Author:
Ayravainen, Laura E M

Title:
An Empirical and Computational Investigation of Generalisation in Nonword Reading

An Empirical and Computational Investigation of Generalisation in Nonword Reading

Laura Eeva Maria Äyräväinen

School of Psychological Science

October 2022

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Life Sciences

Word count: 79,667

Abstract

Reading aloud new words requires an ability to generalise linguistic knowledge acquired via experience in reading. Yet, the exact cognitive mechanisms by which this happens are still unknown. In this PhD project, I investigated generalisation in reading aloud in English, focusing on pronunciations assigned to nonwords by skilled readers. This work consisted of computational, empirical and methodological investigations.

Firstly, I developed a new, symbolic model of reading aloud – the Weighted Segments Pronunciation (WSP) model. This model converts letter strings into speech sounds based on different statistical properties of the writing system, across varying sized print-to-sound correspondences. The WSP model simulated central tendencies in human nonword reading responses comparably to prominent computational models of reading (the DRC and the CDP++ models). Furthermore, the WSP model showed some promise in simulating variability in nonword reading, and the present work illustrated some ways to evaluate models that produce variable output. Issues in the performance of the WSP model were identified and several avenues for improving the model were discussed.

Secondly, I conducted two empirical studies, aiming to clarify which statistical properties of the writing system skilled readers are sensitive to. Both type and token frequency measures of print-to-sound correspondences were shown to be influential in nonword processing, with likely larger influence of type frequency.

Thirdly, I compared two methods of collecting information about how skilled readers process nonwords: the traditional nonword naming method (where participants read aloud nonwords) and a relatively new nonword rating method (where participants give acceptability ratings to pronunciations assigned to nonwords). These comparisons revealed that the rating method is a feasible alternative to the naming method, and it may reveal aspects about skilled readers' knowledge of print-to-sound correspondences that the nonword naming method cannot.

These findings bear relevance to future empirical investigations and theory development of reading aloud.

Acknowledgements

A PhD candidature is surely an unusual time for anyone undertaking such a challenge. But when this endeavour coincides with a global pandemic, the experience becomes all the more peculiar. Uncertainty in all its forms and feelings of isolation were undoubtedly present in my journey, both due to the nature of the beast itself and due to the external circumstances. However, several people helped me and kindly offered the necessary support for me to finish this enormous undertaking.

I thank my supervisors Prof. Colin Davis and Prof. Anne Castles for their guidance, encouragement and helpful comments. I had considerable freedom to shape this project into what it became, but I would certainly not have finished it had it not been for my supervisors steering me away from the rabbit holes I occasionally fell into.

I am also grateful for all the support and silliness from my friends, those on the same journey, those who shared their wisdom 'from the other side' and those who brought the real life back to my attention. Flakey Friends, thank you for reminding me that learning and life happens in and outside of academia. I would surely not be where I am now without all our chats in the pub and occasional road trips. I really needed to be ‘forced’ to party from time to time. Priory Road Punks, I very much appreciated your company, the peer support, and most of all the lunch breaks that were often more like sessions of stand-up comedy. It was a joy to be surrounded by such witty and imaginative individuals! All my climbing buddies, thank you for contributing to my time off being such a wholesome experience – I cannot emphasise enough what a valuable, balancing force it was during this project. DELTA girls, I am in awe of the connection and understanding we still share, despite the distance and such different life circumstances. Our chats during this time helped me keep things in perspective. I learnt to read with some of you and look where it took me! Jacob, you were my much-needed sanctuary in my era of ascetism. Thank you for your jovial, unexpected perspectives about my PhD project and about life in general.

Finally, I am forever grateful to my family in Finland for all their support, countless video calls, online boardgames and jokes. It was a gift to grow up in such a silently accepting environment, which undoubtedly was a big part of me believing in myself enough to attempt something as monumental as a doctoral degree.

Author’s declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE:…28/10/2022.....

Table of Contents

List of Tables

List of Figures

Chapter 1 : General Introduction

Reading requires an ability to connect the written form of a language with its spoken form. In complex writing systems, such as English, print-to-sound correspondences (PSCs) do not follow a simple pattern – the same orthographic segment can correspond to several different phonological segments. As a result, reading aloud new words in such writing systems is accompanied with a degree of uncertainty. Nevertheless, literate individuals manage to assign pronunciations to unknown words and nonwords (pronounceable letter strings). In doing so, not only do they convert text to speech sounds, but they also generalise their experience with separate instances of written and spoken words, producing pronunciations to letter strings they have never encountered before. The cognitive mechanism behind this ability has attracted considerable amount of research interest, and the topic has been approached both via empirical investigations and computational modelling.

Empirical evidence from studies of nonword reading suggest that readers utilise statistical properties of the writing system to read aloud new orthographic material (e.g., Andrews & Scarratt, 1998; Seidenberg et al., 1994; Siegelman et al., 2020). The kind of information readers extract from their experience with reading, however, is yet to be fully determined. Computational models of reading offer a way to test mechanisms that might be at play when skilled readers assign pronunciations to nonwords. Current computational models of reading differ in several ways, for instance, in terms of whether reading words and nonwords are considered distinct processes and in terms of the statistical properties of the writing system that influence the models' print-to-sound conversion (e.g., Coltheart et al., 2001; Perry et al., 2010, Plaut et al., 1996). As a result, each model has different strengths and weaknesses in simulating aspects of nonword reading. Comparisons of human nonword reading behaviour to output from computational models continue to reveal areas for improvement in the current models (e.g., Pritchard et al., 2012, Treiman et al., 2003).

This chapter provides an overview of the empirical investigations of reading aloud, focusing on aspects of nonword reading that have attracted considerable amount of research interest – different statistical properties of the PSCs and unit size in nonword reading. Additionally, the review of empirical findings covers an aspect of nonword reading that has been relatively neglected in the literature until recently – variability in nonword reading. Following this, a

brief description of current computational models of reading is provided, specifically regarding the aspects of nonword reading covered in the summary of the empirical work. This section ends with an overview of recent evaluations of the current computational models against human nonword reading responses. Finally, the aims of the current PhD project are outlined.

1.1 Empirical investigations of reading aloud

This section provides an overview of empirical investigations of nonword reading, particularly regarding the type of pronunciations assigned to nonwords by skilled readers. First, empirical findings regarding the role of key properties of the PSCs in nonword reading are reviewed. Following this, I turn to variability in nonword reading, a ubiquitous finding, which has only recently attracted some research interest, and attempts to incorporate this aspect of print-to-sound conversion into computational models of reading.

1.1.1 Regularity

The English writing system has been described as a quasi-regular domain – a structure that consists of systematic relationships between its elements as well as exceptions to these regularities (Seidenberg & McClelland, 1989). It would thus seem that generating new pronunciations to letter strings in such a system would require either 1) reducing the relationships it contains into a set of rules, largely disregarding the exceptions or 2) relying on the probabilistic nature of these relationships, so that reading responses broadly reflect the frequency of particular relationships in the writing system.

The former option, the rule-based approach, focuses on the regularity found in the English writing system. Although these regularities can be considered at several grain sizes, regularity in English PSCs is traditionally defined at the level of a single letter or letter cluster (grapheme) that corresponds to a single speech sound (phoneme), referred to as grapheme-phoneme correspondence rules (GPC-rules). The pronunciation associated with a given grapheme in majority of the words in which this grapheme occurs is the regular or standard pronunciation for this grapheme, such as *ea* → /i/ (as in *heal*)[1] and as such, this grapheme-phoneme pair makes up a GPC-rule. GPC-rules are also position-specific, such that a different pronunciation may be the most common for a given grapheme in the word initial

[1] Throughout this dissertation, orthographic segments will be formatted in italics and phonological segments presented in DISC phonetic character set, preceded and followed by forward slashes, e.g., *dog* (orthographic), /dQg/ (phonological). See Appendix 1, Table 1A for a list of phoneme characters in DISC and IPA, accompanied by example words for each phoneme.

position than in a middle or final position. Pronunciations deviating from this GPC-rule are irregular pronunciations, such as *ea* → /E/ (as in *head*). By extension, any word that can be pronounced correctly applying these GPC-rules is called regular (e.g., *heal*) and any word for which this is not the case is irregular (e.g., *head*). Regularity is a key concept underpinning processing in the Dual-Route Cascaded model of reading (Coltheart et al., 2001) described in Section 1.2.1.

If skilled readers are aware of and employ GPC-rules when reading aloud new words, this should be seen in the type of reading responses generated. Indeed, several studies report considerable fidelity to the GPC-rules in nonword reading (e.g., Andrews & Scarratt, 1998; Brown & Deavers, 1999; Coltheart & Leahy, 1992; Glushko, 1979; Kay, 1983, cited in Patterson & Morton, 1985), as demonstrated by a sizeable proportion of regular pronunciations assigned to nonwords for which other, plausible pronunciation options are available. However, the very same studies demonstrate that skilled readers also rely on information that goes beyond the GPC-rules, as demonstrated by the proportion of irregular pronunciations assigned to some of the nonwords. For instance, Coltheart and Leahy (1992) investigated the type of nonword naming responses given by developing and skilled readers to monosyllabic nonwords with vowel and the following consonant clusters (word bodies) that are always pronounced irregularly in existing words (e.g., *thild*, sharing a body with *child, wild* etc.). If the irregular nonwords[2] are pronounced irregularly (e.g., *thild* pronounced as /T2ld/ instead of regularly as /TIld/), this is suggestive of utilisation of word body sized segments in reading, either as a word body-rime[3] analogy to existing words, or as a PSC rule that is based on word body sized segments. Coltheart and Leahy reported that the percentage of regular pronunciations assigned to these irregular nonwords by skilled readers was 49%, thus demonstrating considerable reliance on GPC-rules in adult print-to-sound conversion. Yet, the percentage of irregular pronunciations assigned to the same items was 28%, showing that GPC-rules are not the only approach taken. If these irregular pronunciations are to be explained by PSC rules based on larger unit size, such as word bodies, it remains unclear how the choice between these, at times conflicting rules would be made.

Thus, while regularity can explain some findings about the types of pronunciations assigned to nonwords, it alone cannot capture the pattern of nonword reading responses reported in the

[2] While nonwords naturally do not have a correct pronunciation, either regular or irregular, I refer to nonwords that share an irregularly pronounced orthographic segment in existing words as 'irregular nonwords', for brevity.

[3] Rime refers to the phonological counterpart of a word body – e.g. word body *ild* can have a rime /2ld/ or /Ild/

literature. Further evidence regarding the regularity of PSCs is considered in the following section.

1.1.2 Consistency

As the regularities found in the English writing system co-exist with exceptions, the categorical distinction into regular and irregular PSCs seems insufficient for capturing the way in which the orthographic and phonological patterns relate to each other. Another useful statistical property of the writing system is consistency, used to quantify how reliably a pronunciation is associated with an orthographic segment. Throughout the dissertation, the term consistency is used to refer to the association strength of a PSC or a degree of consistency of a PSC, as described below (the proportion consistency). It is worth noting, however, that the term consistency has also been used to refer to types of words in the literature (e.g., Andrews & Scarratt, 1998; Brown & Deavers, 1999; Glushko, 1979) – those with only one pronunciation for (typically) the word body (e.g., *flame, game, name* etc.,) are consistent words and those with more than one pronunciation for the word body (e.g., *gave, save,* etc. versus *have*) are inconsistent. This categorical use of the term consistency should be clear from reference to whole words or nonwords rather than PSCs.

Two definitions of PSC consistency are described next, starting with the definition used throughout this dissertation. *The proportion consistency* of a PSC is the number of words in which the relevant orthographic segment is associated with the same pronunciation relative to all the words in which the relevant orthographic segment occurs, i.e., the number of friends (words that share the spelling and pronunciation of a given segment) relative to the number of friends and enemies (words that share the spelling but differ in pronunciation of a given segment). For instance, the PSC *ould* – /Ud/ has a consistency of .75, because out of the four monosyllabic words in which the orthographic segment *ould* occurs, three are pronounced as /Ud/ (*could, should, would*). The PSC *ould* – /5ld/ has a consistency of .25, as only the word *mould* contains this PSC. This measure of consistency was used in Treiman et al. (1995, Analysis A), and it can be used as is, in which case only the number of words included in the calculations matters, or it can be used weighed by the frequency of occurrence of the words included in the calculations (these options are considered further in Chapter 2, Section 2.3.1). *The entropy H consistency* of a PSC takes into account the number of different pronunciations associated with the given orthographic segment and the similarity of the

probabilities (the proportion consistencies) of different pronunciations associated with the orthographic segment. The measure can be calculated with the following formula:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

where p_1 is the probability of the first pronunciation option, p_2 the probability for the second pronunciation option (if one exists) and so forth up to the number of different pronunciation options. Fully consistent PSCs have an H value of 0, which increases the more uncertainty there is about a pronunciation for an orthographic segment. For instance, the word body *ould* has an H value of 0.81, with two pronunciation options, with probabilities of 0.75 and 0.25. By contrast, another orthographic segment with also two pronunciation options, but probabilities of 0.5 and 0.5 would have a higher H value of 1. This is because the options are equiprobable, thus increasing the uncertainty associated with choosing between them. The entropy H was first introduced in information theory (Shannon, 1948) and has been used as a measure of consistency of PSCs and as a measure of variability in nonword reading responses (e.g., Andrews & Scarratt, 1998; De Simone et al., 2021; Siegelman et al., 2020; Treiman, et al., 1995, Analysis B). As seen in these definitions, consistency of PSCs is a matter of degree, unlike regularity, which is categorical measure. Consistency of PSCs is a central feature of connectionist models of reading (e.g., Plaut, McClelland, Seidenberg & Patterson, 1996), some of which are described below in the Section 1.2.3.

The idea that skilled readers are sensitive to the consistency of PSCs suggests that skilled readers are aware of several, competing pronunciations for the same orthographic segment, such as that the letter *a* is often pronounced as /{/ as in *cat* but also as /#/ as in *fast*.

Evidence supporting the notion that consistency of PSCs plays a role in human print-to-sound conversion comes from studies reporting longer naming latencies and more variability in nonword responses for inconsistent words and nonwords compared to consistent ones (e.g., Andrews & Scarratt, 1998; Glushko, 1979; Seidenberg et al., 1994). Furthermore, pronunciations assigned to nonwords tend to be pronunciations that are most consistently associated with the orthographic segments of the nonwords – either on a grapheme or a word body level (e.g., Andrews & Scarratt, 1998; Brown & Deavers, 1999; Glushko, 1979). For instance, Seidenberg et al. (1994) report a nonword naming study in which participants assigned non-standard pronunciations to some of the 590 nonwords they read aloud. Furthermore, when a nonword received several different pronunciations, the less common

options (i.e., pronunciations given by a smaller number of participants) had longer naming latencies than the more common options. The naming latencies of nonwords also increased as the number of different naming options increased (i.e., when a nonword was pronounced the same way by all participants, it was pronounced faster than the most popular pronunciation option for a nonword that received several different pronunciations). Seidenberg and colleagues interpret this increase in naming latencies as competition between different pronunciation options that skilled readers have experience of.

However, consistency of PSCs alone does not explain the patterns of empirical findings in nonword reading. Andrews and Scarratt (1998, experiment 2) also compared regular-consistent nonwords (nonwords with word bodies that are always pronounced regularly in existing words, e.g., *beal*) to inconsistent nonwords (nonwords with word bodies that are sometimes pronounced regularly and sometimes irregularly in existing words e.g., *basp*), and found increased response latencies and response variability for inconsistent items over regular-consistent items. However, Andrews and Scarratt also collected naming responses to nonwords with word bodies that are consistently pronounced irregularly in several existing words (Irregular-many items, e.g., *nalm* based on *palm*, *calm* etc.) or in a single existing word (Irregular-single items, e.g., *sonth* based on *month*). While these items are also consistent, the naming latencies and the variability of the naming responses for these items were higher than those to the inconsistent items. A potential explanation for this pattern of findings may be in part the regularity-irregularity distinction – skilled readers have a strong tendency to assign regular pronunciations to nonwords whenever possible, and when an irregular pronunciation is assigned to a nonword, this comes with additional processing cost, reflected in increased naming latencies. Another explanation would rely on consistency, but at different grain sizes – the regular-consistent nonwords are consistent both at the level of graphemes and at the level of word bodies, whereas the irregular-consistent nonwords are inconsistent at the level of graphemes and consistent at the level of word bodies. The latter explanation relates to a topic that will be discussed in more detail in the Unit size section (Section 1.1.4). Andrews and Scarratt also report the incidence of regular and irregular pronunciations assigned to the different groups of nonwords. The regular-consistent items and inconsistent items were predominantly pronounced according to the GPC-rules or, in other words, the most consistent grapheme-sized PSCs (regular-consistent: 93%; inconsistent: 87%), whereas irregular-consistent nonwords were mostly pronounced according to the most consistent body sized PSCs (65% of responses). The only group of items that did not conform to this pattern were

the irregular-single items, where the most consistent body sized PSC was used in 40% of the responses and most consistent grapheme-sized PSC in 41% of the responses. Thus, interpreting these findings as consistency of PSCs considered at different grain sizes also leads to the question of how is the grain size chosen? While a large part of these findings could be explained in terms of consistency, something else is needed for the full picture.

The role of regularity and consistency of PSCs has not always been easy to tease apart – studies demonstrating longer naming latencies for irregular words than regular words (e.g., Baron & Strawson, 1976; Waters & Seidenberg, 1985) do not rule out the possibility that these regularity effects are in fact consistency effects. This is because many irregular words are also inconsistent. However, evidence from word naming studies with more carefully controlled experimental stimuli suggest that these two properties are separate (e.g., Andrews, 1982; Jared, 2002). For instance, Andrews (1982, Exp. 2A) inspected naming latencies for words that were either regular-consistent, regular-inconsistent, irregular-consistent or irregular-inconsistent. The most important findings from this factorial design were the main effects of regularity and consistency in the naming latency data, such that longer latencies were seen for irregular items than regular items, matched in consistency and, similarly, longer naming latencies were found for inconsistent items over consistent items, matched in regularity. Thus, it appears that both regularity and consistency play a role in word naming.

In summary, consistency of PSCs clearly has an influence on nonword reading – the type of naming responses, the variability of naming responses, and the speed at which naming responses are given to nonwords largely reflect the consistency of PSCs. However, some findings are difficult to explain solely in terms of consistency (Andrews & Scarratt, 1998, Exp. 2).

1.1.3 Frequency

The difference in rule and probability-based approaches to reading English can also be seen in measures of frequency of PSCs, that is, the prevalence of PSCs in a language. The frequency of a given PSC can be quantified based on types – i.e., the number of different words that contain the given PSC, or based on tokens – i.e., the number of times the given PSC occurs in a corpus, regardless of whether the words that embody this PSC are the same or different. Thus, type frequency captures the exposure to a given PSC as it occurs in distinct linguistic units, while token frequency captures the overall exposure to a given PSC. The rule-based approach would only take into account the most frequent PSCs, whereas the

probability-based approach would also consider PSCs with lower (type) frequencies. The question of whether type or token frequency would better capture the knowledge skilled readers extract from their experience with language is still open – for instance, would a PSC *ave* → /1v/ be a more reliable correspondence for skilled readers, because they have encountered it in many words (*pave*, *cave*, *crave*) or would a PSC *ave* → /{v/ be more reliable as skilled readers encounter this correspondence very often in a single word *have*?

The importance of token frequency is well documented in visual word recognition and word naming studies – words with high token frequency are recognised faster and more accurately than words with lower token frequencies (e.g., Balota, et al., 2004; Forster & Chambers, 1973; Howes & Solomon, 1951). However, attempts at teasing apart the relative importance of type and token frequency in consistency[4] effects of reading aloud, that is, the finding that consistent words are read aloud faster and more accurately than inconsistent words, has produced mixed results (Jared, et al., 1990; Treiman et al., 1995).

In nonword reading, the role of token frequency has been contrasted with that of type frequency, and most results point to the importance of type frequency over token frequency (Andrews & Scarratt, 1998; Johnson, 1970; Kay, cited in Kay & Marcel, 1981). Perhaps the most thorough investigation comparing the role of type frequency with the role of token frequency in nonword reading was carried out by Andrews and Scarratt (1998). In their second experiment, participants read aloud nonwords with word bodies that were either regular-consistent (e.g., *beal*), inconsistent (e.g., *heaf*), or irregular-consistent, so that the nonword's body either occurred in only one existing word (irregular-single, e.g., *sonth*) or in several words (irregular-many, e.g., *dask*). The proportion of regular pronunciations assigned to nonwords was regressed with a number of properties, such as the proportion of regularly pronounced word body neighbours and consistency of different segments of the nonwords. The consistency measure used was entropy H (see Section 1.1.2). Two separate analyses were run, where the consistency measures and the proportion of regular neighbours were computed either as a type-based metric (number of regularly pronounced neighbours / number of all neighbours) or as a token-based metric (where the number of neighbours was multiplied with the token frequency of the neighbours). Most importantly, these analyses revealed that the best predictor for the proportion of regular pronunciations assigned to nonwords was the proportion of regular body neighbours (accounting for 23% of unique variance), followed by word body consistency (accounting for 2% of unique variance). Crucially, the overall

[4] Here consistency refers to two types of words or nonwords: consistent and inconsistent (see Section 1.1.2)

goodness of fit of the regression models was higher when type-based measures of consistency and proportion of regular neighbours were used (R^2 = .69), compared to the token-based measures (R^2 = .62). As such, although both types of frequency measures were found useful, these results suggest that type frequency is more important in nonword processing than token frequency. It is worth noting, however, that only a handful of studies have investigated the issue directly.

1.1.4 Unit size

As briefly mentioned in Section 1.1.2, the statistical properties of PSCs depend on the unit size of PSCs. How a letter string is parsed can change the most consistent pronunciation associated with each segment of the letter string. Investigations into the properties of English PSCs often emphasise the importance of word body sized segments (Kessler & Treiman, 2001; Stanback,1992; Treiman et al., 1995). While consonants can typically be pronounced utilising a simple one-to-one mapping, most vowels have several alternative pronunciations. As the consonants following a vowel often help in determining how the vowel is pronounced, word body sized segments can serve as a particularly useful PSC unit in reading. This idea is supported by an analysis of monosyllabic words carried out by Treiman et al. (1995, Analysis A). The authors analysed a total of 1329 items with a single consonant – vowel – single consonant (CVC) phonemic structure. Each item was divided into initial consonant (i.e., onset), vowel (i.e., nucleus) and final consonant (i.e., coda) segments, as well as larger segments of onset + nucleus (i.e., antibody)[5] and nucleus + coda (i.e., the word body). In the type-based analysis, the consistency of these segments was calculated as the number of words with the given orthographic segment that were pronounced the same way relative to the total number of words with the given orthographic segment. This analysis revealed that the mean consistency of initial and final consonants was over 90% while the vowel consistency was only 62%. This demonstrates how the unpredictable nature of GPCs in English is largely due to the vowels. Most importantly, the antibody segment had a mean consistency of 55%, whereas the body segment had a mean consistency of 80%. Thus, considering the consonants following the vowel provides more consistent PSCs than the vowel alone or considering the consonant preceding the vowel. Using conditional consistency measures, Kessler and Treiman (2001) investigated how much more consistent the PSCs for an onset, nucleus or coda of a monosyllabic word become when other parts of the syllable are taken into account. Their first analysis of a total of 3117 words revealed that out of 68 possible vowel letter

[5] I borrowed this term from Forster and Taft, 1994

strings, 39 were consistent. Of the remaining 29, 23 were significantly more consistent when the coda was considered as well. By contrast, only two vowel segments became more consistent when the onset was taken into account (*w* and *qu* preceding *a*). Using unconditional consistency measures as in Treiman et al. (1995), Kessler and Treiman arrived at comparable consistency measures, confirming the finding that word bodies are more consistent than antibodies, even though onsets and codas alone are equally consistent. The structure of the English writing system would thus encourage adopting a larger grain size reading style, where word body sized units are particularly important.

The importance of word body sized segments has been demonstrated with visual word recognition and word naming tasks in skilled readers of English (e.g., Bowey, 1990; Treiman et al. 1995). For instance, in their large-scale word naming study, Treiman et al. (1995, Part 2) found that higher consistency of onset and word body segments was associated with faster response times (RTs) and lower error rates, while consistency of other segments, such as the vowel alone or the antibody segment did not have as reliable influence on word naming performance (although higher consistency of the vowel segment alone was associated with lower error rates).

