111.3.721

Newport, E.L. (1990), Maturational Constraints on Language Learning. *Cognitive Science, 14,*11-28. https://doi.org/10.1207/s15516709cog1401_2

Norris, D. (1994). A quantitative multiple-levels model of reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 20*(6), 1212–1232. https://doi.org/10.1037/0096-1523.20.6.1212

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Paap, K. R., & Noel, R. W. (1991). Dual-route models of print to sound: Still a good horse race. *Psychological Research, 53*(1), 13–24. https://doi.org/10.1007/BF00867328

Perfors, A., Ransom, K., & Navarro, D. J. (2014). People ignore token frequency when deciding how widely to generalize. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society, Canada.*

Perry, C., Ziegler, J. C., & Zorzi, M. (2007) Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review, 114*(2), 273-315. https://doi.org/10.1037/0033-295X.114.2.273

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology, 61*(2) pp. 106-151. https://doi.org/10.1016/j.cogpsych.2010.04.001

Pinker, S. (1999). *Words and Rules*. Basic Books.

Pinker, S. (2006). Whatever Happened to the Past Tense Debate? *UC Santa Cruz: Festschrifts*. https://escholarship.org/uc/item/0xf9q0n8

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition 28*(1-2), 73-193.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6*(11), 456–463. https://doi.org/10.1016/S1364-6613(02)01990-3

Plaut, D. C., & McClelland, J. L. (1993) Generalization with componential attractors: Word and nonword reading in an attractor network. *Proceedings of the fifteenth annual conference of the Cognitive Science Society,* (pp. 824-829). Erlbaum.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56-115.

Plunkett, K. (1991). Connectionists Approaches to Language Processing and Acquisition. *HERMES - Journal of Language and Communication in Business, 4*(6), 31–63. https://doi.org/10.7146/hjlcb.v4i6.21455

Powell, D., Plaut, D. C., & Funnell, E. (2001). A developmental evaluation of the Plaut, McClelland, Seidenberg, & Patterson (1996) connectionist model of single word reading. *Meeting of the British Psychological Society, Developmental and Education Sections, UK.*

Powell, D., Plaut, D. C., & Funnell, E. (2006). Does the Plaut, McClelland, Seidenberg and Patterson (1996) model of reading learn to read in the same way as a child? *Journal of Research in Reading, 29*(2), 229-250. https://doi.org/10.1111/j.1467-9817.2006.00300.

Powell, D., Stainthorp, R., & Stuart, M. (2014). Deficits in orthographic knowledge in children poor at rapid automatized naming (RAN) tasks? *Scientific Studies of Reading, 18*(3), 192-207. https://doi.org/10.1080/10888438.2013.862249

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes, 8*(1), 1–56. https://doi.org/10.1080/01690969308406948

Pritchard, S. C., Coltheart, M., Palethorpe, S., & Castles, A. (2012). Nonword reading: Comparing dual-route cascaded and connectionist dual-process models with human data. *Journal of Experimental Psychology. Human Perception and Performance, 38,* 1268-1288. https://doi.org/10.1037/a0026703

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rastle, K., & Coltheart, M. (2000). Lexical and nonlexical print-to-sound translation of disyllabic words and nonwords. *Journal of Memory and Language, 42*(3), 342-364, https://doi.org/10.1006/jmla.1999.2687.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *Quarterly Journal of Experimental Psychology, 55A,* 1339-1362.

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

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review, 111*(1), 159–182. https://doi.org/10.1037/0033-295X.111.1.159

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2017). Distributional learning of subcategories in an artificial grammar: Category generalization and subcategory restrictions. *Journal of Memory and Language, 97,* 17-29. https://doi.org/10.1016/j.jml.2017.07.006.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II* (pp. 64-99). Appleton-Century-Crofts.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*(5), 465–658. https://doi.org/10.1016/0005-1098(78)90005-5

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning & Verbal Behavior, 9*(5), 487–494. https://doi.org/10.1016/S0022-5371(70)80091-3

Rueckl, J. G., & Dror, I. E. (1994). The effect of orthographic-semantic consistency on the acquisition of new words. In C. Umilta & M. Moscovitch (Eds.), *Attention and performance, XV* (pp. 571–588). Erlbaum.

Rumelhart, & D. E., & Mcclelland, J. L. (1986). *Parallel distributed processing: explorations in the microstructure of cognition.* Volume 1. MIT Press.

Ryder, R., & Pearson, P. (1980). Influence of type-token frequencies and final consonants on adults' internalization of vowel digraphs. *Journal of Educational Psychology*, *72*(5), 618-624. https://doi.org/10.1037/0022-0663.72.5.618.

Saffran J. R. (2003) Statistical Language Learning: Mechanisms and Constraints. *Current Directions in Psychological Science, 12*(4), 110-114. https://doi:10.1111/1467-8721.01243

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*(4), 606–621. https://doi.org/10.1006/jmla.1996.0032

Saffran, J. R., 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:10.1111/j.1467-9280.1997.tb00690.x

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

Schuler, K. D. (2017). The acquisition of productive rules in child and adult language learners. [Unpublished doctoral dissertation]. Georgetown University.