Larger unit sizes have also been shown to be relevant in nonword naming (Andrews & Scarratt, 1998; Johnson & Venezky, 1975; Ryder & Pearson, 1980; Taraban & McClelland, 1987, Experiment 3; Treiman & Zukowski, 1988). For instance, Taraban and McClelland (1987, Experiment 3) demonstrated that pronunciation assigned to a nonword can be primed by a preceding word, depending on the type of orthographic overlap the nonword and the prime share. The nonwords in the experiment shared an orthographic segment with an exception word (e.g., *come*), so that the overlap between the two items was either the antibody (e.g., *coze*), the body (e.g., *zome*) or the vowel (e.g., *vole*) segment. In a speeded reading task, these nonwords were preceded by primes that were either irregular (e.g., *come*) or regular (e.g., *home*). Taraban and McClelland found that the percentage of regular pronunciations assigned to the antibody-nonwords preceded by regular and irregular primes were 89% and 79%, respectively. For the body-nonwords, these percentages were 96% and 63%. The vowel-nonwords were not affected by primes. These findings suggest that both antibody and word body segments can influence nonword reading, although the latter yields larger effects. Similarly, Treiman and Zukowski (1988) created nonwords that overlapped with an exception word, such as *friend*, in the antibody (e.g., *frieth*), body (e.g., *chiend*) or vowel segment (e.g., *chieth*). Naming responses to these nonwords showed, most

importantly, that more pronunciations congruent with the exception words were assigned to body-nonwords than to antibody- or vowel-nonwords (e.g., *chiend* read as /JEnd/ more often than *frieth* or *chieth* read as /frEth/ or /JET/). Treiman and Zukowski suggest that one reason for the special role of word bodies in nonword reading might be implicit or explicit knowledge about how post-vocalic consonants can alter the pronunciation of the vowel, whereas pre-vocalic consonants rarely affect the vowel pronunciation.

However, some evidence for the importance of antibody segments has also been reported (e.g., Schmalz et al., 2014**;** Treiman et al., 2003). For instance, a study by Treiman, et al. (2003) found that skilled readers are sensitive to regularities in both antibody and body segments of monosyllabic words. Informed by their analysis on statistics of English, their nonword stimuli were constructed using only antibody and body segments where the onset or coda altered the vowel pronunciation to a non-standard vowel (e.g., the vowel in the antibody *wa* in *wasp* is pronounced as a non-standard vowel /Q/ as opposed to the standard vowel /{/ and the vowel in the body *ead* in *head* is pronounced as /E/ as opposed to the standard vowel /i/). The critical nonwords (e.g., *wasb, clead*) were pronounced reliably more often with a non-standard vowel compared to control nonwords (e.g., *trabs*, *cleam*). The authors suggest that previous mixed results for the role of antibody segments is due to the fact that regularities between onset and nucleus are rare in English, not because readers are not sensitive to them.

Finally, utilising larger unit size in reading, especially word bodies, seems to become more prevalent as reading skills develop: when children with lower reading age read irregular nonwords, they produce less word body analogy-based pronunciations than children with more developed reading skills or adults do (Coltheart & Leahy, 1992; Brown & Deavers, 1999; Steacy et al. 2019). Coltheart and Leahy (1992) investigated nonword reading with nonwords that shared a word body with regular existing words or irregular words (where most words with the given word body are pronounced irregularly)[6], referred to as regular and irregular nonwords, respectively. The percentage of irregular pronunciations assigned to the irregular nonwords by 1st grade children, 3rd grade children and adults were 10%, 23% and 28%, respectively. Notably, the percentages of regular pronunciations assigned to the irregular nonwords were 38% (1st graders), 43% (3rd graders) and 49% (adults). Thus, while word body units seem to become more important with increasing reading skills, utilisation of

[6] The study also included ambiguous nonwords, with word bodies that are pronounced regularly in some existing words and irregularly in others. This was excluded from the description for brevity.

grapheme-phoneme sized segments is still quite prevalent even in skilled readers. However, as pointed out by Brown and Deavers (1999), the regular nonwords in Coltheart and Leahy's study had significantly more orthographic friends (words with the same word body and pronunciation) than the irregular nonwords. As demonstrated by Andrews and Scarratt (1998, described in detail in the Section 1.1.3), the number of regularly pronounced neighbours is one of the most important predictors of the proportion of regular pronunciations assigned to a nonword.

With this limitation in mind, Brown and Deavers (1999, Experiment 1) investigated how children (aged five to nine years) and adults read regular-consistent and irregular-consistent nonwords when the number of orthographic friends for these items was controlled. The children were divided into skilled readers (mean reading age of 11 years and 6 months) and less skilled (mean reading age of 8 years and 8 months) based on their reading ability test score. The percentage of reading irregular nonwords as an analogy to the irregular words they resembled by children with less developed reading skills, children with more developed readings skills and adults were 39% 53% and 58%, respectively. The proportion of regular pronunciations for the irregular nonwords by the three groups were 50% (less skilled), 44% (more skilled) and 41% (adults). These studies demonstrate two important aspects about reading. Firstly, it is quite common for adult skilled readers to utilise grapheme sized segments in nonword reading. Secondly, considering larger orthographic segments becomes more prevalent as reading skills develop.

While this pattern of development might show how more experience with different spelling patterns in English encourages adopting a reading style that takes into account varied sized PSCs, it may also be a reflection of how reading is taught in schools. Indeed, the type of reading instruction children are exposed to in school seems to be linked to the way they read unknown words: Deavers and colleagues (2000) compared the incidence of reading nonwords irregularly (as a word-body analogy to irregular words) between groups of children that had received reading instruction focusing on word body sized units (Word Body group) and children that had received instruction where grapheme-phoneme sized segments were central (Grapheme-Phoneme group). In a task where the nonwords were read aloud in isolation, when vocabulary knowledge of the irregular words that the nonwords were based on was taken into account, the Word Body group did produce reliably more irregular pronunciations (31%) than the Grapheme-Phoneme group (19%). Deavers et al. discuss the benefits of using grapheme-phoneme sized segments in early stages of reading, when many words are still

unknown, whereas relying on word-body sized segments should be more successful when a reader's vocabulary is large enough to accommodate efficient use of word body-based analogies. Interestingly, the influence of reading instruction received in childhood seems to also have ramifications to the way individuals continue reading unknown words as adults (Thompson et al., 2009). Thompson and colleagues compared nonword reading performance of adults who had received reading instruction focusing on grapheme-phoneme correspondences in school (phonics group) to that of adults who had not received such reading instruction in childhood (non-phonics group). Both groups of participants read aloud regular (e.g., *beal*), regular-inconsistent (e.g., *dush,* where word the body is either pronounced regularly or irregularly in existing words) or irregular-consistent (e.g., *bealm*) nonwords. The phonics group produced fewer irregular pronunciations to regular-inconsistent and irregular-consistent nonwords and more regular pronunciations to these items than did the non-phonics group.

In summary, empirical evidence of reading clearly points to the importance of word body sized segments and grapheme sized segments in print-to-sound conversion. The role of the antibody segments is less clear (Taraban & McClelland, 1987, Experiment 3; cf. Treiman & Zukowski, 1988), but likely serves as a useful unit in reading in limited cases (Treiman et al., 2003). Furthermore, the unit size of print-to-sound conversion seems to be influenced by reading instruction in schools, which may have a long-lasting effect on reading behaviour.

1.1.5 Variability in nonword reading

Nonword reading task has been widely employed in investigations tapping into print-to-sound conversion. A robust finding in this line of research is that nonword reading is highly variable: the same item can be pronounced in several different ways by different participants (Andrews & Scarratt, 1998; Mousikou et al., 2017; Pritchard et al., 2012; Seidenberg et al., 1994) and the same nonword can be pronounced in different ways by the same participant (Ulicheva et al., 2021). Furthermore, the same sub-lexical unit, such as a word body, in different nonwords can be pronounced in different ways by the same participant (for instance, in Pritchard et al., 2012 data set). As an example of the between-participants variability in nonword reading, Pritchard et al. (2012) reported that for their 412 nonwords, read aloud by 45 participants, the number of different pronunciations for a single nonword ranged from 1 to 24. Similarly, Mousikou et al. (2017) reported 1-22 different pronunciations assigned to each of their 915 disyllabic nonwords.

Coltheart and Ulicheva (2018) investigated the potential sources of variability in nonword reading using the Pritchard et al. (2012) data set. They concluded that skilled readers differ in how they parse a letter string into graphemes as well as how they assign phonemes to the graphemes. This analysis was carried out with the assumption that nonwords are parsed into grapheme sized segments, and graphemic parsing was assessed based on pronunciations produced for each item. However, as discussed in the previous section (1.1.4), skilled readers seem to parse letter strings in more varied ways, such as into onset and word body sized segments. As such, the unit size of the orthographic parsing can also influence phoneme assignment. Considering the studies demonstrating that early reading instruction can influence the way nonwords are named in childhood and adulthood (Deavers et al., 2000; Thompson et al., 2009, described in Section 1.1.4), it appears that the global tendency for parsing letter strings into larger or smaller segments is, at least in part, a systematic source of variability in nonword reading. Phoneme assignment, on the other hand, may vary in a less systematic fashion. Given that individuals differ in their vocabularies and experience in reading, it is likely that individuals also differ in their personal PSC knowledge, an idea brought forward by Seidenberg et al. (1994).

Furthermore, the individual PSC knowledge or tendency to parse letter strings in certain ways do not seem to completely determine how nonwords are named by an individual. Instead, naming responses may be influenced by the context in which nonwords are named. Several studies have demonstrated priming effects in nonword reading (e.g., Taraban & McClelland, 1987; Rosson, 1983), suggesting that pronunciations assigned to nonwords are susceptible to the influence of previous stimuli. For example, Rosson demonstrated how naming ambiguous nonwords is influenced by preceding word primes: *louch* was pronounced more often regularly (like *couch*) when it was preceded by a prime *sofa*, whereas a preceding prime *feel* would increase the incidence of irregular pronunciation (like *touch*) of the nonword.

Additionally, nonword reading responses have been shown to vary by list context, either as the type of pronunciations given (Brown & Deavers, 1999; Glushko, 1979) or as differences in response times (Rastle & Coltheart, 1999, Exp. 2; Zevin & Balota, 2000). These findings are often interpreted as different reading strategies adopted by skilled readers. For instance, Brown and Deavers (1999, Exp. 4) showed that skilled adult readers assign irregular pronunciations to irregular nonwords more often if these nonwords are presented intermixed with exception words (different than the exception words the nonwords were based on) than when the same nonwords are read aloud intermixed with other nonwords. However, other

studies have failed to find reliable influence of list context on types of pronunciations assigned to nonwords (Andrews & Scarratt, 1998).

These contextual influences on nonword reading responses demonstrate that some of the within-participants variability in nonword reading (i.e., the same participant pronouncing the same nonword or sub-lexical segment in different ways) may be relatively systematic.

However, there also appears to be unpredictable within-participants variability, as suggested by findings from a recent study by Ulicheva et al. (2021). In this study, 22 participants read aloud 50 disyllabic nonwords from Mousikou et al. (2017) materials, in five different testing sessions (the same 50 nonwords in each session). It was found that some nonwords elicited more varied naming responses than others and that some participants gave more varied naming responses than others. Using linear mixed effects regression analysis, Ulicheva and colleagues investigated whether characteristics of the items (item variability) and/or characteristics of the participants (literacy skill) might explain the variability in the nonword reading responses. The variability in nonword naming was operationalised as an entropy H of each participants' response to each nonword, across all five testing sessions. Only item consistency measure (averaged entropy H for each grapheme within a syllable) predicted naming variability, whereas literacy skill (average spelling and vocabulary test scores) did not.

Ulicheva and colleagues' analysis focusing on session-to-session changes in nonword naming responses revealed that new pronunciations for a given item were less likely in the later sessions compared to the earlier sessions. In this analysis, session number and item variability were reliable predictors for whether new pronunciations were used or not, whereas literacy skill was not. Furthermore, comparing the dissimilarity between naming responses for the 50 nonwords between each participant pair revealed that the naming responses between participants became more similar throughout the sessions.

Most importantly, Ulicheva et al. (2021) thus demonstrate an association between item characteristics and naming variability, but fail to find one for participant characteristics and naming variability. Furthermore, the naming responses tended to stabilize across sessions, which bears relevance to how one interprets naming responses from single session studies. Ulicheva and colleagues suggest that the variability in nonword reading reported in previous studies has been taken as between-participants variability, even though some of this

variability may be within-participants variability, which cannot be differentiated from between-participants variability in single session studies.

Overall, these studies demonstrate that there is considerable variability in nonword reading. Some sources of this variability have been identified, but more research is clearly needed to uncover how much of the variability from these – and perhaps additional – sources is systematic. As nonwords or sub-lexical segments with multiple plausible pronunciations tend to produce more varied reading responses, and the same participant can produce several different pronunciations for the same item, there seems to be an element of randomness in orthographic parsing and phoneme assignment. These "stochastic processes that occur within individuals" (Ulicheva et al., 2021, General Discussion section, para. 4) will be returned to in Chapters 2 and 3, where attempts to simulate variability in nonword reading is described (Section 2.1.2) and tested (Section 3.3).

1.2 Computational modelling of reading aloud

In recent decades, theories of reading have increasingly been implemented as computational models (e.g., Coltheart et al., 1993; Coltheart et al., 2001; Harm & Seidenberg, 1999; Norris, 1994; Perry et al., 2010), which have several benefits compared to verbal theories. Firstly, this form of conceptualising the cognitive process of reading requires explicit expression of every aspect of the process. Secondly, as computational models simulate human reading behaviour, they have the potential to generate new predictions about reading that have not yet been considered or tested in human participants. Thirdly, simulations of human reading also allow testing and comparing theories more directly.

Current computational models of reading are typically divided into symbolic and connectionist models, which broadly speaking differ in how the linguistic information is represented, processed and acquired. Symbolic models contain representations of words or sub-lexical units of words, which are processed following explicit principles, such as rules or decision trees (e.g., Coltheart, et al., 2001; Norris, 1994). The architecture of symbolic models is specified by the modeller, and typically these types of models do not learn from experience (cf. Coltheart et al., 1993; Pritchard et al., 2016). Connectionist models represent linguistic information as patterns of activation in connections between processing units (e.g., Harm & Seidenberg, 1999; Perry et al., 2010). A connectionist model is trained with input and desired output pairings (i.e., the spelling of the word and the phonemic transcription of the word), during which the weights of the connections between units are adjusted based on

the difference between the model's actual output and the desired output, using, for instance, a backpropagation algorithm (e.g., Seidenberg & McClelland, 1989).

Both symbolic and connectionist approaches to modelling reading can offer valuable insights to the topic. For instance, the connectionist models include accounts on developing reading skills. However, it is not easy to determine what exactly has been learnt by a connectionist model and what exactly is the process of reading aloud in these models, that is, "how the trained network has been structured by the learning algorithm so as to be able to perform the task it has learned" (Coltheart, et al., 2001, p. 205). By contrast, symbolic models provide a clearer account of the exact process involved in reading, yet often lack mechanism for learning.

Most computational models of reading produce both naming responses and reaction times to word or nonword stimuli. Numerous effects of naming latency are reported in empirical work, such as regularity and consistency effects (e.g., Andrews, 1982; Seidenberg et al., 1994) and relatively successful simulation of them achieved by different computational models (e.g., Coltheart et al., 2001; Perry et al. 2010)[7]. While these findings are informative about the processes of reading, the current work focuses only on the type of naming responses given, not on how long it takes for participants or models to produce them.

Four contemporary computational models of reading aloud are described next, particularly regarding how they function in relation to the aspects of nonword reading addressed by empirical work in reading aloud research. Two of these models, the dual-route cascaded model (Coltheart et al., 2001) and the connectionist dual process model (Perry et al., 2010) were chosen because they are widely studied and publicly available. As such, these models' output to different nonwords can be compared to that of human participants and a new model, the Weighted Segments Pronunciation (WSP) model, developed as part of the current PhD project (described in Chapter 2). The third model described, a parallel distributed processing model of reading (Plaut et al., 1996, Simulation 1) is an example of another connectionist approach taken to modelling reading. The fourth model, multiple-levels model of reading (Norris, 1994) is described due to the similarities the approach taken in this modelling work bears to the WSP model.

[7] Regularity and consistency effects refer to shorter naming times for regular or consistent words than to irregular or inconsistent words.

1.2.1 The Dual-Route Cascaded (DRC) model

At its core, the DRC model (Coltheart et al., 2001) assumes two parallel routes for converting letters into speech sounds: the lexical route – based on word specific knowledge, and the nonlexical route – based on grapheme-phoneme correspondence rules (GPC-rules). Both processing systems receive input from the letter system, and both routes influence activation in the phoneme system. The activation in the phoneme system determines the final pronunciation the model gives when the activation of all the relevant phonemes reaches a pronunciation threshold. The GPC-rules of the nonlexical route are central to the predictions the DRC model makes about nonword reading: because for each grapheme only the most common, position specific PSC is available, items read via nonlexical route will always be pronounced regularly – i.e., following the standard, mostly context insensitive pronunciation for each grapheme. As a result, the DRC model's prediction about the unit size of PSCs is that only grapheme-phoneme sized units are employed in nonword reading. Note, however, that some of these graphemes correspond to word body sized segments, e.g. *augh* pronounced as /9/ or *igh* pronounced as /2/ (See the full set of GPCs in Rastle & Coltheart, 1999, Appendix B).

Regularity of PSCs is an integral part of the model, as the assembly of pronunciations via the nonlexical route happens solely based on regular PSCs. By extension, the role of type frequency of PSCs is represented in the model in an all-or-nothing fashion: if the given PSC is the most common one (i.e., has the highest type frequency), it is the pronunciation the model assigns to any nonword with this grapheme. The less common PSCs do not affect the pronunciations (e.g., *a* → /#/ as in *bath* would not be an option since *a* → /{/ as in *cat* has the highest type frequency). The same all-or-nothing principle applies to how consistency influences nonword reading because the PSC with the highest type frequency is also the PSC with the highest consistency value – less consistent PSCs are not available to the model. Although token frequency of words affects the DRC's lexical route, this property does not play a role in pronunciations assigned to nonwords. Finally, the same letter string presented to the DRC model will always be pronounced the same way – that is, there is no variability in the model's reading output.

1.2.2 The Connectionist Dual Process (CDP++) model

The CDP++ is a recent version of the CDP family of connectionist models (Perry et al., 2010). The model also assumes that reading known and unknown words are two parallel

processes. The lexical processing route of the model consists of localist representations of known words, almost identical to the one in the DRC model, and the sublexical processing route is implemented as a two-layer assembly network. The influence of both lexical and sublexical processing come together in the phonological output buffer, where the activation of phonemes is the summed activation produced by the two processing routes. The model learns statistical relationships between orthography and phonology from exposure to a vocabulary that consists of both monosyllabic and disyllabic words and from a pre-training of simple PSCs. During training, the connection weights in the network are adjusted according to a learning algorithm called the delta rule.

To summarise which properties of interest (i.e., regularity, consistency, frequency, unit size and variability in nonword reading) CDP++ is sensitive to, I consider properties that are known due to the network architecture or learning mechanisms specified by the modellers, as well as properties that could be inferred from the model's performance against human data sets. Firstly, the emerging PSCs as a result of training are not restricted to the most common mappings nor to only grapheme-phoneme sized correspondences. The efficiency with which different correspondences are learnt during the training of the network is weighted by the normalised log frequency of each word. As such, the relative strength of the learnt PSCs should show sensitivity to token frequency. The CDP++ is also sensitive to consistency of PSCs (Perry et al., 2010, Appendix D.3), which is demonstrated, for instance, by the proportion of regular pronunciations the model gives to Andrews and Scarratt's (1998, Exp. 2) regular-consistent items (.90) compared to regular-inconsistent items (.83)[8]. Sensitivity to type frequency of PSCs in the CDP++ model's print-to-sound conversion is exemplified by the model's performance in naming nonwords with word bodies that are always pronounced irregularly (Andrews & Scarratt, 1998, Exp. 2). I categorised the vowel pronunciations as regular or irregular following Andrews and Scarratt's definition (1998, Appendix B). Items with several irregularly pronounced word body neighbours were all pronounced irregularly by the CDP++ model, whereas only half of the items with word bodies occurring in only a single irregularly pronounced word were pronounced irregularly. This sensitivity to type frequency also implies at least some sensitivity to regularity, although this feature is not as central in the model's print-to-sound conversion as it is in the DRC model, given that

[8] This simulation was run by myself, and I classified the model's responses following the definition of regularity used in Andrews and Scarratt, 1998

grapheme-sized PSCs are not the only ones available to the CDP++. Finally, there is no variability in the model's output, as the same nonword is always pronounced the same way.

1.2.3 The Parallel Distributed Processing (PDP) models of reading aloud

The PDP models of reading are a group of connectionist models based on the same theoretical assumptions put forward by Seidenberg and McClelland (1989). Unlike the DRC and the CDP++ models, which are referred to as dual-process models, the PDP models of reading do not reflect the notion of two, qualitatively different processes involved in reading aloud. Instead, all letter strings, from known regular and exception words to nonwords, are read aloud using a single mechanism, supported by the same, inter-connected processing units. While the most recent models within the PDP framework for modelling reading include orthographic, phonological and semantic levels (e.g., Harm & Seidenberg, 2004), a brief summary of an earlier model, focusing on orthography to phonology conversion is provided as an example of the PDP models of reading. This model is chosen for two reasons – firstly, it is more widely studied than the more recent models and secondly, more direct comparisons to other models focusing on print-to-sound conversion can be made.

This model, developed by Plaut et al. (1996, Simulation 1), consists of three layers of processing units: the orthographic layer, the hidden layer and the phonological layer. The orthographic units represent graphemes, and the phonological units represent phonemes. Both the orthographic and the phonological units are arranged into onset, vowel and coda clusters of units, and the units within a cluster into ordered sets of units, so that only orthographically or phonotactically legal sequences are allowed. This approach allows the model to represent the relative positions of graphemes and phonemes. The feedforward connections between units in each layer of the network go from each grapheme unit to each hidden unit and from each hidden unit to each phoneme unit.

The model was trained using 2998 monosyllabic, mostly monomorphemic words. During training, the weights of the connections were adjusted using the backpropagation algorithm, based on how similar the model's output (the activation pattern in the phoneme layer) was to the correct pronunciation of each word. Thus, the weight changes during training aimed to optimise the model for correctly pronouncing the training item in question. Words with overlapping orthography and phonology (e.g., *cave* and *pave*) would activate the same units in the network and cause similar weight changes that are beneficial for the correct pronunciation of these words. By contrast, a word with similar orthography but different

pronunciation (e.g., *have*) would steer the weight changes away from it and towards weights that are beneficial for correct pronunciation of this particular item. Furthermore, words with higher token frequency influence the model's learning more than words with lower token frequency. As such, the model is sensitive to consistency and token frequency of PSCs. This dynamic would suggest that type frequency is also influential, but to a lesser extent than token frequency. Indeed, the tendency for the model to use the most frequent PSCs in nonword reading is exemplified by the type of responses the model gave to Glushko's (1979) inconsistent nonwords: "all of the irregular responses to inconsistent nonwords matched some other pronunciation in the training corpus for the same body, with half of these being the most frequent pronunciation of the body" (Plaut et al., 1996, p. 70). This finding also suggests that the unit size of PSCs learnt by the model are not restricted to grapheme-phoneme sized units, but also exhibit mapping of larger segments, such as word bodies. As consistency over varied unit sizes and token frequency have an important influence on the model's print-to-sound conversion, regularity, which is closely linked to type frequency and smaller unit sizes, plays less of a role in this model. Finally, variability in reading letter strings is not represented in the model, as the model's output is always the same for the same letter string.

1.2.4 Multiple-levels Model of Reading Aloud

According to the multiple-levels theory, reading a letter string involves parallel analysis of PSCs on multiple levels (i.e., different sized units of the letter string), which each contribute to the final reading aloud response by facilitating compatible units and inhibiting conflicting units (Shallice & McCarthy, 1985, cited in Norris, 1994). I first describe an implementation of this approach as a symbolic model (Norris, 1994), and then briefly as a connectionist model. In the symbolic model, the relevant units of analysis for nonword reading are onset, nucleus, coda, a combination of onset and nucleus (antibody) and a combination of nucleus and coda (body). The model has a set of PSC rules, based on the aforementioned units (e.g., vowel cluster *ea* pronounced as /i/), which were derived from 2897 monosyllabic, mostly monomorphemic words (used in Seidenberg & McClelland, 1989). Each orthographic unit with multiple corresponding pronunciations is arranged in the model's rule set by frequency (either type or token frequency can be used).

Reading a letter string starts with parsing it into the units described above and the pronunciation rules are then applied based on the following principles: 1) larger units are

preferred over smaller ones, and 2) an agreement between two complementary units is sought before considering rules based on smaller units. For instance, if antibody and body segments agree on the vowel pronunciation, the antibody and body rules are applied in pronouncing the letter string. If the vowel pronunciations differ between the two rules, less frequent antibody and body rules are searched for a matching vowel pronunciation between them. If none are found, a rule based on a smaller unit (onset) and body are then applied to produce the model's pronunciation. As such, the multiple-levels model employs PSCs of varying unit sizes. The model is also sensitive to consistency and type (or token) frequency[9], but this sensitivity depends on the level of agreement between the units of analysis. Regularity is also represented in the model's print-to-sound conversion, when smaller units are involved in the output. However, this property appears to play a smaller role given that the larger units are favoured over smaller units. Finally, the model does not exhibit variability in nonword reading.

The multiple-levels model has also been implemented as a connectionist network (Norris, 1994), where the same units of analysis (onset, nucleus, coda, antibody and body) were used as orthographic input nodes, connected to three sets of phonological output nodes (onset, nucleus and coda) via feedforward, facilitatory connections. The sets of output nodes consist of competing, mutually inhibiting pronunciations, if such exist. For instance, since the model has two competing rules, *i* → /I/ (as in *hint*) and *i* → /2/ (as in *pint*), these two pronunciations are represented in the set of nucleus output nodes and they receive activation from any corresponding input node, that is, orthographic units *pi*, *i* and *int*. The input activation is determined by the log frequency of the PSC rules and the connection weights between input and output nodes are determined by the unit size.

1.2.5 Evaluation of Current Computational Models of Reading

The way in which computational models are evaluated has changed over time. Many of the earlier assessments of model performance in nonword naming (e.g., Coltheart et al., 2001; Perry et al., 2007; Plaut et al, 1996) have since been judged as too lenient (e.g., Gubian et al., 2022; Treiman et al., 2003). For instance, Coltheart et al. (2001) scored any nonword pronunciation produced by the DRC model as correct if it followed the GPC rules of the model. Similarly, Plaut et al. (1996) accepted any nonword pronunciation produced by their PDP models (Simulations 2 and 3) that corresponded to any pronunciation of the same word

[9] Either type of frequency can be used, but after some initial comparisons, Norris concluded that relying on type frequency yields better performance of the model.

body segment in the model's training set (e.g. nonword *mave* was correct if the pronunciation was /m{v/ as in *have* or /m1v/ as in *cave*). However, rather than showing that the models produce plausible pronunciations relative to PSCs in existing words, the comparisons should show whether the models produce the same kind of pronunciations that humans do for a specific group of items. Thus, more formal and fine-grained evaluations with unified criteria for success have since been employed.

Andrews and Scarratt (1998) compared an earlier version of the DRC model[10] and the PDP (Plaut et al., 1996, Simulation 1) model against their human nonword reading data. The main empirical findings that bear relevance to their assessment of the models' performance were as follows. Firstly, nonwords with regular or inconsistent bodies are mostly pronounced regularly, whereas considerable proportion of nonwords (from 40% to 65%) with irregular bodies were pronounced irregularly (i.e., as a word body analogy). Secondly, whether a nonword is pronounced regularly or irregularly was better predicted by type-based measures of different properties of nonwords (such as consistency or proportion of regular body neighbours) than token-based measures. The PDP model would predict irregular pronunciations to inconsistent items, especially if these items shared a word body with a highly frequent word (the authors give examples of the PDP model's performance on Glushko's (1970) nonwords: *lome* pronounced as in *come, plove* pronounced as in *love*). This kind of influence of token frequency or high incidence of irregular pronunciations assigned to inconsistent nonwords was not found in Andrews and Scarratt's data. In light of these findings, the authors argue that the influence of token frequency in PDP models needs to be reconsidered and that the continuous effect of inconsistent neighbours on nonword naming predicted by PDP models was not supported by the empirical findings.

The DRC model, on the other hand, produces regular pronunciations to most of the nonword items, thus following the general pattern found in the human data. However, almost 20% of the items that most participants (60-90%) pronounced irregularly were pronounced regularly by the DRC model. Closer inspection of the irregularly pronounced items by the DRC model revealed that these are a result of applying a multi-letter rule (e.g., *igh* → /2/) to the nonwords, rather than influence of the lexical route on the model's pronunciation. Nevertheless, the lexical effects can be seen in the model's naming latencies, which are longer for the items pronounced irregularly by the majority of the participants. Thus, the main issue of the DRC model's performance with the Andrews and Scarratt's data is

[10] This version was a further development from the DRC model by Coltheart et al. (1993)

overestimation of regular pronunciations for some of the items, which may be alleviated by allowing stronger influence from the lexical route. The authors also consider the DRC model's rule extracting algorithm (Coltheart et al., 1993) as an explanation of how PSCs are learnt by humans and conclude that the current implementation of the rule learning in the DRC model lacks psychological plausibility (such as learning all the single letter rules before multi-letter rules).

Treiman and colleagues (2003) compared several models against their human nonword reading data. These models included the DRC (Coltheart et al., 2001), the CDP (Zorzi, et al., 1998), several PDP models (e.g., Plaut et al., 1996; Harm & Seidenberg, 2004) and the Multiple-levels model (the connectionist versions of the model were used here: one optimised for words from Waters and Seidenberg (1985), which is referred to as Multiple-levels-WS, and one optimised for words from Taraban and McClelland (1987), referred to as Multiple-levels-TM). The authors designed an experiment to investigate the role of consonantal context in vowel pronunciations of nonwords (See Unit Size, Section 1.1.4). Eight groups of items were created, where either the onset or the coda is associated with a non-standard (i.e., irregular) pronunciation of the vowel in existing English words. Each item group consisted of critical (e.g., *wasb, clead*) and control items (e.g., *trabs, cleam*). One of the assessment methods compared the proportion of items for which a given model's pronunciation matched the most common human pronunciation. While all assessed models performed reasonably well on the control items (with the minimum match proportion being .86), the performance of the models was considerably weaker on the critical items, where the match proportions ranged from .38 (the DRC model) to .68 (Multiple-levels-TM model).

Treiman and colleagues also inspected the weaknesses of each model included in the comparisons. The DRC model produced no irregular pronunciations, and as such, the authors point out the weakness of restricting the model to grapheme-phoneme sized segments: the model does not show human-like context sensitivity in nonword reading. As Andrews and Scarratt (1998), Treiman et al. identify the issue as a too weak (or non-existent) influence of the lexical route on nonword naming. While this property of the DRC model could be adjusted, it is unclear whether the model would then retain its success in simulating other effects in reading (e.g., Coltheart et al., 2001). Moving on to the CDP model, as the role of orthographic word body segments are central in this model, Treiman and colleagues see this as a hinderance for picking up regularities in the antibody segments (see Chapter 3 for assessment of a more recent model, CDP++, on the Treiman et al., 2003 data set). The PDP

connectionist models (e.g., Plaut et al., 1996; Harm & Seidenberg, 2004) tended to overestimate the influence of consonantal context on vowel pronunciations. The authors suggest training the models with simple PSCs, akin to phonics training many children receive in school. This, together with the traditional training regime, may allow emergence of a more balanced, in other words, human-like nonword reading performance. The Multiple-levels model produced some errors that are not part of the English PSCs. The authors suggest including a mechanism to remove any illegal phonotactic sequences, as is done in some of the other models (e.g., Coltheart et al., 2001; Plaut et al., 1996). The authors conclude that none of the evaluated models capture the patterns in human nonword reading when the items investigated require consideration of the consonantal context of the vowel segments.

Pritchard et al. (2012) compared the performance of the DRC model and CDP models (CDP++ and different versions of CDP+) against their human nonword naming data set. In order to contrast the two models, the nonwords were chosen based on whether the output from the DRC and the CDP+ models diverged from one another. Reading aloud responses from 45 Australian participants for 412 nonwords were collected. A number of comparisons were made. When responses for each nonword were grouped into the most, the second most and the third most common pronunciations amongst the human participants, the highest percentage of human modal matches (a model producing the most popular human pronunciation for an item) was 74% (the DRC model), while the lowest percentage was 12% (CDP+). Additionally, the percentage of items for which the models produced pronunciations that no human participant produced was at best 2% (DRC) and at worst 49% (CDP+). Another type of comparison consisted of the rate of lexicalisations – i.e., giving a pronunciation of an existing word in response to a nonword – in the human responses and the models outputs. The percentage of lexicalisations in the human data was 9%, which was notably different from the DRC model's 0% or the CDP models' 20% (CDP+) or 17% (CDP++). Overall, the authors conclude that the DRC model outperformed the CDP models, yet, none of the models performed particularly well.

In a recent study, Mousikou, et al. (2017) compared the performance of computational models that can read aloud disyllabic letter strings. Both pronunciation and stress assigned to disyllabic nonwords were investigated, but I will focus only on the former, as stress is beyond the scope of the current PhD project. Mousikou and colleagues collected naming responses from 41 participants for 915 disyllabic nonwords. These naming responses were compared to the output from the CDP++ model and a rule-based algorithm based on the dual-route

framework (Rastle & Coltheart, 2000), referred to as RC00 hereafter. The participants' responses were categorised by frequency (i.e., the first most common pronunciation, the second most common pronunciation etc.) and the output of the CDP++ and RC00 was scored as a percentage of matches in each category. Most notably, the human modal responses were produced only for 44% (CDP++) and 55% (RC00) of the items and total number of matches remained at 76% for CDP++ and 88% for RC00. The success of the models remains considerably low even though a match to a model's output was considered any sequence of phonemes produced by at least one human participant, regardless of stress assignment. Mousikou and colleagues list the types of errors produced by each model and conclude that the models do not provide a good account of the processes involved in disyllabic nonword reading.

Finally, new approaches to evaluating models of reading have been developed in recent years, where the model output is not compared to the average or the most common human reading responses, but to individual participants (Mousikou et al., 2017; Robidoux & Pritchard, 2014). For instance, Mousikou et al. (2017) calculated the proportion of matching pronunciations between each participant pair in their sample (similarity values), as well as between each participant and a computational model (CDP++ and RC00 described above). These similarity values allowed investigating the extent to which the output from the computational models was within the range of human responses. Mousikou et al. conclude that both CDP++ and RC00 yielded lower similarity to the participants than the participants did to each other. However, there was more overlap between the similarity values of RC00 and humans than between CDP++ and humans, indicating that the performance of the RC00 corresponds to that of a more typical human participant than the CDP++ (see Chapter 3, Section 3.2.3 for application of this approach). Similarly, Robidoux and Pritchard (2014) used hierarchical cluster analysis to uncover groups of participants with similar reading profiles in the Pritchard et al. (2012) nonword set. The DRC and CDP++ models were also used as 'participants' to determine how similar the pattern of reading responses from the participants was to the output from these models. Robidoux and Pritchard present results using the Ward's method of cluster analysis, using the proportion of items disagreed on by two participants as the measure of distance between them (participants or models). In this method, the most similar pair in the data is merged into a cluster, then the second most similar pair in the data is merged together, and so on. Robidoux and Pritchard showed that the DRC model was merged with a participant very early in the process and that the DRC

belonged to a final cluster with the largest number of participants. By contrast, the CDP++ merged with a small cluster of participants, late in the analysis. As such, the authors conclude that the DRC model produces more similar reading responses to the human participants than the CDP++ model does.

In summary, the available evaluations of the current computational models of reading demonstrate that pronunciation of nonwords by skilled readers is not fully captured by these models. As such, the theoretical accounts and/or computational implementations of these accounts regarding generalisation of PSC knowledge need to be refined.

1.3 Aims of the Thesis

How do skilled readers generalise their linguistic knowledge when reading aloud new words? This question has been explored by inspecting how skilled readers assign pronunciations to nonwords. The empirical investigations and the computational modelling of nonword reading reviewed above both indicate that more work is needed in this area. In an attempt to contribute to our understanding of generalisation in nonword reading, I approached this question computationally, empirically and methodologically.

Firstly, in an attempt to gain further insight into the mechanisms by which print-to-sound conversion happens in skilled readers, I developed a new symbolic model of reading, referred to as the Weighted Segments Pronunciation (WSP) model. Drawing from previous empirical and modelling work, the WSP model combines some of the strengths of the previous models and simulates variability in nonword reading, an aspect of skilled reading behaviour which is not addressed by nearly any of the previous models (cf. Zevin & Seidenberg, 2006).

Secondly, two empirical studies were carried out in order to investigate which statistical properties of the writing system skilled readers are sensitive to when reading aloud nonwords. I focused on the role of type and token frequency in nonword reading, as this question is still open, with only a handful of studies directly addressing this question.

Thirdly, as the PSC knowledge skilled readers have is central to understanding how this knowledge is generalised in nonword reading, I evaluated a relatively understudied method of assessing skilled readers' PSC knowledge. This method, referred to as the nonword rating method, focuses on acceptability ratings for nonword pronunciations rather than verbal naming responses to nonwords.

The following research aims are addressed by the current PhD project, with chapters covering each aim in parenthesis:

1. Can the WSP model simulate central tendencies of nonword reading in skilled readers? (Chapter 3, Chapter 4, Chapter 5)
2. Can the WSP model simulate variability in skilled nonword reading? (Chapter 3)
3. Does token frequency of PSCs influence nonword processing? (Chapter 4)
4. Does type frequency of PSCs influence nonword processing? (Chapter 5)
5. Can PSC knowledge of skilled readers be assessed using a nonword rating method instead of a nonword naming method? (Chapter 6)

1.4 Conclusion

This chapter summarised important empirical findings regarding nonword reading by skilled readers. Out of the statistical properties focused on, it appears that both regularity and consistency of PSCs are influential in pronunciations assigned to nonwords, as neither alone can explain the pattern of findings in the literature. The role of type frequency seems to be more important in nonword reading than the role of token frequency. However, more direct investigation of the contrast between type and token frequency is needed (see Chapters 4 and 5). The empirical evidence clearly shows that the nonword reading responses as well as the unit size used in nonword reading are variable. Several computational models of reading were presented, particularly regarding the properties of the writing system listed above. Evaluations of the computational models were outlined, with a clear conclusion that more work is needed for the models to fully capture the pattern of reading behaviour found in skilled readers. Finally, aims of the current PhD project were listed, which cover computational, empirical and methodological investigations in nonword reading.

Chapter 2 : The Weighted Segments Pronunciation model

Based on the empirical investigations of nonword reading (Chapter 1), a successful model of reading aloud should accommodate variability and flexible unit size in nonword reading. Out of the statistical properties described, it appears that consistency and type frequency of print-to-sound correspondences (PSCs) are particularly important in nonword reading, while token frequency may be influential to a lesser extent. Regularity of PSCs also seems to play a role in print-to-sound conversion, as skilled readers tend to favour grapheme-phoneme sized PSCs in nonword reading, a preference likely stemming from emphasis on grapheme-phoneme-correspondences (GPCs) in early reading instruction (Thompson et al., 2009). A combination of some or all of these properties may be needed for an accurate simulation of human print-to-sound conversion. Focusing on the flexible unit size in nonword reading, the empirical work suggests that grapheme and word body sized segments are clearly utilised by skilled readers, whereas antibody sized segments may be limited to only a few cases, namely, the vowel *a* preceded by *w* or *qu* (Treiman et al., 2003).

In response to these findings, I developed the Weighted Segments Pronunciation (WSP) model, a symbolic model of reading, which reflects the following view of generalisation of PSC knowledge in reading aloud. Skilled readers have knowledge of several, competing PSCs. Reading aloud an unknown word is a result of choices regarding how to parse the letter string, which heavily influences the pronunciation assigned to each segment in the letter string. Exposure to reading equips skilled readers with statistical information about PSCs[11], making certain pronunciations for a given letter string more likely than others. Exposure to reading, as well as other influences, such as reading instruction received in school, result in global tendencies to parse letter strings in certain ways.

As seen in the description of current computational models (Chapter 1, Section 1.2), variability in nonword reading is widely neglected: the same nonword will always be pronounced the same way by these models. As such, the current models attempt to simulate the most popular human responses for nonwords (i.e., the human modal response), but not the variability in pronunciations assigned to the same nonwords. While focusing on the human

[11] 'PSC knowledge' and 'statistical information about PSCs' refer to mostly implicit knowledge, although some of this knowledge is also explicit (e.g., that the grapheme *ch* as onset is mostly pronounced as /J/ (as in *chat*), but sometimes as /S/ (as in *chef*).

modal response provides important information about the central tendencies in reading aloud, it is also a reductionist approach, as categorisation of naming responses to modal responses sometimes disregards a considerable proportion of other pronunciations, produced by a considerable number of skilled readers (e.g., Pritchard et al., 2012). Furthermore, extracting the most common response is sometimes arbitrary, when the numbers of participants giving the most common and the second most common naming responses are almost equal. To address this issue, the WSP model operates in two different modes – the deterministic and the variable mode. The deterministic mode produces a single, constant output to a letter string, comparable to output from other contemporary models. By contrast, the output from the variable mode may change each time the same letter string is read aloud by the WSP model. The variable mode can thus simulate reading performance of a single participant, and multiple simulations of the same data set allows for the generation of output comparable to responses from a group of participants.

2.1 Print-to-sound conversion in the WSP model

2.1.1 Deterministic mode

The core of the WSP model relies on parsing a letter string into different sized orthographic segments, assigning a pronunciation to these segments (thus forming potentially different pronunciation options for the full letter string between different parsing styles), and resolving the competition between the parsing styles based on their overall strength. The strength of a parsing style is based on two factors: 1), the different statistical properties of the PSCs in each parsing style and their relationship to each other (if more than one property is used) and 2), the weight applied to each parsing style. The output of the model for any letter string is the pronunciation corresponding to the strongest parsing style.

More specifically, any nonword given as input will be parsed into consonant and vowel clusters, forming three parsing styles: antibody-coda (e.g., *stra-nd*), onset-word body (e.g., *str-and*) and onset-nucleus-coda (e.g., *str-a-nd)*. Thus, the final pronunciation of the model can be based on PSCs at the level of larger segments (antibody and word body parsing styles) or at a smaller, grapheme-sized level[12]. The pronunciation most consistently associated with a given orthographic segment is assigned to each orthographic segment within each parsing style. PSCs with a minimum type frequency of 1 are available to the model, which means that the model's pronunciation can be based on a PSC occurring in only a single word. This

[12] Note, however, that the consonant clusters in the onset-nucleus-coda parsing style are sometimes a combination of several graphemes, e.g., a coda *lm* would be a single cluster, rather than two separate graphemes.

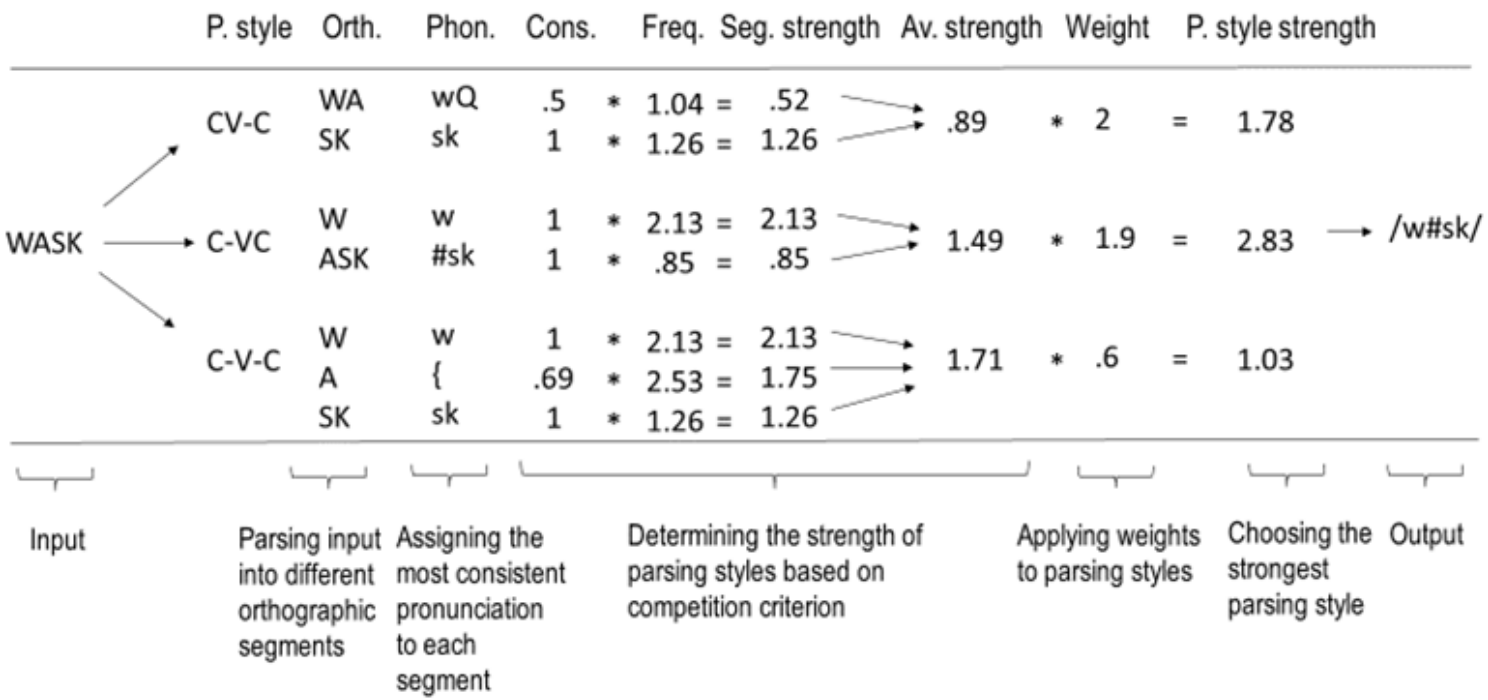
parameter can be changed by the user to include only PSCs occurring in, for instance, at least two different words.

The competition between the parsing styles will be described next, through an example. The first factor defining the strength of a parsing style are the statistical properties of the PSCs of each segment and how they relate to each other, referred to as the competition criterion hereafter. In this example, the competition criterion is the consistency and the type frequency of each segment, and their relationship is multiplicative. In order to bring the scales of consistency (ranging from 0 to 1) and type frequency (ranging from 1 to several hundreds) closer together, the measure of type frequency used will be logarithmic. Additionally, a constant 1 is added to each type frequency value to ensure that each value is above 0. Thus, the competition criterion is consistency * log10(type frequency +1). Applying this competition criterion to each segment results in a strength value for each segment (i.e., the higher the consistency and type frequency of the PSC, the higher the strength of the segment). The resulting strength values of each segment within a parsing style are then averaged together, and finally the weights are applied to the overall strength of each parsing style. The weight for each parsing style is constant across items, such that the same weight is always applied to the same parsing style.

The weights are added to reflect a tendency to favour a particular parsing style over others, over and beyond the differences the competition criterion might create. As demonstrated by empirical investigations of reading, the tendency for parsing a new letter string in a particular way can be influenced by, for instance, reading instruction received in school (Deavers, et al., 2000). As such, the weights for different parsing styles in the WSP model could be adjusted to reflect different populations and their parsing preferences. See Figure 2.1 for an example of the competition between the parsing styles, for a letter string *wask.* Note that the three parsing styles do not always diverge. For instance, for a letter string *sall*, both the antibody and small segment options would be /s{l/ while the word body option would be /s$l/. As another example, all three parsing styles would produce the same pronunciation, /mEst/ for a letter string *mest*.

Figure 2.1

*Example of parsing style competition in WSP model with competition criterion Consistency * log10(Type Frequency + 1)*



Note. P. style = parsing style; CV-C = antibody-coda; C-VC = onset-word body; C-V-C = onset-nucleus-coda; Orth. = orthographic segments, Phon. = phonological segments, Cons. = consistency value, Freq. = type frequency value (note: the value in the figure is log10(type frequency + 1), Seg. strengths = segment strengths, Av. strength = average of all the segment strengths within a parsing style.