Schuler, K. D., Reeder, P.A., Newport, E.L., & Aslin, R.N. (2017). The effect of Zipfian frequency variations on category formation in adult artificial language learning. *Language Learning and Development, 13*(4), 357-374. doi:10.1080/15475441.2016.1263571

Schuler, K., Yang, C., & Newport, E. (2021). Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient. PsyArXiv. https://doi.org/10.31234/osf.io/utgds

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 523-568. https://doi.org/10.1037/0033-295X.96.4.523

Seidenberg, M. S., Plaut, D. C., Petersen, A. S., McClelland, J. L., & McRae, K. (1994). Nonword pronunciation and models of word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 20*(6), 1177–1196. https://doi.org/10.1037/0096-1523.20.6.1177

Seidenberg, M. S., Waters, G. S., Barnes, M. A., & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning & Verbal Behavior*, *23*(3), 383–404. https://doi.org/10.1016/S0022-5371(84)90270-6

Shankweiler, D., & Liberman, I. Y. (1972). Misreading: A search for causes. In J. F. Kavanagh & I. G. Mattingly (Eds.), Language by ear and by eye: The relationship between speech and reading (pp. 239–317). Cambridge, MA: MIT Press

Shannon, C. E. (1949). The mathematical theory of communication. In C.E. Shannon, & W. Weaver (Eds.), *The mathematical theory of communication* (pp. 29-125). University of Illinois Press.

Shapiro, L. R., & Solity, J. (2008). Delivering phonological and phonics training within whole-class teaching. *British Journal of Educational Psychology,78*(4). 597-620. https://doi.org/10.1348/000709908X293850

Shapiro, L.R., & Solity, J. (2016). Differing effects of two synthetic phonics programmes on early reading development. *British Journal of Educational Psychology, 86*(2), 182-203. https://doi.org/10.1111/bjep.12097

Solity, J. (2020). Instructional psychology and teaching reading: Ending the reading wars, *Educational and Developmental Psychologist, 37*(2) 123-132. https://doi.org/10.1017/edp.2020.18

Solity, J., Deavers, R., Kerfoot, S., Crane, G. & Cannon, K. (2000). The Early Reading Research: The impact of instructional psychology. *Educational Psychology in Practice, 16*(2) 109-129. https://doi.org/10.1080/02667360050122190

Solity, J., & Vousden, J. (2009). Real books vs reading schemes: a new perspective from instructional psychology. *Educational Psychology, 29*(4) 469-511. https://doi.org/10.1080/01443410903103657

Steacy, L. M., Elleman, A. M., Lovett, M. W., & Compton, D. L. (2016) Exploring differential effects across two decoding treatments on item-level transfer in children with significant word reading difficulties: A new approach for testing intervention elements. *Scientific Studies of Reading, 20*(4), 283-295. https://doi.org/10.1080/10888438.2016.1178267

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

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say “broke”? A model of learning the past tense without feedback. *Cognition*, *86*(2), 123 – 155. https://doi.org/10.1016/S0010-0277(02)00176-2

Tabossi, P., & Laghi, L. (1992). Semantic priming in the pronunciation of words in two writing systems: Italian and English. *Memory & Cognition, 20*(3), 303–313. https://doi.org/10.3758/BF03199667

Tamminen, J., Davis, M. H., Merkx, M., & Rastle, K. (2012). The role of memory consolidation in generalisation of new linguistic information. *Cognition, 125*(1), 107-112. https://doi.org/10.1016/j.cognition.2012.06.014

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology, 79*, 1–39. https://doi.org/10.1016/j.cogpsych.2015.03.003

Taylor, J., Davis, M., & Rastle, K. (2017). Comparing and validating methods of reading instruction using behavioural and neural findings in an artificial orthography. *Journal of Experimental Psychology,* 146(6), 826–858

Taylor, J., Davis, M., & 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., 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

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Bulletin, 139*(4), 792–814. https://doi.org/10.1037/a0030801

Thiessen, E. D., & Pavlik, P. I., Jr. (2013). iMinerva: A mathematical model of distributional statistical learning. *Cognitive Science, 37*(2), 310–343. https://doi.org/10.1111/cogs.12011

Thompson, G. B., Connelly, V., Fletcher-Flinn, C. M., & Hodson, S. J. (2009). The nature of skilled adult reading varies with type of instruction in childhood. *Memory & Cognition, 37*(2), 223–234. https://doi.org/10.3758/MC.37.2.223

Tong, S., Zhang, P., & He, X. (2020). Statistical learning of orthographic regularities in chinese children with and without dyslexia. *Child Development, 91*, 1953-1969. https://doi.org/10.1111/cdev.13384

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). Test of Word-Reading Efficiency - Second Edition (TOWRE 2). Austin, TX: Pro-Ed.