2.1.2 Variable mode

As mentioned in Chapter 1 (Section 1.1.5), variability in skilled nonword reading is a robust finding and reading responses to nonwords vary between and within participants. The type of variability in nonword reading that the WSP model's variable mode aims to simulate is within-participants variability. Here the focus is on the randomness of pronunciation assignment when a letter string has several plausible pronunciation options, while the underlying probabilities for the pronunciation options are based on statistical properties of the writing system. However, another method of extracting variable output is also described.

The assembly and strength of pronunciation options corresponding to the three parsing styles described for the deterministic mode also apply to the variable mode. However, while only the most consistent pronunciation for any orthographic segment was available in the deterministic mode, all the pronunciations associated with a given orthographic segment that have a consistency of at least .3[13] are available in the variable mode. For example, the most consistent PSC *ear* → /7/ (as in *clear*) and the second most consistent PSC *ear* → /3/ (as in *learn*) are both available to be combined with the onset and coda of a given nonword. If several pronunciations are available for several different segments, all combinations of these segments are assembled and included as pronunciation options from the given parsing style. For instance, for the letter string *gear* (let's assume this is an unknown word for the model), the small segment parsing style would have two possible pronunciations for the onset, /_/ and /g/, and two possible pronunciations for the vowel segment , /7/ and /3/, thus resulting in four pronunciation options from this parsing style: /_7/, /g7/, /_3/ and /g3/. The strength of each pronunciation option for the full letter string (as a product of the competition criterion and the weight of each parsing style) is then converted into a probability, so that the strength of each unique pronunciation option is relative to the summed strengths of all the options, from all three parsing styles.

The output produced by the variable mode of the model contains two elements: the different pronunciation options for a letter string, and the probabilities for each option. This output can

[13] The minimum consistency for including PSCs can be adjusted, but as it should reflect the correspondences likely to be known by and easily available to a skilled reader: very low consistency thresholds for including a PSC can result in unlikely pronunciations (e.g., pronouncing *ie* as /E/ (consistency of 0.06), as this is the second most consistent PSC after ie - /i/, based on only one item, *friend,* in WSP's vocabulary. Note, however, that the *iend* - /End/ PSC as a word body unit would still be included in reading items with a word body *iend*, as the consistency for the body unit would be higher than .3).

be achieved in two different ways. Firstly, the probabilities for each pronunciation option can be used directly as proportions for each pronunciation option (the raw probabilities method), in which case these probabilities can be thought of as the proportion of participants assigning each pronunciation to the letter string. Alternatively, the final pronunciation for any letter string can be chosen at random, weighted by the probabilities for each option. As such, while a single reading response of the WSP model can be any of the pronunciation options for a letter string, the most probable option will be chosen more often than the less probable options, if the same letter string is read aloud several times by the model (the multiple simulation runs method). Multiple simulation runs of the same set of nonwords thus result in output with similar structure to naming responses from a group of participants – extracting the proportions for each pronunciation option for each item summarizes this type of data into the form in which the output from the raw probabilities method already is. While both the raw probabilities and the multiple simulations methods produce output that can be compared to a group of participants (both in terms of the pronunciation options and the proportion of participants producing each option), only the multiple simulation runs method produces individual reading responses, which allows comparisons to individual participants, for instance, by using the methods described in Section 1.2.5 in Chapter 1 (Mousikou et al., 2017; Robidoux & Pritchard, 2014).

The weights used in the deterministic mode are not necessarily the same as the ones used in the variable mode. This is because the process of producing a representative human modal response is different from the process of producing representative proportions for different pronunciation options. For instance, let us assume that a letter string *rall* is pronounced as /r$l/ by 44% of participants and as /r{l/ by 42% of participants. In the WSP model, only the pronunciation corresponding to the word body parsing style matches the human modal response, /r$l/. For the deterministic mode of the WSP to produce the correct human modal response, the word body parsing style needs to be advantaged enough to win the competition between the three parsing styles (let us only focus on the weights in this example and ignore the difference in parsing style strengths). This could be achieved by a significantly large weight for the word body parsing style, as the relative strength of the other parsing styles does not matter, beyond the requirement that they are smaller than the word body parsing style. In the variable mode, by contrast, the word body parsing style should also be the strongest (so that the proportion of /r$l/ pronunciations would be the highest), but the relative strength of the antibody and small segment parsing styles, both of which produce the

pronunciation /r{l/, should be almost equally strong. Thus, the optimal balance between the weights in each mode of the WSP is different, depending on whether the aim is to simulate the human modal responses only (deterministic mode) or different pronunciation options and the relative frequency in which the different options are assigned to a letter string (variable mode).

2.2 Print-to-sound knowledge of WSP model

If skilled readers base their pronunciation choices for nonwords on statistical properties of the writing system, these properties must be extracted from readers' experience with words that they already know how to read. In other words, the vocabulary of each reader serves as the repository from which the statistical properties of different PSCs are derived. This idea is employed in the WSP model, as the PSC knowledge of the model is based on a vocabulary.

2.2.1 Vocabulary

In current computational models, the language experience that print-to-sound conversion is based on has traditionally consisted of monosyllabic, often monomorphemic words (Coltheart et al., 1993; Coltheart et al., 2001; Seidenberg & McClelland 1989). The same, albeit more lenient approach is taken for the WSP model. The vocabulary of the WSP model consists of 3921 monosyllabic and mostly monomorphemic words. The orthographic and phonological forms of the words were retrieved from the web interface of the Celex database (Baayen et al., 1995, retrieved at http://celex.mpi.nl/) and a logarithmic token frequency measure (Zipf) for each item was extracted from the SUBTLEX-UK database (Van Heuven, Mandera, Keuleers & Brysbaert, 2014; retrieved at https://psychology.nottingham.ac.uk/subtlex-uk/). Only items found in both databases were included. Items included as monosyllabic were any items that had one orthographic or one phonetic syllable only, based on the WebCelex database. Thus, items with two orthographic but one phonetic syllable (e.g., *fuel*) were included as well as items with one orthographic but two phonetic syllables (e.g., *hour*). This more lenient criterion for monosyllabic words was adopted to create a more complete set of PSCs that skilled readers are likely to be familiar with. Items without a vowel (e.g., *shh* or names of letters) and contractions (e.g., *ma'am*, *ne'er*) were removed.

The items included were monomorphemic, with the exception of irregular past tenses, which were included to increase the repertoire of PSCs (e.g., *was* pronounced as /wQz/ or *lead* pronounced as /lEd/). There is evidence suggesting that lexical items in irregular past tense are stored as separate entries in the mental lexicon, while past tense for regularly inflected

items are constructed from their base forms, following morphological rules (Ullman, 1999; Newman et al., 2007). As such, because WSP's vocabulary includes items in irregular past tense but not items in regular past tense, it should resemble the mental lexicon that skilled readers draw from when they utilise the statistical information of PSCs in reading.

Token frequency information was sensitive to part of speech (PoS) for items with different pronunciation based on PoS (e.g., *dove* – /d5v/ as a verb but *dove* – /dVv/ as a noun). PoS specific Zipfs were calculated for these items where possible. Items with a different pronunciation but the same PoS (e.g., read in present and past tense) received the PoS specific Zipf (a total of Zipf for all the items with a given orthographic form and PoS). The general Zipf (a total of Zipf for all the items with a given orthographic form regardless of PoS) was assigned to all other items.

2.2.2 Statistical properties of PSCs

In line with the empirical investigations of reading aloud, the PSC knowledge of the WSP model contains information about the consistency, type frequency and token frequency of PSCs of varying sizes. The different sized PSCs mirror the three parsing styles – i.e., consonant and vowel clusters as well as antibody and word body sized segments. The consistency of each PSC was calculated as the proportion of words with a given orthographic segment that are pronounced the same way (friends), out of all the words with the given orthographic segment (friends and enemies). For instance, the PSC *ould* – /Ud/ has a consistency of 0.75, (*could, should, would* vs *mould*). The PSC *ould* – /5ld/ has a consistency of 0.25, as only the word *mould* contains this PSC. This was a consistency measure based on types. Token-based consistency measures were also calculated, where summed token frequency of friends was divided with the summed token frequency of friends and enemies.

The type frequency of each PSC was calculated as the number of words in which a given PSC occurs. Using the example above, the PSC *ould* – /Ud/ has a type frequency of 3 and the PSC *ould* – /5ld/ has a type frequency of 1. Token frequency of each PSC was calculated as a summed token frequency of all the words in which a given PSC occurs.

2.2.3 Exceptions

Although parsing a letter string generally results in consonant only and vowel only clusters (or empty clusters for items with no onset or coda, such as *art* or *why*), some exceptions were introduced due to their special role in PSCs. The same exceptions are part of how a letter

string is segmented as input and what kind of PSCs exist in the WSP model's PSC knowledge. Letters *a, e, i, o* and *u* are always considered vowels, whereas *y* is classified as a consonant when it is the first letter of the input string (e.g., *yeast* is parsed as *y-ea-st*) and otherwise as a vowel. The rest of the letters are considered consonants, with the exception of *r* and *w* following a vowel. In these cases, *r* and *w* are attached to the vowel cluster, e.g., *st-ar-t* and *cr-aw-l.* These exceptions are included because the *r* and *w* change the pronunciation of the preceding vowel. For instance, graphemes *ar* and *aw* are most consistently associated with pronunciations /#/ and /*$*/, as in *far* and *straw*, as opposed to the most consistent pronunciation of *a* as nucleus (/*{*/ as in *cat*).

Additionally, *u* is considered a part of the consonants *g* and *q* when these consonants precede a *u* in the beginning or the end of a word. This results in the most common pronunciations /g/ and /kw/ for graphemes *gu* and *qu* in items like *guest* and *quest.* Equally, graphemes *gue* and *que* are pronounced as /g/ and /k/ in items like *rogue* and *casque*. Furthermore, vowel *e* as the final letter of a letter string, preceded by a vowel and a consonant cluster (i.e., the silent *e*), is categorised as both part of the preceding consonant cluster and the preceding vowel. This categorisation was made to address split vowel graphemes, i.e., a long vowel preceding a silent *e*, such as in *huge* (compared to *hug*). As a result, *e* as part of the coda is silent because the most common pronunciation of consonants followed by a single *e* is the consonant in question (e.g., codas *ce, ste* are pronounced as /s/ and /st/), while the split vowel version of any vowel receives the most consistent pronunciation associated with it (e.g., *a_e* –> /1/ as in *ace*, *o_e* –> /5/ as in *rose*).

Onsets *c* and *g* are most often pronounced as /k/ as in *cat* and /g/ as in *golf*. However, these letters are more often assigned a soft pronunciation (/s/ as in *cell* and /_/ as in *gene*) when followed by *e, i* or *y*. In order to produce this context sensitive pronunciation of *c* and *g*, a separate onset for *c* and *g* when these letters are followed by a letter *e*, *i* or *y* was added to the PSC knowledge of the WSP. Statistically, however, *g* followed by *i* is most often pronounced as /g/ (not as /_/) in monosyllabic words. The PSC knowledge of the WSP model reflects this, that is, *gi* pronounced as /g/ instead of /_/. The soft *c* pronunciation is also assigned to onset clusters where the onset-nucleus boundary is *ce*, *ci* or *cy* such as in *scene, science* and *scythe*. However, there were no monosyllabic words with *sc* onset preceding an *i* in WebCelex. To include this PSC, the word *scion* was included in the WSP's vocabulary.

These exceptions are not solely a feature of the WSP model, but the same or similar exceptions can be found in the DRC model as well, where they are classified as context sensitive or multiletter rules (see Rastle & Coltheart, 1999, Appendix B). Similarly, some of these exceptions are included in the segmentation of letter strings in the Multiple-levels model (Norris, 1994). Although it is important to understand why skilled readers seem to have at least implicit knowledge of these exceptions and how they are learnt, the overall goal of the current modelling work is to test the general principle of the WSP model, and therefore these exceptions are included for the time being.

However, many if not all of these exceptions can be justified empirically – the way in which skilled readers seem to parse letter strings corresponds to the exceptions introduced to the model. For instance, I extracted all of the items with vowel + r from a set of 412 nonwords, named by 45 participants (Pritchard et al., 2012, described in detail in Chapter 3, Section 3.2.3). There were 30 items with this spelling pattern. On average, 74% of the participants (ranging from 20% to 100%) produced a context sensitive pronunciation for these items, where the following *r* modifies the vowel. By contrast, a context insensitive pronunciation, where the vowel is pronounced ignoring the *r*, was produced on average by 0.01% of the participants, at most by four participants for any given item. Thus, this nonword reading behaviour suggests that skilled readers have a strong tendency to parse the vowel and the following r together rather than separately.

Finally, English accents differ in rhoticity, i.e., whether *r* is pronounced when it is preceded but not followed by a vowel (e.g., *star* is pronounced /st#r/ in rhotic accents but /st#/ in non-rhotic accents). Most accents in England – apart from the South-West and a part of Lancashire (Trudgill, 1984) – and standard Australian English (Turner, 1994; Trudgill & Gordon, 2006) are non-rhotic. The WebCelex database, however, is not fully consistent with non-rhotic pronunciations. For instance, items with a word body segment *air* (based on words *air, chair, fair, flair, hair, lair, pair* and *stair*) are all pronounced as /8R/ according to WebCelex (*R* denotes a possible linking *r*) [14]. Due to this inconsistency, the PSC knowledge of the WSP model was modified so that the rhotic *r* was removed from the phonemic transcription, thus ensuring the pronunciations of WSP are always non-rhotic.

[14] Even in rhotic accents, this final r in monosyllabic, monomorphemic items could still be pronounced in multisyllabic words if it serves as a linking r, e.g., *care* pronounced as /k8/ but *caring* pronounced as /k8RIN/.

2.2.4 Assembly of unknown segments

If the input string contains orthographic segments that are not part of the PSC knowledge of the WSP model, a pronunciation for them may be assembled from existing PSCs in limited cases. Currently this procedure is available for any unknown coda or onset that can be divided into two known elements, for instance *sht* as an onset would be assembled from the most consistent pronunciations associated with *sh* and *t,* where any properties included in the competition criterion would be calculated as the mean value of the known elements (*sh* → /S/ and *t* → /t/).

The assembly of unknown segments is included in the model to increase the model's ability to read aloud letter strings. As the models' PSC knowledge is only based on monosyllabic words, this procedure essentially increases the model's PSC repertoire. However, this feature is also considered an exceptional strategy to be used when normal reading procedures fail to produce a full pronunciation for a letter string. As such, if a response time feature was added to the WSP model, assembling unknown segments would result in slower responses than reading via the normal procedure.

2.3 Optimisation of WSP model

The optimisation of the WSP model includes two decisions – firstly, which statistical properties of PSCs will determine the strength of each segment – i.e., what is the competition criterion – and secondly, what weight will be applied to each parsing style. Optimisation of the model involves determining the best set of weights for the parsing styles, based on the model's performance against a set of items. As both the PSC knowledge and the tendency to parse letter strings in certain ways are believed to be shaped by reading experience in skilled readers, the WSP model's vocabulary was used for optimisation. In other words, 3921 monosyllabic words were used as the optimisation set. This way, the weights for each parsing style, which reflect the model's tendency to parse letter strings in certain ways, are based on the spelling patterns in existing words that the model 'knows'. Note, however, that the model could be optimised using a human nonword naming data set and the weights would thus be adjusted based on which combination of the weights would yield the most 'human-like' performance, such as the largest number of human modal responses for the item set. This approach is explored in Chapter 3, Section 3.4.

2.3.1 Choice of competition criterion

The choice of competition criterion was based on the empirical findings in nonword reading: consistency of PSCs is clearly influential, while the role of type and token frequency needs further clarification (see also Chapters 4 and 5). Thus, two versions of the WSP model will be investigated further via model optimisation – both with consistency and frequency as the competition criterion, but these measures are either type or token-based. More formally, the competition criterion is presented in (1) for the type-based version of the model and in (2) for the token-based version of the model:

$$comp.cri. = \frac{\text{sum}(friends)}{sum(friends + enemies)} * log10(sum(friends) + 1) \quad (1)$$

$$comp.cri. = \frac{\text{sum(token freq friends)}}{\text{sum}(token\ freq\ friends + enemies)} * \text{log10}(\text{sum}(token\ freq\ friends) + 1) \quad (2)$$

where comp. cri. is the competition criterion a segment strength is based on. Consistency is calculated as the number of friends (words with the given orthographic segment and pronunciation) relative to the number of friends and enemies (i.e., all the words with the given orthographic segment) for the type-based version of the model (WSP-type hereafter). For the token-based version of the model (WSP-token hereafter), the consistency measure is calculated as summed token frequency of friends relative to the summed token frequency of friends and enemies. As the maximum values for the consistency measures in both type and token-based equations are 1, but the maximum values of frequencies are in several hundreds, it was deemed necessary to logarithmically compress the frequency values. A constant 1 was added to the frequency values to ensure that a logarithm of the frequency value is always defined. This logarithmic compression was applied so that the competition between parsing styles would not be driven by frequency alone, but as a contribution of both consistency and frequency.

While type-based measures of consistency and frequency are relatively straightforward to quantify, the token-based measures could be quantified in different ways, for instance, as a

summed, maximum or mean token frequency. I opted for the summed token frequency because there is empirical evidence suggesting that this specific measure is influential in consistency effects in word naming (Jared et al., 1990).

Finally, another property of PSCs considered in Chapter 1, regularity, is not included in the WSP model directly. However, as frequency is part of the competition criterion, the influence of regularity can be seen in the pronunciations based on the small segment parsing style, particularly in the WSP-type version of the model.

2.3.2 The optimisation procedure

The goal of optimising the model is to find a set of weights for each parsing style that would result in the largest number of matches between the model's output and the pronunciations associated with items in the optimisation set. As existing words are used as the optimisation set, the correct pronunciations of these words serve as the criterion for matches. The optimisation was performed as a grid search, where each of the three weights, one for each parsing style, ranged from 0.1 to 2 in increments of 0.1 (i.e., 8000 different combinations of weights). The combination of weights yielding the best performance was chosen for each version of the model (the WSP-type and the WSP-token). The set of weights with the largest values was chosen if several combinations of weights resulted in identical performance.[15]

The best combination of weights found for the deterministic mode of the WSP model were also used for the variable mode of the model. This was because existing words were used as the optimisation set: in order to find the best weights that produce 'human-like' proportions for different pronunciations assigned to a letter string, the optimisation set needs to contain different pronunciations for the same item. As there is no variability in the pronunciations to existing words (i.e., there is no alternative pronunciation for the word *cat* in the same way as there are alternative, plausible pronunciations for the nonword *cearn*), the 'human-likeness' of the proportions of different pronunciation options cannot be assessed. This need not be the case if a nonword reading data set was used for optimisation (see Chapter 3, Section 3.4).

Optimising the WSP model's type and token versions for the vocabulary of the WSP model resulted in the following pattern of weights: the antibody parsing style had a slightly higher weight than that of the word body parsing style, while the weight for the small segment

[15] In the absence of a better criterion, choosing the largest values was ultimately arbitrary, but this criterion was applied consistently.

parsing style was approximately half of the weights for the former two parsing styles. Table 2.1 summarises the weights and other parameters of the two versions of the WSP model.

Table 2.1

Parameters of the vocabulary-optimised versions of the WSP model

Parameter	**WSP-type**	**WSP-token**
Competition criterion[a]	Consistency-type * log(type freq + 1)	Consistency-token * log(token freq + 1)
Weight: antibody-coda parsing style	2	1.9
Weight: onset-word body parsing style	1.9	1.8
Weight: onset-vowel-coda parsing style	0.6	0.7
min. frequency of PSCs	1	1
min. consistency of PSCs	0.3	0.3

[a] refer to (1) and (2) for more detailed calculation of type and token-based consistency and frequency measures

2.4 Conclusion

In this chapter, I described a new computational model of reading, the Weighted Segments Pronunciation (WSP) model, which I have developed as part of the current PhD project. The WSP model converts letter strings into speech sounds by considering PSCs of varying sizes. Before producing output, several competing pronunciation options, corresponding to different parsing styles, can be available to the model. This competition is resolved based on different statistical properties of the PSCs within the parsing styles. In addition to the statistical properties of the PSCs, the model's tendency to parse letter strings into larger or smaller segments is influenced by weights for each parsing style, which can be optimised for a set of words or nonwords. The WSP model can also operate in deterministic or variable modes, which aim to simulate central tendencies in human nonword reading (deterministic mode) and variability in nonword reading (variable mode). In the following chapters, two versions of the WSP model will be evaluated, where the competition of the parsing styles is based on a multiplicative relationship of either type-based consistency and frequency of PSCs (WSP-type) or token-based consistency and frequency of PSCs (WSP-token). These evaluations include comparisons of the WSP model against other computational models of reading and human nonword reading responses.

Chapter 3 : Evaluation of the Weighted Segments Pronunciation model

In this chapter, the performance of the Weighted Segments Pronunciation (WSP) model will be evaluated, both in the model's deterministic and variable modes, against several data sets of human nonword responses. The performance of the WSP model's deterministic mode is also contrasted with two current computational models, the dual-route cascaded model (DRC) and the connectionist dual process model (CDP++, see Chapter 1, Sections 1.2.1 and 1.2.2). The performance of the WSP's variable mode is compared to the performance of, to my knowledge, the only other model that simulates variability in skilled readers' nonword naming responses (Zevin & Seidenberg, 2006). Finally, performance of the WSP model as well as characteristics of the available nonword reading data sets are investigated by optimising the model for each of them. The discussion of the results from these comparisons focuses on potential avenues for improving the WSP model, as well as providing insights on the characteristics of the nonword reading data sets.

3.1 General considerations in the evaluation of the models

I start by testing the performance of WSP model's deterministic mode against three human nonword reading data sets, which capture different aspects of nonword reading, as described below. The deterministic mode of the WSP model is used in these comparisons, to allow contrasting the model's performance with that of the DRC model (Chapter 1, Section 1.2.1) and the CDP++ model (Chapter 1, Section 1.2.2), two dual-process models of reading that have been studied widely and that are publicly available. DRC.1.2.3 (retrieved from https://maxcoltheart.wordpress.com/drc/) and CDP++.2 (retrieved from https://sites.google.com/site/conradperryshome/) were used in the following comparisons, and in all comparisons reported in this thesis. The WSP-type and WSP-token versions of the WSP model are included in these comparisons.

The DRC, CDP++ and WSP models all produce output in DISC phonetic alphabet. The PSC knowledge of these models is based on the Celex database, i.e., British English. However, the DRC model deviates from the CDP++ and WSP model's transcription in two ways: yod-pronunciations (e.g., /ju/ in *dune* → */djun/*) correspond to a symbol /W/ and there is no difference between the phonemes /9/ and /$/ (as in *tour* and *door*, respectively), but rather

any item with either of these phonemes is pronounced as /9/. To unify the transcription between different models, I have changed the /W/ into /ju/ and /9/ into /$/ in the DRC's output. The same changes have also been made to the transcription of the participant responses in the nonword data set by Pritchard et al. (2012), described below.

3.2 Evaluation of the WSP model's deterministic mode

In the following data sets, the nonwords and the pronunciations assigned to them are referred to as regular and irregular, relative to GPC rules. Regular nonwords contain an orthographic segment (typically the word body), which is always pronounced regularly in existing words. Similarly, irregular nonwords are items that share orthographic segments with existing, irregularly pronounced words. This categorisation is used throughout, as the focus in the comparisons between different models is mostly in the regular vs irregular – or standard vs context sensitive – dimension. This is why the vowel pronunciations of nonwords are the primary interest in these comparisons; the vowel pronunciations allow differentiating between the regular and irregular pronunciations.

3.2.1 Andrews and Scarratt set

This data set consists of 16 nonwords, derived from eight different word bodies that are always pronounced irregularly in existing words (Andrews & Scarratt, 1998, Exp. 1)[16]. The naming responses from 24 Australian participants were categorised as irregular, regular or other, based on the vowel pronunciation of the items. For instance, the vowel in *pould* pronounced as /6/ (as in *loud*) was regular, as /U/ (as in *could*) was irregular and any other vowel pronunciation was categorised as other. I categorised the output to these items from the DRC, CDP++ and WSP models in the same way, extracted the type of the most common pronunciation assigned to each nonword by the participants (i.e., the human modal response) and compared these pronunciation types to the pronunciation types of the models' output. Each word body of these nonwords was irregular in existing words and there was considerable preference for irregular pronunciations (as word body analogies) in the human data for these items: the human modal response was irregular for 62.50% of the items and regular for 18.75% of the items. As such, this data set should be best simulated by models that utilise larger unit size in print-to-sound conversion.

[16] The stimuli in the experiment 1 by Andrews & Scarratt consisted of 216 nonwords, but detailed naming responses were only reported for 16 of these.

Table 3.1 lists the human modal responses and the output from each computational model for these nonwords. The highest proportion of matches with human modal responses was produced by the CDP++ (.75), followed by the WSP-token (.56), the DRC (.44) and the WSP-type (.38).

Table 3.1

Human modal response type and output from computational models to 16 nonwords with irregular word bodies from Andrews and Scarratt (1998, Exp. 1)

item	Human modal	DRC	CDP++	WSP-type	WSP-token
beart	irregular (/#/)	b7t	b3t	b8t	b#t
kneart	other	n7t	n3t	n#t	n#t
chalt	regular (/{/)	J{lt	J$lt	J$lt	J$lt
wralt	irregular (/$/)	r{lt	r$lt	r{lt	r{lt
chigh	irregular (/2/)	J2	J2	J2	J2
jigh	irregular (/2/)	_2	_2	_2	_2
gight	irregular (/2/)	g2t	g2t	glt	g2t
zight	irregular (/2/)	z2t	z2t	z2t	z2t
ginth	regular (/I/)	gInT	gInT	gInT	gInT
jinth	regular (/I/)	_InT	_InT	_2nT	_2nT
vould	other	v6ld	v6ld	vUd	vUd
pould	other	p6ld	p6ld	pUd	pUd
roup	irregular (/u/)	rup[17]	rup	rup	rup
moup	irregular (/u/)	m6p	mup	m6p	mup
searn	irregular (/3/)	s7n	s3n	s3n	s3n
gearn	irregular (/3/)	_7n	g3n	g7n	g7n

Note. Human modal column shows the type of vowel pronunciation assigned to each item, with the DISC transcription of the vowel pronunciation in parenthesis.

Only three of the items (*chalt*, *ginth* and *jinth*) were assigned regular pronunciations in the modal responses of human readers. These responses were predicted perfectly by the DRC model, the CDP++ matched two out of three of the regularly pronounced items, while WSP-type and WSP-token matched only one of the regularly pronounced items. Irregular vowel pronunciations were assigned to 10 items in the human modal responses. An average of 70%

[17] The item *roup* is not included in the analysis as a match, because its irregular pronunciation is a result of this item being part of the DRC model's vocabulary, read aloud based on the output from the lexical route. Because the CDP++ pronounced the nonword *moup* irregularly, I assume the irregular pronunciation of the item *roup* would also remain even if this item was not part of the CDP++ model's vocabulary. Thus, this item was retained for the CDP++. Note that *roup* is not part of the WSP's vocabulary, because even though this item is found in WebCelex, it is not found in SUBTLEX-UK (see Chapter 2, Section 2.2.1).

of participants (from 32 – 100%) pronounced these items irregularly. The CDP++ matched nine of these responses, failing only with the item *beart,* which received the lowest percentage of irregular pronunciations amongst the participants (32%), out of the items with irregular human modal response. The WSP-token matched eight of the irregular pronunciations, but failed to produce an irregular pronunciation to items *wralt* and *gearn*, which were pronounced irregularly by 55% and 86% of the participants, respectively. By contrast, the WSP-type matched only half of the irregularly pronounced items and the DRC matched 40% of the irregularly pronounced items. The average percentage of participants assigning an irregular pronunciation to the items that DRC and WSP-type failed to pronounce irregularly were 67% and 61%, respectively. As such, the items for which the two models fail to produce irregular pronunciations are not items that receive particularly low percentages of irregular pronunciations amongst humans.

All of the irregular pronunciations produced by the DRC model are multiletter-graphemes in the model's GPC rules, namely, *igh* and *ight*. The WSP-type performs poorly on this data set, which seems to be a result of relatively strong influence of the antibody parsing style in the model's output – six out of 16 responses were based on the antibody parsing style, which mostly yields regular word body responses. By contrast, only three responses in the WSP-token output were based on the antibody parsing style. The rest of the responses in the output of both WSP-type and WSP-token were based on the word body sized segments.

In summary, most of the models that include larger unit size in their PSCs perform better on this data set than models that do not. The CDP++ model had the strongest performance, followed by the WSP-token, the DRC and the WSP-type. The CDP++ and WSP-token models predict regular and irregular pronunciations relatively well, and most failures to do so happen with items that are not the clearest exemplars of their pronunciation category amongst humans. However, the DRC model predicts regular pronunciations well, but fails on over half of the items where irregular pronunciations dominate in the human data. Finally, the WSP-type performs poorly on this data set, which is likely explained by a relatively strong emphasis on the antibody parsing style in the competition between the three parsing styles, and as a result, the model's responses.

3.2.2 Treiman set

This data set consists of 158 nonwords with antibody or word body segments that are pronounced either irregularly or regularly (relative to GPC rules) in existing words (Treiman

et al., 2003, Exp. 1). An additional 20 filler nonwords present in the original study were not included in current comparisons. The naming responses from 24 American participants were categorised as irregular or regular based on the vowel pronunciation for each item. The data set was organised into eight item groups with equal numbers of regular and irregular items. For instance, one of the two antibody-item groups (or CV-item groups) consisted of irregular items such as *wark* and regular items such as *tark*. In the irregular items in this group, the onset *w* modifies the pronunciation of the vowel *a* in existing words to /Q/ or /$/ (as in *watch* and *war*) whereas the vowel *a* preceded by most other onsets would be pronounced regularly as /{/ or /#/ (as in *cat* and *car*). Similarly, one of the six word body-item groups (or VC-item groups) consisted of irregular items (e.g., *chead*) and regular items (e.g., *cheal*), where the irregular vowel pronunciation /E/ (as in *head*) is associated with the following coda *d* in existing words, while the regular vowel pronunciation /i/ (as in *seal*) is linked to most other codas.

In order to perform a fair comparison between the American participants' responses and the output from the computational models, all of which were based on Australian or British English, the categorisation of the irregular and regular vowel pronunciations from the models were based on the models' dialect, rather than an exact phonemic match between the participants and the models. Two types of comparisons to output from computational models were made. Firstly, the human modal responses were matched to the output from each model as a total proportion of matches, and as a proportion of matches for the regular and irregular items separately. All matches are based on the vowel pronunciation in these comparisons. Secondly, for each item group, the proportion of irregular vowel pronunciations was calculated for both irregular and regular items and the difference between the proportions was computed. These proportion differences can be seen as a measure of context sensitivity – a large proportion difference within an item group indicates that the irregular items were mostly pronounced irregularly (i.e., considering the context for each vowel) and the regular items were mostly pronounced regularly in the given item group. Thus, I refer to the proportion differences for each item group as context sensitivity score hereafter. The context sensitivity scores could also be calculated for the output from the DRC, CDP++ and WSP models. Out of the irregular items, 60.75% received an irregular (context sensitive) pronunciation as the human modal response, whereas all regular items were pronounced regularly by the majority of the participants. The participant responses reveal considerable preference for irregular pronunciations for the irregular items compared to the regular items,

and as such this data set, too, should be best simulated by models that accommodate larger PSCs in their print-to-sound conversion.

The comparison between the computational models and the human modal responses revealed that the two versions of the WSP model had the highest proportion of total matches, followed by the DRC and finally the CDP++ (see Table 3.2 for details). The WSP model's versions also produced the best match for irregular items, followed by the CDP++ and then the DRC. However, the DRC matched the human modal responses for regular items perfectly, while the other three models reached at most 87% of matches to human modal responses for the regular items.

Table 3.2

Proportion of matches between Treiman et al. (2003) human modal responses and output from DRC, CDP++ and WSP models

	Match to human modal responses		
Model	*Regular*	*Irregular*	*Total*
DRC	1.00	.38	.69
CDP++	.87	.44	.66
WSP-type	.85	.66	.75
WSP-token	.86	.71	.78

For the DRC model, all mismatches for irregular items were a result of regular pronunciations produced by the model, except for one item that was pronounced irregularly, when the majority of the participants pronounced it regularly (*wadge* as /wQ_/ by the DRC and as /w{_/ by the majority of the participants). As noted by Treiman et al. (2003), this single irregular response from the DRC model is because the item *wadge* is part of the model's vocabulary, and thus it is pronounced using the lexical information about the item's pronunciation rather than assembled via the GPC rule system. This item is thus not included in the calculations of context sensitivity scores, reported below. For the CDP++, most mismatches were irregular pronunciations (23 items), or regular pronunciations (15 items) produced by the model when the majority of the participants gave the opposite type of response. However, the model also produced 16 other responses, which were mostly not produced by any of the participants (such as *wark* as /wUk/ or *warse* as /w8z/ and several items with a word body *ance* pronounced as /#ns/). For the WSP-type, most of the mismatches were due to regular responses (13 items) or irregular responses (13 items) when

the majority of the participants did not produce that pronunciation. There were also 13 other responses, as all the items with a word body *ance* were pronounced as /#ns/ (like in *chance*, *dance* etc.), which was a response not given by any of the participants, these items were all pronounced with a regular vowel in the human data. Additionally, three items were pronounced according to antibody segments – *yeab* and *yead* as /j8b/ and /j8d/ / (based on *yeah*) and *swean* as /swEn/ (based on *sweat*). WSP-token performed similarly, except that there were only 12 mismatches in the other category, nine items with the word body *ance* and the same three antibody-segment based pronunciations as in the WSP-type's output. The WSP-token version also produced slightly more irregular responses (15 items) than regular responses (10 items) when the majority of the participants produced the opposite type of response.

Note, however, that the other responses for items with the word body *ance* in the CDP++, WSP-type and WSP-token models' output are based on the pronunciations corresponding to this word body in the Celex data base (Baayen et al., 1995). As such, rather than categorising these responses as errors, they are better understood as PSCs appropriately learnt by these models, but that are a mismatch to the PSCs utilised by the participants in Treiman et al. (2003) data, due to dialect differences.[18]

Next, I compared the context sensitivity scores in the human data to those produced by the computational models. In the human data, some item groups elicited far more context sensitive pronunciations than others. An important question in evaluating the performance of a computational model is, then, whether it produces the same pattern of context sensitive responses as humans do and whether the model's context sensitivity scores are of a similar magnitude to those found in the human data. To this end, the context sensitivity scores from the human participants were correlated with those produced by each model and the root mean square error (RMSE) was computed for the context sensitivity scores from each model, relative to the human data (see Table 3.3). In addition to comparisons between the DRC, the CDP++ and the two versions of the WSP model, Table 3.3 also includes all the models evaluated in Treiman et al. (2003). The output from the DRC model in this table is based on simulations from the version 1.2.3, which yielded identical results to the earlier version

[18] Pronouncing the word body *ance* as /#ns/ is not a ubiquitous pronunciation in British English, but this correspondence is followed unchanged in the WSP and CDP++, whereas the pronunciation for words containing this particular correspondence was changed in the vocabulary of the DRC model (as stated in the DRC's documentation, retrieved in How DRC 1.2 Differs from DRC 1.0 (wordpress.com), Appendix 1, Derivation of DRC 1.2's Vocabulary)

reported in Treiman et al., 2003, i.e., context sensitivity values of 0 for all item groups. The human context sensitivity scores had the strongest, positive correlation with the context sensitivity scores produced by the two versions of the WSP model (WSP-type: $r(6) = .68$, $p = .06$; WSP-token: $r(6) = .72$, $p = .04$). These versions of the WSP model also had the smallest RMSE. Apart from the WSP model, only the correlations between human data and Multiple-levels model (Norris, 1994, WS parameters and Norris, 1994, TM parameters in Table 3.3) approached significance ($r(6) = .67$, $p = .07$ for WS parameters and $r(6) = .65$, $p = .08$ for TM parameters).

In summary, models that include larger unit size PSCs have a better overall performance with this data set than models that do not. Furthermore, it seems that models with explicit focus on antibody and body segments, such as the WSP and the Multiple-levels model, capture more of the patterns in the human naming data than models where the unit size is more flexible, emerging as a result of training rather than being a fixed part of the model architecture (i.e., connectionist models such as CDP++ or PDP models by Plaut et al., 1996).

Comparisons between the WSP-type, the WSP-token, the CDP++ and the newest version of the DRC model revealed that the two versions of the WSP model have stronger performance in four out of five of the assessment criteria used: the WSP outperformed the other models in proportion of matches to human modal responses for all the items and the irregular items alone, as well as in human-model correlation and RMSE for context sensitivity scores. The WSP-token showed stronger performance in almost all of these measures compared to the WSP-type, except for the RMSE, which indicated that the WSP-type produced slightly more similar absolute values to the human context sensitivity scores than the WSP-token did. The DRC model outperformed the other models in proportion of matches to human modal responses for regular items, but its ability to account for naming responses for irregular items was considerably weaker than that of the other models. Even though the CDP++ model produces context sensitive responses to some of the items in this data set, the pattern of context sensitivity scores for each item set tended to be opposite to the pattern found in the human data, as indicated by a (statistically non-significant) negative human-model correlations of the context sensitivity scores. The CDP++ did not have the strongest performance in any of the assessment criteria used, although it produced a higher proportion of matches to irregularly pronounced items than the DRC model did.

Table 3.3

Context sensitivity scores by computational models and participants for Treiman et al. (2003) data set

	Item group								Human-model correlation	Human-model RMSE
	CV1	*CV2*	*VC1*	*VC2*	*VC3*	*VC4*	*VC5*	*VC6*		
Human data	.58	.16	.55	.86	.12	.33	.83	.70		
DRC	.00	.00	.00	.00	.00	.00	.00	.00	-	-
CDP++	.56	.60	.90	.40	.70	.80	1.00	.40	-.11	0.39
WSP-type	.33	.50	1.00	.80	.20	.40	.90	.60	.68	0.23
WSP-token	.33	.50	1.00	.90	.20	.20	.80	1.00	.72 *	0.25
Zorzi et al., 1998	.00	.60	.90	.20	.00	1.00	1.00	.20	.06	0.48
Plaut et al., 1996, Simulation 2	.33	.50	1.00	.30	.60	.50	.90	.80	.15	0.35
Plaut et al., 1996, Simulation 3	.89	.80	.90	.20	.40	.70	.80	.70	-.08	0.40
Plaut et al., 1996, Simulation 4	.44	.00	.60	.10	.20	.50	1.00	.60	.51	0.30
Plaut & McClelland, 1993	.56	.50	1.00	.70	.00	.90	1.00	.70	.61	0.30
Powell et al., 2001	.56	.60	.70	.30	.10	.70	1.00	.80	.42	0.30
Harm & Seidenberg, 2003	.56	.70	.90	.70	.40	1.00	1.00	.80	.41	0.35
Norris, 1994, WS parameters	.00	.10	1.00	.90	.30	.50	.90	1.00	.67	0.30
Norris, 1994, TM parameters	.00	.20	1.00	.90	.30	.40	.80	1.00	.65	0.29

Note. Item groups CV1 and CV2 are antibody-items (e.g., *wark*, *tark*) and item groups VC1-VC6 are word body-items (e.g., *chead*, *cheal*). Human-model correlation = correlation between each model's and human participants' context sensitivity scores, calculated as a proportion difference of irregular pronunciations between irregular and regular items within each of the eight item groups. Human-model RMSE = the root mean squared error of each model's context sensitivity scores, relative to those in the human data. The human data and output from all other computational models except for DRC, CDP++, WSP-type and WSP-token are taken from Treiman et al. (2003). * denotes statistical significance at alpha level .05.

3.2.3 Pritchard set

This data set consists of 412 nonwords for which the output from the DRC and CDP+ (Perry et al., 2007) models differed (Pritchard et al., 2012). The nonwords were read aloud by 45 Australian participants. The resulting naming responses were categorised as the first, second, third and fourth or lower most popular pronunciations. I investigated the incidence of irregular word bodies in this data set by using the PSC knowledge of the WSP model: any word bodies with a matching vowel pronunciation between the vowel segment alone and the word body segment were considered regular, while mismatch in the vowel pronunciations between these two segments were considered irregular. After removing 18 items that did not have a word body segment in the WSP model's PSC knowledge (e.g., *deche* or *floz*) and thus could not yield different pronunciations between the vowel and body segments, a total of 302 items (76.65% of the considered items) had regular bodies, and 92 items (23.35% of the considered items) had irregular bodies. Furthermore, comparing the vowel pronunciations of the human modal responses to the context insensitive vowel pronunciations from the WSP model revealed that 86.89% of the items received a regular vowel pronunciation as a human modal response. By contract, 8.98% of the items received an irregular pronunciation as the human modal response (this included both antibody and word body analogies)[19]. As such, successful simulation of a large proportion of this data set does not require the use of larger PSCs.

As most of the nonwords received several different pronunciations in this data set, the human modal response, the second most common response, the third most common response and the fourth or lower most common responses were extracted from the data set. The proportion of items for which the first, second, third and fourth most common human responses agreed with a model's output were then calculated. Additionally, the proportion of mismatches were calculated, i.e., items for which a model's pronunciation was not produced by any of the participants. Table 3.4 summarises the proportion of matches for each model. As seen in this table, the DRC model outperforms the other models both in terms of the proportion of human modal responses and the total proportion of items regardless of their frequency category (1^{st}, 2^{nd}, 3^{rd} or 4^{th}). The proportion of mismatches is also the lowest for the DRC model. The

[19] These calculations differ from the categorisation of participant responses in Pritchard et al. (2012), because I only considered vowel pronunciations and the human modal responses and because I categorised regularity based on the WSP model's PSCs, whereas Pritchard et al. considered the whole pronunciation of nonwords from all participants and based their regularity on the output from the DRC model. Thus, multi-letter graphemes such as *igh* pronounced to rhyme with *sigh* was a regular response in Pritchard et al., but irregular in the current analysis.

performance of the WSP-type was the second strongest, closely followed by the WSP-token, and finally the CDP++. The performance of the CDP++ model was considerably lower than that of the other models.

Table 3.4

Proportion of matching naming responses between human participants and the DRC, CDP++ and WSP models for Pritchard et al. (2012) data set

Model	First	Second	Third	Fourth	Total	Mismatch
DRC	.74	.15	.05	.05	.99	.01
CDP	.38	.18	.10	.09	.74	.26
WSP-type	.68	.16	.05	.06	.95	.05
WSP-token	.66	.16	.06	.06	.94	.06

A model's pronunciations that no human participant agrees with are the most concerning ones and a closer inspection of these mismatches may provide valuable insight into the aspects of the model's print-to-sound conversion that need adjustment. I identified several classes of pronunciations in the output from the two versions of the WSP model. Similar investigations for the DRC and CDP++ models are reported in Pritchard et al. (2012). For the WSP model, the largest group of items with mismatches were nonwords with the onset *th*. The WSP's pronunciation for several of these items was /D/ (as in *this*), based on the antibody parsing style. For instance, both WSP-type and WSP-token pronounced nonwords *thaque* and *thet* as /D{k/ and /DEt/, respectively. By contrast, the most typical pronunciation assigned to items with *th*-onset amongst human participants was /T/ (as in *think*). Campbell and Besner (1981) also report human tendency to pronounce the onset *th* as /T/ rather than as /D/ in nonwords. The PSC *th* → /D/ in existing words is mostly found in function words (e.g., *the, that, these*), which tend to have higher token frequency than content words. Thus, it is not surprising that the WSP-token produced more mismatches of this category (for 12 items) than the WSP-type (for 7 items).

Another category of mismatches was irregularly pronounced word bodies, such as *dauche* pronounced as /d5S/ (based on *gauche*), or *scrolk* pronounced as /skr5k/ (based on *folk* and *yolk*). With these items, WSP's pronunciations are driven by the most consistent word body sized segments, even when only a single exemplar is available (such as *gauche,* as the word bodies in these items have a perfect consistency). Human participants were not this sensitive to word body consistency, which resulted in mismatches for six items for both versions of the

WSP model. The third category of mismatches was onset *g* followed by the vowel *e*. There were three items for which both versions of the WSP model gave a soft pronunciation, such as *geech* pronounced as /_iJ/, whereas no human did this. Considering the statistical properties of English monosyllabic words, the soft pronunciation of *g* is expected when it precedes certain vowels (*e* or *y*). However, skilled readers do not seem to show similar sensitivity to it, a finding also reported by Treiman, Kessler and Evans (2007). There were also a handful of items without a clear category. Finally, it is worth noting that four out of the six items that were mismatches for the DRC model (*frymph*, *geech*, *gert*, and *que*) were also mismatches for both versions of the WSP model, as the two models pronounced these items the same way.

As unusual responses to nonwords are given not only by computational models, but also by skilled readers, another way of quantifying a models' performance against human data is to treat each model as a 'participant' and compare the similarity of the model's responses to each of the participants' responses. To gauge what kind of similarity would be acceptable for a computational model, these human-model measures of similarity can be compared to the similarity of each participant's responses with that of other participants (an approach taken by Mousikou et al., 2017). Thus, I calculated the number of matching pronunciations each participant shared with each of the other participants in the Pritchard data set. The matches between any human participant and each of the models were then calculated separately for each model, in order to exclude potential model-model matches. Table 3.5 summarises the similarity between participants and the models calculated this way. Most importantly, the minimum, maximum or mean proportion of matches between any two human participants were very similar to the proportions of matches between any human participant and the DRC or the WSP models. By contrast, the proportion of matches between any participant and the CDP++ model was below that of human participants in all of these measures. As such, while the performance of the DRC and the WSP models (both type and token versions) were well within the range of an individual participant according to this analysis, the performance of the CDP++ model fell below it. Given that the minimum and the mean proportion of matches to participants was generally higher for the DRC and the WSP models than for the human-human matches, these results can also be interpreted as the DRC and the WSP models being more likely to match a randomly chosen participant than a randomly chosen participant would be.

Table 3.5

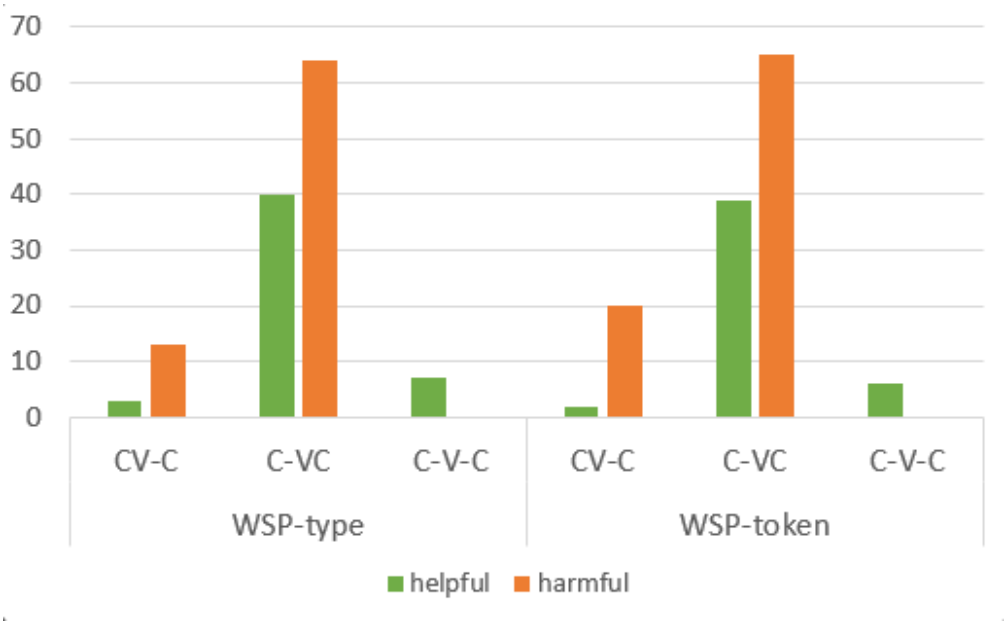
Proportion of matching human-human pronunciations and human-model pronunciations for each participant in Pritchard et al. (2012) data set

	Human	DRC	CDP++	WSP-type	WSP-token
min	.24	.29	.17	.28	.26
max	.68	.68	.38	.64	.56
mean	.44	.52	.30	.48	.44

Finally, as WSP model's output is a result of competition between different parsing styles, it is important to ask whether this competition results in the kind of pronunciations humans produce. This was investigated by contrasting the model's final output with the different pronunciations (corresponding to different parsing styles) available for the model. As the Pritchard set consisted of a large number of nonwords with regular bodies, there is a considerable overlap between responses based on small or large segment parsing styles in the WSP model (i.e., all parsing styles result in the same pronunciation). At the same time, this data set still has a sizeable sample of irregular items, where the word body parsing style results in a different pronunciation compared to the small segment or antibody parsing styles. As such, the Pritchard set lends itself well to investigations of whether the WSP model uses the same reading style that humans do for different nonwords. To answer this question, the responses from the WSP model (both type and token versions) were categorised as helpful or harmful, based on whether the model's response to each item matched the response produced by the highest number of human participants, compared to other pronunciations based on other parsing styles available for the WSP model. Only items where the pronunciation corresponding to the winning parsing style differed from those corresponding to the other parsing styles were considered, as these items would show the unique contribution each parsing style had in increasing or decreasing the model's performance (See Figure 3.1 for a summary).