Treiman, R. (2018). What Research Tells Us About Reading Instruction. *Psychological Science in the Public Interest, 19*(1), 1–4. https://doi.org/10.1177/1529100618772272

Treiman, R., Goswami, U., & Bruck, M. (1990). Not all nonwords are alike: Implications for reading development and theory. *Memory & Cognition 18*, 559–567. https://doi.org/10.3758/BF03197098

Treiman, R., & Kessler, B. (2019). Development of context-sensitive pronunciation in reading: The case of ‹c› and ‹g›. *Journal of Experimental Child Psychology, 182*, 114–125. https://doi.org/10.1016/j.jecp.2019.02.001

Treiman, R., Kessler, B., & Bick, S. (2003). Influence of consonantal context on the pronunciation of vowels: A comparison of human readers and computational models. *Cognition, 88*(1), 49–78. https://doi.org/10.1016/S0010-0277(03)00003-9

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology, 124*(2), 107–136. https://doi.org/10.1037/0096-3445.124.2.107

Uddén, J., Araújo, S., Forkstam, C., Ingvar, M., Hagoort, P., & Petersson, K. M. (2009). A matter of time: Implicit acquisition of recursive sequence structures. In N. Taatgen, & H. Van Rijn (Eds.), *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society* (pp. 2444-2449).

Vousden, J. (2008). Units of English spelling-to-sound mapping: A rational approach to reading instruction. *Applied Cognitive Psychology, 22*(2), 247-272. https://doi/org/10.1002/acp.1371.

Vousden, J.I., Ellefson, M.R., Solity, J. & Chater, N. (2011). Simplifying Reading: Applying the Simplicity Principle to Reading. *Cognitive Science, 35*(1), 34-78. https://doi.org/10.1111/j.1551-6709.2010.01134.x

Wittenberg, E. & Jackendoff, R. (2018). Formalist modeling and psychological reality. *Linguistic Approaches to Bilingualism, 8*(6), 787–791. https://doi.org/10.1075/Lab.18077.Wit

Wonnacott, E., Brown, H., & Nation, K. (2017). Skewing the evidence: The effect of input structure on child and adult learning of lexically based patterns in an artificial language. *Journal of Memory and Language, 95*, 36–48. https://doi.org/10.1016/j.jml.2017.01.005

Xu, F., & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, *22*(3), 531–556. https://doi.org/10.1017/S0305000900009946

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*(2), 245–272. https://doi.org/10.1037/0033-295X.114.2.245

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press

Yang, C. (2018a) A formalist perspective on language acquisition. *Linguistic Approaches to Bilingualism 8*(6), 665–706. 10.1075/lab.18014.yan

Yang, C. (2018b) Some consequences of the Tolerance Principle. *Linguistic Approaches to Bilingualism 8*(6), 797 - 809. doi.org/10.1075/lab.00022.yan

Yang, C. (2021, October) A user's guide to the Tolerance Principle [Blog post]. Retrieved from https://ling.auf.net/lingbuzz/004146

Yang, C., Crain, S., Berwick, R. C., Chomsky, N., Bolhuis, J. J. (2017). The growth of language: Universal Grammar, experience, and principles of computation, *Neuroscience & Biobehavioral Reviews, 81*(B),103-119, https://doi.org/10.1016/j.neubiorev.2016.12.023.

Yelland, G.W. (1994). The Processes of Lexical Access. In R. E. Asher (Ed.), *The Encyclopaedia of Language and Linguistics* (pp. 31-36). Pergamon Press.

Yip K. & Sussman, G. (1997). *Sparse Representations for Fast, One-Shot Learning* [Conference presentation]. National Conference on Artificial Intelligence, Orlando, FL, United States.

Zeno, S. (Ed.). (1995). *The educator's word frequency guide*. Touchstone Applied Science Associates.

Zevin, J. D., & Seidenberg, M. S. (2004). Age-of-acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory & Cognition, 32*(1), 31–38. https://doi.org/10.3758/BF03195818

Zevin, J. D., & Seidenberg, M. S. (2006). Simulating consistency effects and individual differences in nonword naming: A comparison of current models. *Journal of Memory and Language, 54*(2), 145–160. https://doi.org/10.1016/j.jml.2005.08.002

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

Ziegler, J. C., Perry, C., Jacobs, A. M., & Braun, M. (2001). Identical words are read differently in different languages. *Psychological Science, 12*(5), 379–384. https://doi.org/10.1111/1467-9280.00370

Ziegler, J., & Perry, C., & Zorzi, M. (2014). Modelling reading development through phonological decoding and self-teaching: Implications for dyslexia. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 369,*(1634), 20120397. https://doi.org/10.1098/rstb.2012.0397.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.

Zorzi, M., Houghton, G., & Butterworth, B. (1998). Two routes or one in reading aloud? A connectionist 'dual-process' model. *Journal of Experimental Psychology: Human Perception and Performance, 24*(4), 1131 -1161. https://doi.org/10.1037//0096-1523.24.4.1131