Figure 3.1

Unique contribution of parsing styles by two versions of the WSP model (deterministic mode) with Pritchard et al. (2012) data set



Note. CV-C = antibody-coda parsing style, C-VC = onset-word body parsing style, C-V-C = onset-vowel-coda parsing style. Y-axis depicts the number of items that either increased (helpful) or decreased (harmful) the number of participants agreeing with the WSP model's output, compared to other pronunciation options (based on other parsing styles) available for the WSP model.

The overall pattern of results from both WSP-type and WSP-token was very similar, and therefore the following description of the findings applies to both versions of the model. The analysis revealed that the word body parsing style won most often, followed by the antibody parsing style, whereas the small segment parsing style won only a few times. The word body parsing style resulted in clearly higher number of uniquely helpful pronunciations compared to the other two parsing styles. However, the number of harmful pronunciations produced by the word body parsing style were almost twice the number of helpful pronunciations. When the antibody parsing style won, it produced harmful pronunciations four times as often as it produced helpful pronunciations. By contrast, the small segment parsing style did not produce harmful pronunciations at all, but the number of helpful pronunciations was also low. Overall, each parsing style was uniquely beneficial for the WSP's performance, but the model also overestimated the incidence of pronunciations based on word body and antibody parsing styles in this data set. See the Discussion (Section 3.5.1) for further consideration of these findings.

3.3 Evaluation of WSP model's variable mode

As described in Chapter 2 (Section 2.1.2), two methods can be used for extracting proportions of different pronunciation options for a nonword from the variable mode of the WSP model – the raw probabilities method and the multiple simulation runs method. The proportions from the raw probabilities method can be thought of as the probabilities that are expected from an infinite number of participants, if the participants' knowledge of PSCs is extracted from English monosyllabic words. The benefit of this method is that a stable estimate of the proportions is achieved, as the same nonword will always receive the same proportions of pronunciation options. However, the disadvantage of this approach is that the proportions do not vary – an unrealistic feature, considering the variability in nonword reading demonstrated by empirical studies (e.g., Coltheart & Ulicheva, 2018; Ulicheva et al., 2021). The proportions for pronunciation options obtained from the multiple simulation runs method, by contrast, will be different every time output is generated using this method. While this approach allows more realistic production of responses from a 'group of participants', the weakness of this approach is the variability of the proportions achieved – namely, how will the model's performance be assessed against human data, when the simulation data can be very different from one set of simulation runs to the next? While the available data from human participants might be very different from a particular set of simulation runs, there might be a group of human participants that would produce comparable proportions. In order to gain some idea about the range of performance using the multiple simulation runs method, five sets of simulation runs were generated for each data set the WSP model was tested on.

The performance of the WSP's variable mode on the three data sets used for testing the WSP's deterministic mode will be described next, using both raw probabilities and multiple simulation runs methods. For the Andrews and Scarratt set, proportions of regular pronunciations and proportions of irregular pronunciations for each item in the human data and those in the model output were compared separately. Additionally, the proportion differences between irregular and regular pronunciations for each item were also compared, as this measure should capture more about the relative preference for either pronunciation type in the human data (given that the responses for each item consisted of a proportion of regular, irregular and other responses). For the Treiman set, the correlation of the context sensitivity scores in human data and model output were used. For the Pritchard set, the comparisons between human and model pronunciations were made separately for the 1st, 2nd

and 3rd most frequent human responses for each item, if such were available (some items only had one or two pronunciation options in the human data). For instance, if the model had produced a matching pronunciation for an item in the 1st response group (e.g., *brask* pronounced as /br{sk/), this pronunciation would count as a match and the proportion associated with this pronunciation option in the model's output would be compared to the proportion of human participants that had produced this pronunciation. Similarly, if the model had produced a matching pronunciation for an item in the 2nd response group (e.g., *brask* pronounced as /br#sk/), this pronunciation would count as a match in the 2nd response analysis and the proportion associated with this pronunciation option in the model's output would be compared to the proportion of human participants that had produced this pronunciation. For this analysis of the Pritchard set, only responses given by at least three participants were included. This stricter criterion for valid responses was chosen in order to gain more reliable response options and proportions for them, as responses given by only one or two participants may be a result of pronunciation or transcription errors.

3.3.1 Evaluation of the WSP model's variable mode against three data sets

3.3.1.1 Raw probabilities method. The performance of the WSP model's variable mode using the raw probabilities method was compared against the naming responses from Andrews and Scarratt set, Treiman set and Pritchard set. The outcomes of these comparisons are summarised in Table 3.6, for both WSP-type and WSP-token versions. As seen in the table, the two versions produce very similar results.

Table 3.6

Comparison of WSP model (variable mode – raw probabilities) against three data sets of human nonword reading

Data set	Item group	Human-model correlation		RMSE		Match proportion	
		WSP-type	*WSP-token*	*WSP-type*	*WSP-token*	*WSP-type*	*WSP-token*
Andrews & Scarratt set	regular	.57 *	.58 *	0.33	0.32	1.00	1.00
	irregular	.32	.35	0.31	0.30	1.00	1.00
	reg-irreg diff.	.45	.48	0.60	0.58	1.00	1.00
Treiman set		.77 *	.74 *	0.26	0.27	-	-
Pritchard set	1st response	.37 *	.37 *	0.28	0.28	0.92	0.92
	2nd response	.21 *	.21 *	0.33	0.33	0.40	0.40
	3rd response	.24	.23	0.37	0.37	0.26	0.26

Note. Match proportion for Andrews and Scarratt set is 1.00 as the WSP always produced the regular and the irregular pronunciation options found in the human data. Match proportion for Treiman set is not calculated because the proportion differences of irregular pronunciations for the regular and irregular items within each item group were of interest in this analysis, not whether the human modal response or less frequent human responses were found as the WSP's pronunciation options. * denotes statistically significant correlation at alpha level of .05. RMSE = the root mean squared error of each model's proportion for a given pronunciation option, relative to those in the human data.

In the Andrews and Scarratt set, the proportions of regular responses as well as the proportion differences between irregular and regular responses were fairly similar to the corresponding proportions produced by the WSP model, as demonstrated by moderate, positive human-model correlations. For instance, the two items that were pronounced regularly by all the participants (*ginth* and *jinth*), also received the highest proportion of regular pronunciations by the WSP-token (.81 and .72, respectively) and amongst the highest proportions of regular pronunciations by the WSP-type (.81 and .70, respectively). However, the human-model correlations for the irregular responses were somewhat lower. For instance, the two items that were pronounced irregularly by over 99% of the participants (*zight* and *searn*) received much lower proportions of irregular responses from the two versions of the model: WSP-token (.41 and .65, respectively) and WSP-type (.39 and .64, respectively). Closer inspection of the proportions of regular and irregular responses revealed that the model tended to overestimate the proportions for regular responses: for 13 out of 16 items in the output from both versions of the model the regular proportions were larger than the irregular proportions. By contrast, this was the case for only three items in the human data.

In the Treiman set, the human-model correlations of context sensitivity scores were strong for both versions of the WSP model. The six word body item groups were particularly well matched: the largest context sensitivity scores found in the human data tended to also be the largest in the model's output, and the smallest sensitivity scores in the human data were the smallest in the model's output. However, the model's context sensitivity scores for the two antibody item groups were less similar to those from humans: the larger score in human data for these items (.58) was the smaller in the model output (WSP-type: .36, WSP-token: .35) and the smaller score in the human data (.16) was larger in the model output (WSP-type: .37, WSP-token: .36). In other words, the model produced more context sensitive vowel pronunciations to items for which humans showed less context sensitivity, and vice versa. Finally, the absolute values of the context sensitivity scores varied less in the model output (WSP-type: .27-.46; WSP-token: .29-.48) compared to the range of values in the human data (.12-.86).

In the Pritchard set, the human-model correlations for the proportions of naming responses were modest. Firstly, the proportion of items that the WSP model pronounces the same as humans do declines as the pronunciation options amongst humans become less common – while 1st responses by humans are matched by the WSP for 92% of the items, only 40% of the items are matched for the 2nd most common human pronunciations. Secondly, the human-model correlations for the 1st responses remain just under .4 for both versions of the WSP, with even weaker correlations to the 2nd and 3rd pronunciations.

To provide an estimate of the strength of human-model correlations that should be expected from a successful model, I extracted human-human correlations from two groups of participants, in the following way: I split the participants in the Pritchard data set into two groups – participants with an odd participant number (n = 27) and participants with an even participant number (n = 26). Then, I calculated the proportions of participants in each group that produced the 1st pronunciation option for each item, the proportions for the 2nd pronunciation option for each item, from both groups, and so on for the 3rd pronunciation option. As some items only received one or two different pronunciation options, the number of items to be correlated in these comparisons were 412 for the 1st pronunciation options, 407 for the 2nd option and 393 for the 3rd option. The correlations for the proportions of the pronunciation options between the two groups of participants decreased from the 1st pronunciation option ($r(410) = .83$, $p < .001$) to the 2nd pronunciation option ($r(405) = .52$, p

< .001) and the 3rd option ($r(391) = .22$, $p < .001$). Compared to these human-human correlations, the WSP model is clearly below adequate levels of performance.

Closer inspection of the type of items for which the human-model correlations were particularly low may shed light on how to improve the model. There were 30 items (for WSP-token, 28 for WSP-type) in the 1st response category for which the difference between the human and the model proportions was over .5. Nearly two thirds of these items received such different proportions between humans and the WSP model because the WSP produced less pronunciation options for them than humans did. Both versions of the WSP model produced only one pronunciation option for 19 of these items, thus resulting in a proportion of 1, which was much higher than the proportion of participants producing the corresponding pronunciation for these items. While no clear, dominant categories were found for these 19 items, three types of pronunciations are worth noting. Firstly, there were four items with the vowel *au* (e.g., *wauce*, *shrauk*) for which the second most popular human response was /6/. The PSC *au* → /6/ does not exist in the WSP's repertoire, and it is not a common PSC in English in general (cf. *ablaut*, *degauss*, *Gaussian*, *Nauru*, *Saudi Arabia*, and a few loan words such as *sauerkraut*). Secondly, for a handful of items, the second most common human responses showed context sensitivity in the antibody segment, such as *wa* pronounced as /wQ/ in items *thwalc* and *thwazz* or yod[20] included in the pronunciations for items *bune* or *gneuth* (i.e., pronounced as /bjun/ and /njuT/). The *thwa* antibody segment is part of the WSP's PSC knowledge, but it only occurs in one word, *thwack*, which does not reflect the context sensitive vowel pronunciation. Yod, on the other hand, only exists as a less frequent option for onsets, and as such it is not mapped onto the following vowel phoneme consistently enough to produce pronunciations with yod[21]. Thirdly, there were items that elicited so many different naming responses in humans that the proportion for any one pronunciation option was very small (e.g., *tuise*, with the 1st response proportion of .16).

Apart from the items for which the WSP model produced far less pronunciation options than humans did, there were seven (WSP-type) and nine (WSP-token) items for which the participants gave nearly unanimous naming responses (and thus close to 1 as a proportion for these pronunciations), whereas the model produced more pronunciation options and thus smaller proportions for each of them. Five of the items in this category were due to the word

[20] E.g., the /j/ in the pronunciation of *dune* → /djun/

[21] The yod pronunciations could be added into the WSP output as phonotactic constraints, as they precede the vowel sound /u/ in the presence of certain onsets (e.g., /d/ as in *dune* or /t/ as in *tune*) but not others (e.g., /J/ as in *chew* or /r/ as in *rude*).

body *olk* receiving almost exclusively the regular pronunciation /Qlk/ from the participants. By contrast, the proportions for this pronunciation option from the model was at most .34 (by both versions of the model), while the irregular pronunciation /5k/ for items with this word body tended to get higher proportions than the regular options. There were also a handful of items for which the onset of the nonword received a nearly unanimous pronunciation amongst the participants (e.g., *th* as /T/ or *g* as /g/) whereas the model's proportions for these items reflected the two competing options (/T/ or /D/ and /g/ or /_/), thus producing much lower proportions for these items.

Human nonword reading responses as phonemic transcriptions naturally reflect not only the type of PSC knowledge skilled readers have, but also fatigue, mispronunciations and transcription errors. Thus, it is not surprising that proportions for different pronunciation options in the human data should often be smaller than those produced by a computational model, due to a larger number of different pronunciations generated by skilled readers. This is demonstrated in Table 3.7, which depicts statistics for the number of pronunciation options produced by the participants and the WSP model for the Pritchard set, as well as the number of items in the Pritchard set that received more, less or the same number of pronunciation options by the humans compared to the number of pronunciation options by the WSP model. The same number of pronunciation options was produced for each item by the WSP-type and WSP-token versions of the model.

As seen in this table, skilled readers produce more pronunciation options for a vast majority of the items than the model does. However, there are some items for which the model produces more options. The proportions from participants and the WSP model could be brought closer together for the former category of items, where the human proportions for different pronunciation options are lowered partly due to human error, by introducing sources of error also in the model’s output (e.g., a non-zero probability for making letter confusions).

Table 3.7

Number of pronunciation options produced for Pritchard et al. (2012) nonword set by human participants and the WSP model (variable mode)

	Human – full	Human – min. 3	WSP
Mean (SD)	8.39 (4.5)	2.59 (1.22)	1.9 (0.78)
Range	1 - 24	1 - 7	1 – 5
H > M	395	237	-
H < M	4	54	-
H = M	13	121	-

Note. Human – full = all the human responses are included; human – min.3 = only responses produced by at least three participants are included; (H > M) = number of items for which humans produce more pronunciation options than the model; (H < M) = number of items for which humans produce less options than the model; (H = M) = number of items for which humans and the model produce the same number of options.

Increasing the performance of the model for the latter category of items – where the model generates pronunciation options that are not present in the human responses – requires some consideration of why these differences in the number of pronunciation options occur. As the model's pronunciation options are combinations of all the pronunciations that are associated with a given orthographic segment consistently enough (in the current models, the threshold is consistency of .3), sometimes this generates pronunciation options that are not produced by humans, such as *brolk* pronounced as /brQlk/, /br5k/ and /brQk/. The last option is not a likely response from skilled readers (although there was one participant producing this response in the Pritchard set), but it is produced by the WSP model because the coda *lk* is associated with the pronunciation /k/, due to items like *folk* and *yolk* etc. in the WSP's vocabulary, which makes this correspondence consistent enough to be included in the assembly of pronunciation options. On the other hand, this procedure allows inclusion of important pronunciation options that do exist in human responses, such as *ces* pronounced as /kEs/, /sEs/, /kEz/ and /sEz/ (all of which were also produced by human participants in the Pritchard set). Additionally, for some items this procedure produces both likely and unlikely pronunciation options, such as for the nonword *thelm.* The pronunciation options for this item produced by the model are /TElm/, /DElm/, /TEm/ and /DEm/. As already discussed above (Section 3.2.3), the existence of pronunciation options /T/ and /D/[22] for an onset segment *th* is

[22] Pronunciation options /D/ and /T/ for the onset *th* are due to the antibody parsing style (particularly for segments *the* and *tha*), the onset *th* - /D/ in isolation does not have high enough consistency to be included when the threshold for including a PSC in assembling pronunciation options is .3.

problematic, as humans rarely produce the latter option when reading nonwords in isolation. However, the coda options /m/ and /lm/ are important, because the former is the default pronunciation for *lm*, due to there being slightly more words like *calm, palm* etc., where the *l* is silent than words like *film, realm* etc. where the *l* is pronounced in the WSP's vocabulary. Thus, inclusion of less consistent PSCs is beneficial for the coda in this case.

This brings us to an important consideration about the small segment options in the WSP model, namely, that PSCs for onset and coda clusters are treated as a single unit, rather than as graphemes. Some of the unlikely pronunciation options in the WSP's variable mode could be avoided by increasing the threshold for PSCs included in the assembly of pronunciation options. If the onset and coda segments were based on graphemes, some of the likely pronunciation options, such as *lm* → /lm/ would be included while less likely options (*lm* → /m/) would be left out. However, before modifying the model, it should be ensured that the proportions and pronunciation options in the human data are indeed representative of what different samples of skilled readers produce. This question will be addressed in the Discussion section.

3.3.1.2 Multiple simulation runs method. Out of the five sets of simulation runs for each of the three data sets, Table 3.8 summarises the results for the sets of simulation runs that showed the best and the poorest performance. The choice of the best and the poorest performance was based on the human-model correlations for the 1st responses in the Pritchard set, the proportion difference between regular and irregular items in the Andrews and Scarratt set and the human-model correlation of context sensitivity scores in the Treiman set. Each set of simulation runs was matched with the sample size of the relevant nonword naming data set (i.e., 45 runs for Pritchard set and 24 for Andrews and Scarratt set and Treiman set each).

What can be seen in Table 3.8, is that there is considerable variability in the level of performance different sets of simulation runs have. However, the performance in the Pritchard set does not vary as much as the performance in the other two, smaller data sets. For the Andrews and Scarratt set, the correlations of proportions of regular, irregular and the difference between regular and irregular items from the best performing set of simulation runs (both versions of the model) tended to be slightly higher than those obtained with the raw probabilities method. For the Treiman set, the best performing set of simulation runs produced slightly lower correlation to the human context sensitivity scores from WSP-type, but higher correlation from WSP-token, compared to the correlations from the raw

probabilities method. Finally, the best performing set of simulation runs from both versions of the model produced higher human-model correlations for the Pritchard set than those obtained with the raw probabilities method.

Table 3.8

Comparison of WSP model (variable mode – multiple simulation runs) against three data sets of human nonword reading

Data set	Item group	Human-model correlation		RMSE		Match proportion	
		WSP-type	*WSP-token*	*WSP-type*	*WSP-token*	*WSP-type*	*WSP-token*
		Best performing set of simulation runs					
Andrews & Scarratt set	regular	.57 *	.78 *	0.34	0.31	1.00	1.00
	irregular	.41	.53 *	0.30	0.28	1.00	1.00
	reg-irreg diff.	.50 *	.68 *	0.61	0.55	1.00	1.00
Treiman set		.74 *	.86 *	0.28	0.28	-	-
Pritchard set	1st response	.45 *	.45 *	0.30	0.30	.92	.92
	2nd response	.44 *	.44 *	0.23	0.23	.34	.34
	3rd response	.25 *	.25 *	0.17	0.17	.15	.15
		Worst performing set of simulation runs					
Andrews & Scarratt set	regular	.32	.36	0.34	0.35	1.00	1.00
	irregular	.12	.11	0.33	0.33	1.00	1.00
	reg-irreg diff.	.19	.22	0.64	0.65	1.00	1.00
Treiman set		.49	.33	0.27	0.28	-	-
Pritchard set	1st response	.43 *	.44 *	0.30	0.30	.92	.92
	2nd response	.43 *	.44 *	0.23	0.23	.34	.34
	3rd response	.24 *	.24 *	0.17	0.17	.15	.15

Note. Match proportion for Andrews and Scarratt set is 1.00 as the WSP always produced the regular and the irregular pronunciation options found in the human data. Match proportion for Treiman set is not calculated because the proportion differences of irregular pronunciations for the regular and irregular items within each item group were of interest in this analysis, not whether the human modal response or less frequent human responses were found as the WSP's pronunciation options. * denotes statistically significant correlation at alpha level of .05. RMSE = the root mean squared error of each model's proportion for a given pronunciation option, relative to those in the human data.

Overall, it might be difficult to see the benefit of using the multiple simulation runs method at all, since the output from one set of simulation runs to the next can be so variable (especially for small data sets). However, the next sections demonstrate the value of the multiple simulation runs method in more detail.

3.3.2 Comparison of the WSP model and Zevin and Seidenberg (2006) model

The variable mode of WSP model was an attempt to address a gap in the current modelling work, which mostly neglects variability in nonword reading – both within and between participants. Exceptions to this pattern are a handful of modelling studies used in investigations of dyslexia (Perry et al., 2019; Rueckl, Zevin & Wolf VII, cited in Compton et al., 2019; Welbourne et al., 2011; Ziegler et al., 2008). For instance, in the studies by Perry et al. (2019) and Ziegler et al. (2008) the performance on different components of reading (such as access to orthographic lexicon or phoneme processing) from normally developing or dyslexic children were simulated by adding noise to the corresponding components in a computational model or by hindering learning in these components, such that the impairment of the different components were proportional to the level of deficits in the individual child's performance in these components. Reading aloud output from these individualised models were very similar to the reading performance from the group of developing or dyslexic children.

However, to my knowledge, only the work by Zevin and Seidenberg (2006) covers simulating individual differences in skilled readers. As the current PhD project focuses on non-pathological reading performance, the model developed in Zevin and Seidenberg (2006) is compared to the variable mode of the WSP model. Zevin and Seidenberg used a modified version of the model by Harm and Seidenberg (1999), a connectionist model with orthographic and hidden units, as well as a phonological attractor network, which allows noisy input to gravitate towards phonotactically legal patterns. Zevin and Seidenberg's model (Z&S model from now on) consisted of 133 orthographic, 200 phonological and 100 hidden units. Additionally, 20 clean-up units mediated connections from a phonological unit (corresponding to a set of phonemic features) to itself and to other phonological units. This model was trained multiple times, so that each version of the model was exposed to a slightly different set of words. The same 5870 monosyllabic words were used for training each version of the model, but the number of times each word occurred was random, although based on token frequency of the words. Due to differences in the model's dialect and that of the test sets (Andrews & Scarratt, 1998; Treiman et al. 2003)[23], some items were removed (such as the case 2 CV from Treiman set). As such, I only report the comparisons between Z&S model and the WSP model on the Andrews and Scarratt data set (experiment 2). This

[23] Comparisons to Glushko's (1979) data set were reported as well, but as only response times are available in this data set, it cannot be included in comparisons for WSP, which, in its current form, does not produce response times

data set consists of nonwords with either regular, consistent word bodies (RCB items, e.g., *beal*, pronounced regularly in all words), regular, inconsistent bodies (RIB items, e.g., *heaf*; pronounced regularly in some words such as *leaf* and irregularly in others such as *deaf*), no regular analogy word bodies with many neighbours (NRAM items, e.g., *nalm*; the word body occurs in many irregularly pronounced words) and no regular analogy word bodies that are unique (NRAU items, e.g., *donth*; a word body only occurs in a single, irregularly pronounced word).

The pattern of naming responses from the human participants was compared to the output from the WSP model's variable mode and versions of Z&S model in terms of the proportion of regular pronunciations to each item group and variability in pronunciations assigned to items in each item group. The variability was quantified as H measure of entropy (see Chapter 1, Section 1.1.2). The proportion of regularly pronounced items was calculated based on vowel pronunciation only, using Andrews and Scarratt's definition of regularity (Andrews & Scarratt, 1998, Appendix B). The multiple simulation runs method was used for obtaining variable output from the WSP model and five different sets of simulation runs were generated for both type and token versions of the model. Each set consisted of 24 simulation runs, corresponding to the number of participants in the human data set. As all sets of simulation runs from both WSP-type and WSP-token versions produced the same pattern of results and the same statistically significant results, only the poorest performing group of simulation runs with the lowest human-model correlation across all items in the data set are reported, thus providing a conservative estimate of the WSP's performance.

The analysis of the proportion of regular pronunciations to different item groups revealed that in the human data the proportion of regular pronunciations decreased from one item group to the other, in the following order: RCB > RIB > NRAU > NRAM. (See Figure 3.2). This pattern is expected, given that the number of regular word body neighbours decreases in a graded fashion in these item groups. The difference between the RCB and RIB items was marginally significant in the human data ($t(73.76) = 1.61$, $p = .11$)[24], while the RIB-NRAU difference ($t(29.81) = 6.10$, $p < .001$) and NRAU-NRAM difference ($t(30.83) = 2.49$, $p = .02$) were both statistically significant. Numerically, the simulation runs with WSP-type and WSP-token versions of the WSP model produced the same pattern (See Figure 3.2). Statistically, both WSP-type and WSP-token versions produced reliable RCB-RIB

[24] Welch t-test (2-tailed) is used in all the analyses reported in this section, due to unequal sample sizes of the item groups compared

differences (WSP-type: $t(61.57) = 2.85$, $p = .006$; WSP-token: $t(66.65) = 2.57$, $p = .01$) and RIB-NRAU differences (WSP-type: $t(40.51) = 5.43$, $p < .001$; WSP-token: $t(43.16) = 5.46$, $p < .001$), but non-significant NRAU-NRAM differences (WSP-type: $t(46.00) = 1.11$, $p = .27$; WSP-token: $t(45.98) = 0.81$, $p = .42$). Like the human data, the Z&S model's output also yielded statistically significant RIB-NRAU and NRAU-NRAM differences, but unlike human data, the model also produced a significant RCB-RIB difference. Thus, while both the WSP and the Z&S models reflect the general pattern found in the human data, both models overestimate the proportion of regular pronunciations assigned to items with no regular word body neighbours. Overall, the Z&S model is a closer fit to the human data than the WSP model.

Figure 3.2

Percentage of regular pronunciations assigned to the Andrews and Scarratt nonwords (1998, Exp. 2) by human participants and computational models

Note. reg. pron. % = percentage of regular pronunciations assigned to different item groups; RCB = nonwords with regular and consistent bodies; RIB = nonwords with inconsistent bodies; NRAU = nonwords with unique, irregular bodies; NRAM = nonwords with irregular bodies occurring in several words; Z&S, 2006 = approximate re-creation of the plot of simulation data from Zevin & Seidenberg's model (2006).

Turning to the variability in Andrews and Scarratt set, the variability of pronunciations assigned to different item groups (as measured by H-value) increased from RCB to RIB to NRAU. The NRAM items had slightly lower mean H than NRAU items, but both of the 'no regular analogy' item groups had significantly higher H than the RCB or RIB items (see

Figure 3.3). In the human data, there was just significant RIB-RCB difference ($t(76.20) = 2.03$, $p = .05$), reliable NRAU-RIB difference ($t(31.43) = 4.34$, $p < .001$) and no reliable difference between NRAM and NRAU items ($t(37.55) = -0.29$, $p = .77$).

Figure 3.3

H-values as measures of pronunciation variability in the Andrews and Scarratt nonwords (1998, Exp. 2) by human participants and computational models

Note. H-value = entropy H-value, a measure of pronunciation variability in different item groups; RCB = nonwords with regular and consistent bodies; RIB = nonwords with inconsistent bodies; NRAU = nonwords with unique, irregular bodies; NRAM = nonwords with irregular bodies occurring in several words; Z&S, 2006 = approximate re-creation of the plot of simulation data from Zevin & Seidenberg's model (2006).

Numerically, both versions of the WSP model produced the same pattern of mean H-values than that found in the human data, except that the NRAU items had a slightly lower mean H-value than the NRAM items. Statistically, both versions of the WSP model produced the same pattern of results as found in the human data: reliable RIB-RCB differences (WSP-type: $t(63.47) = 3.25$, p .002; WSP-token: $t(64.21) = 3.07$, $p = .003$) and NRAU-RIB differences (WSP-type: $t(57.68) = 4.39$, $p < .001$; WSP-token: $t(55.06) = 4.58$, $p < .001$), but no reliable differences between NRAM and NRAU items (WSP-type: $t(44.52) = 0.41$, $p = .68$; WSP-token: $t(45.31) = 0.49$, $p = .63$). The Z&S model produced numerically similar pattern of mean H-values as that found in the human data, although the absolute value for the NRAM items was considerably lower in the model output. Statistically, the Z&S model produced a non-significant difference between RIB and RCB items, a significant difference between the items with regular analogies and those with only irregular analogies and a significant NRAM-

NRAU difference[25]. As such, the general pattern of response variability found in the human data is also produced by the WSP and the Z&S models. However, both models underestimate the variability in each of the item categories. Both versions of the WSP model are a closer match to human data than the Z&S model, which is mostly due to a notably lower variability in the NRAM items produced by the Z&S model.

In summary, both Z&S and WSP model produce the same general pattern of regular responses and response variability as found in the Andrews and Scarratt's data (experiment 2). However, the Z&S model is a closer match to human data for proportions of regular pronunciations in different items sets, while the WSP model simulates response variability better. These results will be considered further in the Discussion.

3.3.3 Comparing individual participants to individual simulation runs

The Pritchard et al. (2012) data set provides detailed information about skilled naming responses, as every participant's naming response for every item is available. This level of detail allows further comparisons between these human responses and the output from the WSP model's variable mode, using the multiple simulation runs method.

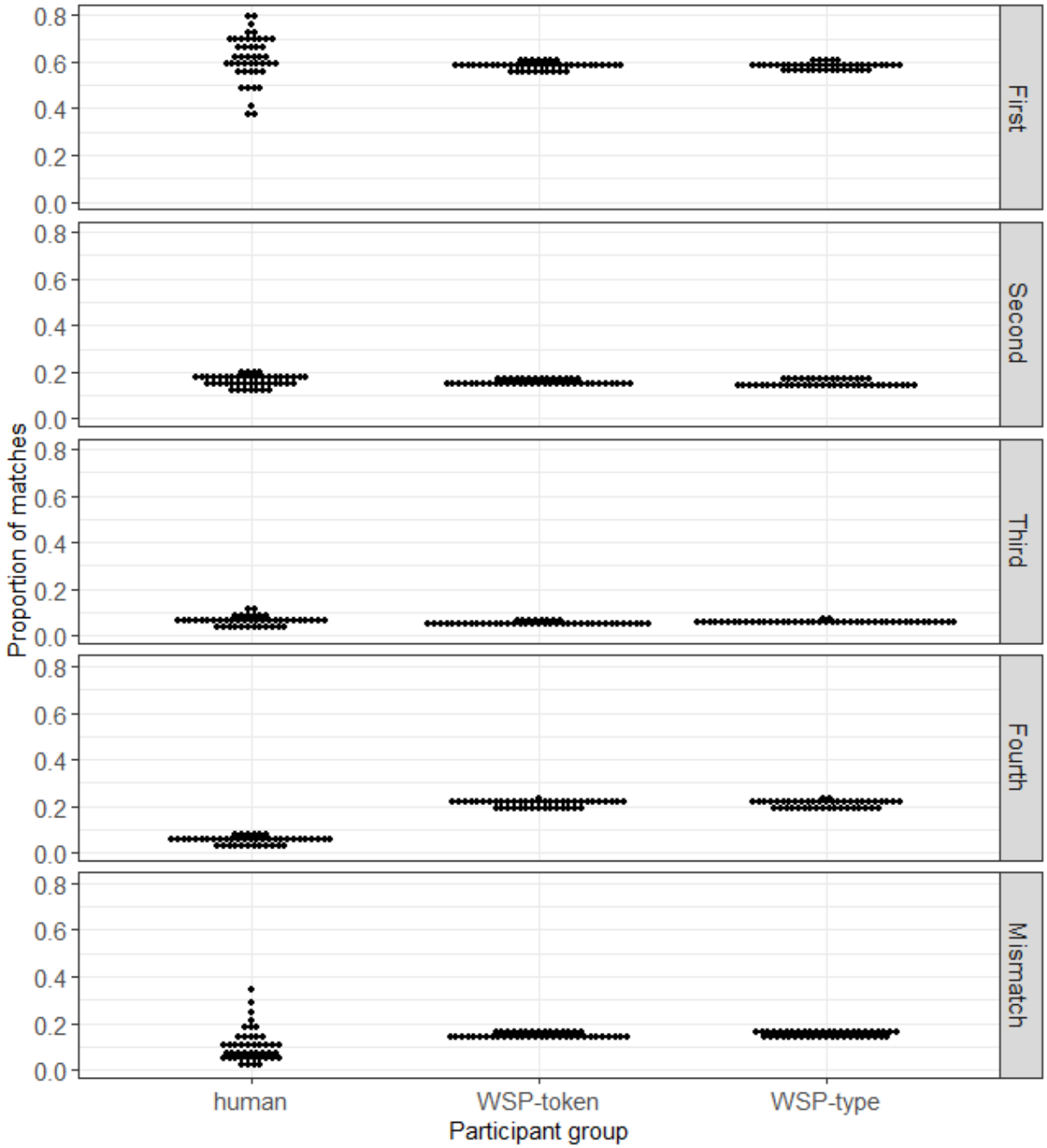
One potentially useful benchmark against which to evaluate the performance of the WSP model's variable mode is how well individual participants match the human modal responses and less frequent response categories and, importantly, for how many items individual participants assign pronunciations that are not produced by any other participant. To produce this type of benchmark, the proportion of naming responses that matched the human modal response, less frequent responses and responses not produced by other participants in the Pritchard set were calculated for each individual human participant. The same proportions were then calculated for each simulation run in the WSP model's output, which consisted of 45 simulation runs, corresponding to the number of participants in the Pritchard set. To use a conservative estimate of the model's performance, the poorest performing sets of simulation runs (out of the five sets generated) were chosen from both the WSP-type and the WSP-token models, based on the mean proportion of human modal responses in the Pritchard set. However, the performance from all five sets of simulation runs from each version of the model were very similar with each other.

[25] I am constrained by the level of detail provided in the Zevin and Seidenberg (2006) – their simulation data is no longer easily accessible (J. Zevin, personal communication, January 26, 2022), and as such the analyses and plots in this section for the human and WSP data were also left without further detail (e.g., confidence intervals in the Figures 3.2 and 3.3).

These comparisons are summarised in Figure 3.4, which depicts the proportion of matches to each response category for each individual participant and each individual simulation run.

Figure 3.4

Proportion of matches to different response categories in the Pritchard et al. (2012) set by individual human participants and individual WSP model simulation runs



Note. Each dot in the graph represents the proportion of responses from a single human participant or a single simulation run that correspond to a specific response category (e.g., human modal response, second most popular response, etc.).

Both versions of the WSP model produced nearly identical proportions in each response category. The results described here thus apply to both versions of the model. Most importantly, the mean proportion of matches to the human modal, the second and the third most common responses produced by the WSP model were almost identical to those produced by human participants, although the average human participant matched the human modal responses slightly better (mean proportion of matches was .61 for humans and .58 for the model). By contrast, the model simulation runs had a higher mean proportion of matches to the fourth or lower response category (humans mean proportion: .05; WSP mean proportion: .21) and higher proportion of mismatches (humans mean proportion: .10; WSP mean proportion: .15). This suggests that the WSP model produces more uncommon pronunciations, on average, than humans do, which is likely related to some of the pronunciation options available to the model (see Section 3.3.1.1 for similar findings and discussion). Overall, however, the model's performance is mostly within the range of skilled readers' performance.

3.4 Evaluation of WSP model optimised for nonword reading data sets

Optimising the WSP model for a nonword reading data set serves two purposes. Firstly, it allows further demonstration of the model's behaviour. This should be particularly beneficial for the performance of the variable mode, where optimisation with a nonword data set may yield more human-like proportions than simply using the same weights that were used for the deterministic mode (as was the case when optimising the model with its vocabulary). Secondly, it may provide insights about the type of reading behaviour the composition of the nonword sets in question might encourage.

The WSP-type and the WSP-token versions of the WSP model were optimised using each of the three nonword reading data sets described above, separately. The performance of the model when optimised for each of these data sets was then compared against all three data sets. The optimisation was performed as described in Chapter 2 (Section 2.3.2), except that this time the success of the output from the model's deterministic version was compared against the number of matches to the human modal response in each data set. Because there were three items with unknown pronunciation (categorised as 'other') as the modal response in the Andrews and Scarratt set, I used the second most common response for these items. The success of the variable output was assessed based on the human-model proportion correlations for the difference in irregular and regular pronunciations in the Andrews and Scarratt set, the context sensitivity scores in the Treiman set and the 1st responses in the

Pritchard set. The weights for the resulting final versions of the model are listed in Table 3.9, for both WSP-type and WSP-token.

Table 3.9

Parsing style weights of the two versions of the WSP model optimised for nonword data sets

Version	Optimisation set	Deterministic mode weights			Variable mode weights		
		cv-c	*c-vc*	*c-v-c*	*cv-c*	*c-vc*	*c-v-c*
WSP-type	*Andrews & Scarratt*	1.90	2.00	1.30	0.10	2.00	1.10
	Treiman	1.70	2.00	1.40	1.80	2.00	0.10
	Pritchard	1.80	1.60	1.50	0.60	1.00	1.90
WSP-token	*Andrews & Scarratt*	1.90	1.90	1.30	0.10	2.00	1.30
	Treiman	1.80	1.90	1.50	1.80	2.00	0.10
	Pritchard	1.80	1.60	1.80	0.40	0.90	2.00

Note. cv-c = antibody-coda parsing style (e.g., *wa-sk*); c-vc = onset-word body parsing style (e.g., *w-ask*); c-v-c = small segment parsing style (e.g., *w-a-sk*).

As seen in this table, the word body sized parsing style is generally the most advantaged in versions of the model that were optimised for the Andrews and Scarratt or Treiman sets. By contrast, the small segment parsing style tended to be advantaged in the versions of the model optimised with Pritchard set. These patterns of weights fit with the general characterisation of the three data sets, namely, that consideration of larger segments is beneficial for the Andrews and Scarratt and Treiman sets, whereas smaller segment reading style is more important for good performance in the Pritchard set. Additionally, as the Treiman set was the only one that also focused on items with irregularly pronounced antibody segments, the considerably larger weights for the antibody parsing style would be expected in the versions of the model that were optimised for the Treiman set compared to the other versions, where this parsing style should generally be less important. However, this was only true for the variable mode of the model, whereas the advantage of the antibody parsing style in the deterministic mode was surprisingly large regardless of the nonword set used for optimisation.

3.4.1 Deterministic mode

The performance of the different versions of the WSP model on the three data sets is summarised in Table 3.10. The vocabulary-optimised versions of the model are also included for ease of comparison.

Table 3.10

Performance of versions of WSP model (deterministic mode) optimised for different nonword reading data sets

	Data set				
	Andrews & Scarratt 1998	**Treiman et al. 2003**		**Pritchard et al. 2012**	
Model version	*1st Human pron match*	*1st Human pron match*	*Human-Model correlation*	*1st Human pron match*	*Total match*
WSP-type-AS	.69	.77	.70	.71	.96
WSP-token-AS	.69	.77	.68	.68	.94
WSP-type-T	.63	.82	.73*	.73	.97
WSP-token-T	.69	.85	.79*	.71	.96
WSP-type-P	.19	.70	-	.79	.98
WSP-token-P	.19	.70	-	.78	.98
WSP-type-vocab	.38	.75	.68	.68	.95
WSP-token-vocab	.56	.78	.72	.66	.94

Note. Model's optimised for AS = Andrews & Scarratt set, T = Treiman set, P = Pritchard set, vocab = WSP's vocabulary. Human-model correlation = correlation between human and model sensitivity scores (difference in proportion of irregular pronunciations for irregular and regular items within each of the eight item groups in Treiman et al., 2003). Human-model correlation for WSP-type-P and WSP-token-P versions of the WSP model could not be computed as these models always produced a regular response to items, thus resulting in a 0 difference between proportions of irregular pronunciations for regular and irregular items in each item group. * denotes statistically significant correlation at an alpha level of .05.

Unsurprisingly, the versions of the model optimised for a given nonword reading data set had the strongest performance on this data set. The versions optimised for Andrews and Scarratt set or Treiman set maintained a relatively strong performance across the three data sets. For instance, the performance on Pritchard set was still generally higher than that of the vocabulary-optimised versions of the WSP, the performance on the Andrews and Scarratt set for Treiman-optimised versions was higher than it was for the vocabulary-optimised versions and the performance on the Treiman set for the Andrews and Scarratt-optimised versions was

generally at least as good as it was for the vocabulary optimised versions. By contrast, while the Pritchard-optimised versions of the model performed very well on the Pritchard set – both WSP-type and WSP-token exceeded the DRC model's proportion of matches for human modal responses – the performance on the other data sets suffered noticeably. For instance, no human-model correlations could be computed for the context sensitivity scores for the Treiman set, as every item was pronounced regularly by the Pritchard-optimised versions of the model, and the proportion of matches for the Andrews and Scarratt set was lower than those for the DRC or CDP++. Comparing the WSP-type and the WSP-token versions of the WSP model as they were optimised for the different nonword data sets, the WSP-token version tended to outperform the WSP-type on the Andrews and Scarratt and Treiman sets, but the WSP-type performed better than the WSP-token on the Pritchard set.

3.4.2 Variable mode

Only the proportions for pronunciation options from the raw probabilities method are reported here for brevity. The performance of difference versions of the model on the three data sets is summarised in Table 3.11. As can be seen in this table, the Andrews and Scarratt-optimised and Treiman-optimised versions of the model do not complement each other the same way they did in the deterministic mode – the Treiman-optimised versions produce worse correlations to proportions of regular, irregular and irregular-regular difference in Andrews and Scarratt set than what was found for the vocabulary-optimised versions. Similarly, the Andrews and Scarratt-optimised versions produce poorer human-model correlations of the context sensitivity scores in the Treiman set than the vocabulary-optimised versions did. By contrast, the Pritchard-optimised versions produce mostly better performance in the Andrews and Scarratt set than the vocabulary-optimised version did. Overall, even when the model is optimised for Treiman or Pritchard set, the model's performance on these data sets is only slightly increased compared to the vocabulary-optimised versions of the model. The only data set for which data set specific optimisation seems to yield larger benefits is the Andrews and Scarratt set. Finally, the WSP-type and WSP-token versions of the model produce very similar results for all data sets, which was also the case with the vocabulary-optimised versions of the WSP model in the variable mode (see Table 3.6).

Table 3.11

Performance of versions of WSP model (variable mode) optimised for different nonword reading data sets

data set	item group	human-model correlation		RMSE		match proportion	
		optimised for Andrews and Scarratt set					
		WSP-type	*WSP-token*	*WSP-type*	*WSP-token*	*WSP-type*	*WSP-token*
Andrews & Scarratt set	regular	.85 *	.85 *	0.25	0.25	1.00	1.00
	irregular	.74 *	.75 *	0.23	0.23	1.00	1.00
	reg-irreg diff.	.84 *	.85 *	0.44	0.44	1.00	1.00
Treiman set		.55	.53	0.27	0.27	-	-
Pritchard set	1st response	.37 *	.37 *	0.29	0.29	.92	.92
	2nd response	.18 *	.18 *	0.35	0.35	.40	.40
	3rd response	.10	.09	0.36	0.36	.26	.26
		optimised for Treiman set					
		WSP-type	*WSP-token*	*WSP-type*	*WSP-token*	*WSP-type*	*WSP-token*
Andrews & Scarratt set	regular	.53 *	.55 *	0.28	0.27	1.00	1.00
	irregular	.12	.14	0.34	0.34	1.00	1.00
	reg-irreg diff.	.33	.35	0.57	0.56	1.00	1.00
Treiman set		.82 *	.78 *	0.21	0.21	-	-
Pritchard set	1st response	.31 *	.30 *	0.32	0.32	.92	.92
	2nd response	.15	.15	0.38	0.38	.40	.40
	3rd response	.21	.19	0.43	0.44	.26	.26
		optimised for Pritchard set					
		WSP-type	*WSP-token*	*WSP-type*	*WSP-token*	*WSP-type*	*WSP-token*
Andrews & Scarratt set	regular	.55 *	.56 *	0.45	0.44	1.00	1.00
	irregular	.63 *	.65 *	0.36	0.35	1.00	1.00
	reg-irreg diff.	.61 *	.62 *	0.78	0.77	1.00	1.00
Treiman set		.62	.55	0.40	0.40	-	-
Pritchard set	1st response	.42 *	.42 *	0.28	0.28	.92	.92
	2nd response	.19 *	.19 *	0.34	0.35	.40	.40
	3rd response	.13	.11	0.36	0.37	.26	.26

Note. Match proportion for Andrews and Scarratt set is 1.00 as the WSP always produced the regular and the irregular pronunciation options found in the human data. Match proportion for Treiman set is not calculated because the proportion differences of irregular pronunciations for the regular and irregular items within each item group were of interest in this analysis, not whether the human modal response or less frequent human responses were found as the WSP's pronunciation options. * denotes statistically significant correlation at alpha level of .05. RMSE = the root mean squared error of each model's proportion for a given pronunciation option, relative to those in the human data.

Due to the little or no improvements in the WSP model's performance when it was optimised for different data sets, it was also tested whether the range of weights in the optimisation procedure were too restricted. To test this possibility, the variable mode (WSP-type) was optimised with weights ranging from 1 to 50, in increments of 5 for all three data sets. The performance of the model did not improve noticeably for any of the data sets (see Appendix 2).

3.5 Discussion

In this chapter, I evaluated a new computational model of reading aloud, the WSP model, against different human nonword reading data sets and other computational models of reading. In the simulations reported in this chapter, two versions of the WSP model were considered in parallel – these were the WSP-type, where the strength of the competing pronunciation options was determined by consistency and type frequency of the PSCs, and WSP-token, where the competition was determined by consistency and frequency measures based on summed token frequency. The model's tendency to parse letter strings into larger or smaller segments is also influenced by weights for each parsing style, which can be optimised for a set of words or nonwords. The model was optimised first with existing monosyllabic words (vocabulary-optimised WSP, Section 3.2) to allow fair comparisons between the WSP, the DRC and the CDP++ models on three human nonword reading data sets. The WSP was also optimised specifically for these nonword reading data sets (Section 3.4), which allowed further inspection of the model's performance as well as gaining more information about the data sets themselves, as discussed below. Finally, the model can produce invariable responses to nonwords, aiming to simulate central tendencies in human nonword reading (deterministic mode), and it can produce variable pronunciations to nonwords, aiming to simulate naming behaviour from a group of participants (variable mode). These two modes of operation were tested against the three nonword reading data sets and the performance of the variable mode of the model was also compared to a model that also produces variable output (Zevin & Seidenberg, 2006).

3.5.1 The WSP model optimised for its vocabulary

The performance of the vocabulary-optimised WSP model in the deterministic mode compared favourably to the DRC and CDP++ models. While each of the three models outperformed the other two models on one of the three data sets, both versions of the WSP model were never the poorest performing models (only the WSP-type version performed

poorly on the Andrews and Scarratt set). However, none of the models tested in this chapter captured the pattern of responses in the three human data sets adequately. This is exemplified, for instance, as the maximum percentage of matches between a model output and the human modal responses out of the four models and across the three data sets, which was 78% (WSP-token model's performance on the Treiman set).

The usefulness of different parsing styles available for the vocabulary-optimised WSP model (deterministic mode) when reading aloud the Pritchard set nonwords was assessed by categorising the model's responses as helpful or harmful. This analysis revealed that although all parsing styles were uniquely beneficial for the model's performance, there was considerable overestimation of the antibody and word body sized parsing styles. Two important points are to be made from this finding.

Firstly, the overestimation of responses based on the antibody parsing style seemed to be more severe than the overestimation of the word body parsing style, as the former had far more harmful wins than the latter. This is likely related to the WSP's vocabulary used as an optimisation set. Inspection of the spelling patterns found in the WSP's vocabulary may help explaining why vocabulary-optimised WSP model emphasises the larger segment reading styles over the smaller segments. Firstly, out of 3921 items in the WSP's vocabulary, all three parsing styles produce the same, correct pronunciation for 76.6% of the items. The number of items that are pronounced correctly by only one of the parsing styles is 187 for the antibody parsing style, 315 for the word body parsing style and only three items for the small segment parsing style. Furthermore, there were 109 items for which both of the large segment parsing styles produce the correct pronunciation, 71 items for which both the antibody and small segment parsing styles produce the correct pronunciation and 69 items for which both the word body and the small segment parsing styles produce the correct pronunciation. As such, the maximal number of matches between the WSP's output and the correct pronunciation for the monosyllabic words in the model's vocabulary clearly requires an advantage for the larger PSCs. These characteristics still do not explain why the antibody parsing style was so advantaged compared to the word body parsing style. I do not have a satisfying explanation for this. However, it is worth noting that even though there are 315 items for which the correct pronunciation can only be obtained via the word body parsing style, this does not mean that a set of weights exists which would result the word body parsing style winning only for these items and not others, where this parsing style would lead to an incorrect pronunciation of a word.

I now turn to the second point regarding the finding that the WSP model seemed to overestimate the incidence of large segment reading style on the Pritchard set. This finding can partly be explained by the possibility that the composition of the Pritchard set encourages applying small segment reading style in skilled readers. This explanation is supported by aspects of the data set itself and the WSP model's simulations on this data set. Firstly, as noted in the description of this data set (Section 3.2.3), the majority of the items had regular word bodies and only less than a quarter of the items had irregular bodies. Thus, a large proportion of these items had only a single plausible pronunciation of the vowel. Secondly, inspection of the vowel pronunciations in the human modal responses revealed that 87% of the items received a regular vowel pronunciation as a human modal response, while only 9% were classified as irregular. Thirdly, when the WSP model (deterministic mode) was optimised for the Prichard set, over 95% of the model responses to the Pritchard set were based on the small segment reading style (for both type and token versions). Furthermore, the results of the WSP that was optimised to fit the Pritchard set showed that the model performed well on the Pritchard set items, but notably worse on other nonword data sets, where the incidence of irregular pronunciations (i.e., responses reliant on larger segment PSCs) was considerably higher amongst participants. The same, although less extreme pattern of performance was seen for the DRC model, which produces almost exclusively regular pronunciations. Thus, two models that mostly apply small segment reading style have the strongest performance on the Pritchard data set while simultaneously showing poor performance on data sets where large segment reading style is beneficial. Finally, Perry (2018) conducted an analysis of the Pritchard set and concluded that many of the items in this set are 'orthographically strange', lacking significant overlap with existing words (for instance, 60% of the nonwords had one or no orthographic neighbours), and thus potentially encouraging different reading strategies to the ones employed in reading more typical nonwords. Perry (2018) suggests that a grapheme-by-grapheme reading strategy might become more widely used when reading style based on larger orthographic segments fails. Together these findings point to the conclusion that the small unit reading style is likely overrepresented in the human naming responses to the Pritchard set.

The performance of the vocabulary-optimised WSP model in the variable mode was assessed by comparing the proportions of different pronunciation options for nonwords produced by the model to the proportion of participants producing the same pronunciations. The human-model correlations were relatively strong for some data sets (e.g., the Treiman set, and

regular responses for the Andrews and Scarratt set), but considerably weaker for others, such as the Pritchard set and irregular responses in the Andrews and Scarratt set. Several issues in the model's performance were identified. Most importantly, the model's proportions for regular and irregular body responses to nonwords did not always converge with those from human participants – the model overestimated the incidence of regular-body pronunciations in the Andrews and Scarratt set but overestimated the incidence of irregular-body responses in the Pritchard set. As shown with the data set specific optimisation of the model, some of these issues are not resolved with adjusting the pronunciation option weights - which is seen particularly clearly in the performance on the Pritchard set. The suggestion that the compositions of the Pritchard set may encourage small segment reading style argued above also bears relevance to the question of whether the proportions of participants producing a certain naming response in the Pritchard set are representative enough to demand modifications to the WSP model, so that the model would better reflect the patterns of responses found in this data set.

While some of the shortcomings of the WSP model are worth focusing on (see below), some of the differences between the human and model proportions may need to be considered more carefully – namely, the pattern of naming responses for the Pritchard set can be seen as an example of list context effects in nonword reading, which have also been reported in previous studies (e.g., Brown & Deavers, 1999, Exp. 4; Rastle & Coltheart, 1999, Exp. 2). It is thus difficult to determine what exactly would count as 'standard reading behaviour' and what would be considered influence from the context in which nonwords are presented. If computational models aim to capture the general principles by which skilled readers name letter strings, list context effects may be considered as noise. However, if skilled readers are influenced by the context in which they name letter strings, an accurate account of skilled reading should also include this feature (for instance, by advantaging certain types of pronunciations for items based on recently presented items to simulate priming effects). On the other hand, list context effects, as well as other influences such as priming effects in nonword reading are likely a result of complex interactions of several factors, including semantic involvement in print-to-sound conversion (e.g., Rosson, 1983). As such, it may be wise to aim to model reading behaviour without these influences first.

Yet, this simplified goal brings us back to the question: what is considered 'standard reading behaviour'? If some nonword reading data sets encourage certain kind of reading behaviour, what kind of nonword data set would be as free from these influences as possible?

Considering that most nonword data sets to date were constructed for a specific purpose – i.e., manipulating certain properties of the nonwords in order to discover their influence on nonword reading (e.g., Andrews & Scarratt, 1998; Treiman et al., 2003) or selecting items based on maximal contrast between computational models (Pritchard et al., 2012), it may be beneficial to construct a nonword naming data set that is a representative sample of the PSCs found in English. While this approach still does not remove all influences of context (e.g., nonwords named in isolation vs. intermixed with words), it may still be informative in search for the 'standard reading behaviour' current computational models aim to simulate.

Comparison of the WSP model's variable mode to the model by Zevin and Seidenberg (2006) showed that, overall, both models capture the general pattern of human naming responses in terms of the proportion of regular pronunciations and pronunciation variability for different nonword groups in the test set (Andrews & Scarrat, 1998, Exp. 2). Focusing on the individual differences in nonword reading, Zevin and Seidenberg (2006) suggest that these differences could arise from different exposure to words and PSCs within them. As such, they suggest that models which include a mechanism for learning PSCs is needed, and multiple runs of these types of models can represent nonword reading responses from participants with varying experience with reading. While there is no such learning mechanism in the WSP model, and the underlying probabilities for each pronunciation option remain the same across simulation runs (or 'participants'), the WSP model can produce reading responses that are variable, reflecting the choice of pronunciations skilled readers make for ambiguous stimuli. Thus, the WSP model's variable mode can be seen as a reflection of the variability found within subjects, on a trial-by-trial basis, even when the model does not capture individual differences in longer-term tendencies skilled readers might have when reading aloud nonwords (e.g., due to their personal PSCs, or reading style adopted as a result of reading instruction in school). By contrast, Zevin and Seidenberg's model simulates the variability in global tendencies of reading aloud, i.e., between subjects, but not the within-subjects variability, as the same item would always be pronounced the same way by the same version of the model that represents a single participant.

As both models showed a relative strength in one of the measures investigated, Zevin and Seidenberg's model in the proportion of regular responses and the WSP in the response variability, it seems that a model simulating either between-subjects or within-subjects variability in nonword reading is sufficient for capturing the general pattern in the human data. Yet, this level of analysis is not sufficient for differentiating between the two types of

variability; as pointed out by Ulicheva et al. (2021), nonword naming studies based on single testing sessions do not allow differentiating between within and between-subjects variability. I therefore suggest that more detailed data sets with naming responses to all the items from all the participants (such as the one from Pritchard et al., 2012) are necessary for further investigations on variability in human nonword reading. With such data sets, the relative success of Zevin and Seidenberg's model and the WSP model could be teased apart, as similarity to patterns of both between-subjects and within-subjects variability could be tested. This would be the case especially with data sets that contain repeating sub-lexical patterns or that are collected during multiple testing sessions with repeating items across sessions.

An idea of how to evaluate the performance of a model that produces variable output with a more detailed data set was presented in Section 3.3.3. Comparison of individual participant's performance or an individual WSP model's simulation run against the group tendencies in the Pritchard set showed that an average simulation run performed comparably to an average human participant. However, the WSP model produced more uncommon pronunciations than an average human participant did.

3.5.2 The WSP model optimised for nonword data sets

When the WSP model was optimised for specific nonword data sets, the performance of the deterministic mode of the model was generally higher than that of the vocabulary-optimised model. This was particularly true when the model was optimised for the Andrews and Scarratt or Treiman sets, which produced considerably high performance across data sets. As mentioned above, optimisation of the WSP model for the Pritchard set produced a less balanced performance, where high incidence of small segment reading style hindered the model's fidelity to human-like naming responses in two of the data sets.

The most important finding about the data set specific optimisation in the variable mode was that the model's performance was not increased noticeably in the Pritchard and Treiman sets, when the model was optimised specifically for these data sets. Only the performance on the Andrews and Scarratt set benefitted considerably from the model optimisation for this data set. This finding suggests that something else than the relative strength of the weights in the WSP model prevents it from reaching higher convergence with the pattern of proportions for different pronunciation options found in the human data.

3.5.3 Limitations and future directions

The finding that skilled readers do not produce context sensitive or irregular pronunciations as often as might be expected based on the PSCs found in existing words (Treiman et al., 2003; Treiman, et al., 2007) is problematic for the WSP model. In the deterministic mode, the context sensitive responses tended to be too frequent compared to human responses (e.g., all the items with an onset *g* followed by *e* received a soft pronunciation by the model) and in the variable mode, the proportions of context sensitive and insensitive responses did not always correspond to those found in the human data. While more human-like proportions of context sensitive responses could be achieved by applying a disadvantage for context sensitive pronunciation options in the model's variable mode, this would only be an ad hoc modification that would not inform us about the mechanisms by which this lower context sensitivity happens in human nonword reading. Thus, a more general solution is needed.

A related issue for the WSP model was briefly discussed in Section 3.3.1, namely, that the model's assembly of pronunciation options in the variable mode sometimes produces responses that are rarely or never produced by humans. This is partly because onsets and codas in the WSP's PSC knowledge and the procedure for naming a letter string are treated as single units regardless of the number of graphemes in them. A potential way forward would thus be to modify the representation of the onset and coda segments so that they correspond to graphemes and to increase the threshold for inclusion of less frequent PSCs in the assembly of pronunciation options. However, it remains to be seen whether this modification would result in reduced performance on other aspects of nonword reading that the model currently simulates well. For instance, the issue of context sensitive onsets involving a soft *g* pronunciation would still remain without a solution: pronunciation associated with a *g* followed by *e* is currently represented in the WSP model as a *ge* segment for all parsing styles with an onset segment (i.e., word body and small segment parsing styles) and in the antibody parsing style whenever the full vowel segment is found in the WSP's vocabulary. Thus, most items with a *ge*-segment would receive a context sensitive pronunciation by all three parsing styles. In the variable mode of the WSP, the same way of representing the *ge*-segments would still result in considerably higher proportion of context sensitive pronunciations than context insensitive pronunciations. Removing this way of representing the context sensitive pronunciations for *g* would require an alternative solution so that these types of pronunciations could be produced at all by the model.

Similarly, the issue of pronunciations assigned to the *th*-onsets by the WSP's variable mode is not solved with higher threshold for inclusion of PSCs for the assembly of pronunciation options. This is because the pronunciation most consistently associated with the antibody segment *the* is /DE/, with a consistency value of .75. Thus, this pronunciation, which is rarely produced by human participants, would have to be included as a pronunciation option. One way to avoid these problematic *th*-onset pronunciations would be to include part of speech information in the PSC knowledge of the WSP model, for instance, such that only the PSCs occurring in content words would be included.

As noted above regarding the nonword data set specific optimisation of the WSP's variable mode, adjusting weights for the different pronunciation options does not increase the model's performance considerably for all data sets. As such, the source of the model's imperfect correspondence to human proportions for different pronunciation options needs to be found elsewhere. One such source may be the vocabulary which the WSP's PSC knowledge is based on. Unlike the model, skilled readers naturally base their PSC knowledge on more than just monosyllabic words. Thus, inclusion of disyllabic and multisyllabic words in the WSP's vocabulary may produce more human-like PSCs. On the other hand, the current vocabulary of the WSP model, on which the PSCs available to the model are based on, contains words that are likely not known by an average reader. More realistic vocabulary for the model could be achieved by using a frequency threshold for inclusion of items, for instance, only items with a Zipf frequency of 4. However, as the frequency of the word does not always predict whether the word is generally known by skilled readers (Brysbaert et al., 2019) a word prevalence measure, the proportion of participants indicating they know a given word (Brysbaert et al., 2019), may be a better criterion for inclusion of words in the WSP model's vocabulary.

Additionally, as pointed out in the comparisons to the Zevin and Seidenberg's model, the fact that the same PSC knowledge is available to each simulation run ('participant') of the WSP model is unrealistic. For the WSP model's variable mode to simulate personal PSCs for each participant, the model's PSC knowledge should be based on slightly different sets of words. Additionally, or alternatively, the weights for each parsing style should vary between simulation runs ('participants'), to allow simulation of the global tendency to parse letter strings into larger or smaller segments that skilled readers appear to differ in.

Inspection of the WSP model's performance as it was optimised for different nonword data sets was informative, particularly in terms of the composition of these data sets. However, the issue of what kind of psychologically plausible data set should be used for optimising the WSP model is still open. This data set should be representative of skilled readers' experience with reading. However, exposure to the WSP's vocabulary reported in this chapter (i.e., the vocabulary-optimised model) demonstrated that using this optimisation set may result in the model over-emphasizing certain aspects in reading, such as the antibody parsing style, which is not the most widely used parsing style amongst skilled readers (e.g., Andrews & Scarratt, 1998, Exp. 1; Kessler & Treiman, 2001; Treiman et al. 1995).

By contrast, optimising the WSP model for the Treiman set appeared to result in a balanced, well performing model across data sets. There are two potential reasons for why the Treiman data set produced such an optimal balance between the three parsing styles in the WSP model. Firstly, this data set consisted of an even number of regular and irregular items, where the irregular items tended to elicit a high proportion of context sensitive pronunciations in skilled readers. Thus, only combinations of weights that produce both irregular and regular pronunciations will result in a strong performance on this data set. For instance, the small segment parsing style needs to override the antibody parsing style for items with regular word bodies (as the antibody parsing style is the only one that might produce a different pronunciation than the other two parsing styles for these items). Secondly, for the purposes of optimising the WSP model, the Treiman et al. data set may have an appropriate proportion of items requiring an antibody analogy (two item groups) compared to items requiring a word body analogy (six item groups). Thus, for the model to produce a maximal number of human modal responses in this data set, the antibody parsing style needs to be strong enough to win for a relatively small number of items, while the word body parsing style needs to win for a substantial proportion of the items. This type of balance between the two large segment parsing styles should be particularly beneficial for the model's performance, because the use of antibody analogies by skilled readers appears to be restricted to a few cases, the *qua* and *wa* antibody segments (e.g., Treiman et al., 2003), while word body analogies are used in reading aloud a wider range of nonword spelling patterns (e.g., Andrews & Scarratt, 1998; Brown & Deavers, 1999; Treiman et al., 2003). To investigate the performance of the WSP model optimised for Treiman set further, the versions of the WSP model optimised for this data set will be included in model comparisons in Chapters 4 and 5, where computational models are compared to new nonword reading data sets.

Finally, the two versions of the WSP model, the WSP-type and the WSP-token, do not seem to differ noticeably in their capacity to capture human data – though some data sets under some optimisation conditions were better simulated by the token version, others were better fit by the type version. As such, these investigations have not provided a definite answer to whether type or token frequency is better suited for capturing the kind of PSC knowledge skilled readers might utilise when reading aloud. Thus, as described below, this issue will be approached empirically in the following chapters.

3.5.4 Conclusion

To conclude, the WSP model provides a flexible approach to investigating print-to-sound conversion in skilled readers. The model can also be used for gaining further insight to characteristics of different nonword reading data sets. Compared to the DRC and CDP++ models, the deterministic mode of the WSP model has an overall, strong performance in simulating nonword reading behaviour of skilled readers, across data sets. However, none of the models compared in this chapter fully captured the pattern of naming responses in the human data. The performance of the variable mode of the WSP model was somewhat promising, but clearly below what could be considered as successful simulation of variability in nonword reading. Yet, when different ways of evaluating the variable output of the model were considered, it was found that the individual simulation runs of the model perform almost as well as individual participants do. Several ways to improve the model's variable mode were discussed, such as improvements to the PSC knowledge of the model (e.g., basing the PSC knowledge on content words only) or modifications on how onset and coda segments are represented in the model (e.g., graphemes rather than consonant clusters). The investigations conducted in this chapter also highlight the potential issues one ought to consider when evaluating a model's performance against nonword naming data sets, namely, other influences that may shape the nonword reading responses besides the PSC knowledge of skilled readers. While the performance of the WSP model on the available data sets was informative, it is important to see how well the performance seen so far would generalise to different types of data and, indeed, whether some of the shortcomings identified with the available data sets would also be as problematic in other data sets. As such, Chapters 4 and 5 present new human nonword reading data and comparisons to WSP and other computational models are made. Furthermore, as the WSP's type and token frequency versions performed very similarly in the three data sets used thus far, the experiments reported in the following

chapters were designed to investigate the difference between the role of type and token frequency in nonword reading empirically.

Chapter 4 : Token frequency in nonword processing

4.1 Introduction

When skilled readers assign pronunciations to nonwords, they are likely to employ knowledge of several statistical properties of the writing system, such as consistency and frequency of print-to-sound correspondences (PSCs; e.g., Andrews & Scarratt, 1998; Seidenberg et al., 1994, see also Section 1.1 in Chapter 1). However, agreement has not been reached about which statistical properties skilled readers are sensitive to when reading aloud nonwords. One of the remaining questions is whether nonword reading is influenced more by type frequency, that is, the number of words embodying a given PSC than by token frequency, that is, the frequency of the words embodying a given PSC. For instance, are skilled readers more likely to name a nonword *strave* to rhyme with words like *brave, cave* and *pave* (using the PSC *ave* → /1v/) because this PSC occurs in several words? Or would skilled readers pronounce this nonword to rhyme with *have* (using PSC *ave* → /{v/) as this PSC is encountered often, although in only a single word. Only a few studies provide direct empirical evidence on the topic and due to limitations of this previous research, more conclusive evidence is still needed.

In this chapter, I report results from an experiment designed to investigate the role of token frequency in nonword processing, specifically regarding pronunciations assigned to nonwords. Apart from the relative importance of type and token frequency of PSCs, it is also valuable to discover whether token frequency has *any* influence in nonword processing and if so, how large this effect is. Implications of these findings bear relevance to theory development of reading aloud. Additionally, due to the type of stimuli used in the experiment, a secondary goal of the study was to investigate the relationship of consistency and type frequency in nonword reading. The secondary goal of the study can be seen as an edge case which might be particularly problematic for computational models of reading with strong reliance on consistency of PSCs. The empirical findings from the study are compared with the output from some current computational models of reading. Empirical evidence from investigations of the relative importance of type and token frequency is outlined first, followed by an overview of the role of type and token frequency in computational models of reading, after which the current study is described.

4.1.1 Token frequency in human nonword reading

The question of whether type or token frequency of PSCs better characterises human nonword reading has been addressed in only a handful of studies (Andrews & Scarratt, 1998; Johnson, 1970; Kay, cited in Kay & Marcel, 1981; Norris, 1994; Treiman, Goswami and Bruck, 1990). While some evidence for the role of token frequency is reported (Andrews & Scarratt, 1998), the findings from these studies mostly suggest that type frequency of PSCs is more influential in nonword reading. However, characteristics of these studies do not always allow strong conclusions to be drawn. Some of the issues in these studies are outlined below.

Johnson (1970) investigated pronunciations assigned to vowel clusters in nonwords by developing readers (second, fourth and sixth graders). The participants chose a pronunciation from four options (four existing words in which the critical vowel cluster was pronounced in different ways). One of the aims of the study was to determine whether the children's pronunciations of the vowel clusters would be associated more closely to type or token frequency-based measures of most common PSCs. Johnson extracted the type and token frequencies from different corpora – the type-based measures came from a 20,000-word corpus, originally compiled by Thorndike as the most frequent English words and updated by Venezky (1963). The token corpus consisted of 1000 most frequent English words, as they occur in written American English (by Kucera and Francis, 1967). The proportions of different pronunciations for each vowel cluster were thus calculated either as the number of words in which a given pronunciation occurred (type-measure) or as the number of words with the given pronunciation, multiplied by their token frequency (token-measure). For example, a PSC *ou* → /U/ proportion was 0.01 in type-based measures, but 0.26 in token-based measures (due to highly frequent items *could*, *should* and *would*). Pronunciation preferences to nine vowel clusters were investigated such that 10 nonwords for each cluster were created. Via descriptive comparisons, Johnson concluded that type frequency predicts human nonword pronunciations better than token frequency. For instance, the most common human pronunciation matched the most common pronunciation for eight out of nine vowel clusters based on type frequency and seven out of nine vowel clusters based on token frequency. Some issues to point out from this study are that the type and token frequency measures were derived from two different corpora, which may lead to inconsistencies between the two measures. Additionally, the participants' reading ability varied considerably. In principle, sensitivity to either type or token frequency of PSCs may change as reading

skills develop. Most importantly, the pronunciations were investigated on a grapheme level, thus ignoring contextual effects to the vowel pronunciations[26]. While this level of analysis is not necessarily a problem for comparing type and token frequency-based PSCs, it demands more careful selection of stimuli. This is because the most common pronunciations for graphemes based on type counts, by definition, reflect the PSCs that occur in the largest number of words. By contrast, the non-standard, or irregular pronunciations for vowel clusters tend to be exemplified by highly frequent words. For example, the most common pronunciation for vowel clusters *ea* or *oo* depend on the coda, such that *ea* followed by *d* and *oo* followed by *k* are mostly pronounced irregularly in existing words (e.g., *head* and *book*), while these vowel clusters combined with most other codas are associated with a regular pronunciation (e.g., *heal* and *boost*). If the test items mostly included orthographic segments without the codas that would elicit an irregular pronunciation, the test set would be biased towards type-based PSCs. For some vowel clusters, this was indeed the case, such as *oo,* for which only two out of the ten nonwords used would encourage irregular pronunciations as they had a coda *k.* As such, while Johnson's study does provide some indication of the importance of type frequency over token frequency, this finding needs to be confirmed by investigations with stricter control of the stimuli, bearing in mind other properties of the PSCs, such as the vowel context, demonstrated by later research (e.g., Brown & Deavers, 1999; Treiman et al., 2003).

Another study investigating the role of type and token frequency was conducted by Andrews and Scarratt (1998, see Chapter 1, Section 1.1.3). Their two experiments showed that when skilled readers name regular-consistent, inconsistent and irregular-consistent nonwords, the likelihood of pronouncing a nonword regularly is best predicted by a type-based measure of the proportion of regular body neighbours. Importantly, token-based measures also served as significant predictors, but the overall fit of the regression models based on type-metrics was higher (R^2 = .69) than those based on token-metrics (R^2 = .62). It is worth noting that the results reported by Andrews and Scarratt were based on summed token frequencies, rather than, for instance, maximum token frequencies. However, the authors note that they also considered maximum token frequency, which was not a significant predictor for the dependent variables (Andrews & Scarratt, 1998, footnote 11, p. 1071). This is important, because summed token frequency is correlated with type frequency. This is because summed

[26] While some qualitative observations about contextual effects of the vowel pronunciations are provided in this study, these are not considered in terms of the main analysis.

token frequency also includes information about the number of items included in the final summed value, not only the token frequency of these items. As a demonstration of this relationship, the Weighted Segments Pronunciation (WSP) model's PSC knowledge is a collection of PSCs of varied sizes for which consistency and different measures of frequency were then calculated. Calculated across all the PSCs in this data base, type frequency and summed token frequency had a strong positive correlation ($r(3211) = .93$, $p < .001$). By contrast, maximum token frequency correlated with type frequency to a far lesser extent ($r(3211) = .3$, $p < .001$). Bearing this in mind, it is possible that token-based measures in Andrews and Scarratt's experiments were significant predictors because they also contain type frequency information, not because token frequency per se predicts nonword reading behaviour. Thus, quantifying token frequency in a different way, for instance, as maximum token frequency, may be more informative in investigations aimed at teasing apart the influence of type and token frequency.

Additionally, a study by Treiman et al. (1990) bears some relevance to the question of token versus type frequency, although this study was designed to answer different questions. In three experiments, adults and 1st and 3rd graders named 48 regular and consistent nonwords which were classified as either high or low. The word body segments in high nonwords were more prevalent in the language based on three measures: 1) type frequency, that is, the number of monosyllabic words in which the word body segment was pronounced the same way), 2) the number of these words also occurring in a list of words in reading materials for children and 3) the summed token frequency of these words. The high and low words contained the same graphemes (e.g., high: *tain*, *goach* and low: *taich*, *goan*) and the antibody segments in the nonwords were similar in frequency. Thus, differences in pronunciations assigned to high and low items can be attributed to the word body segments, rather than graphemes or antibody segments. Correct pronunciations for nonwords were those following GPC rules (as defined by Venezky, 1970). Both children and adults made more errors in naming the low items compared to the high items and lexicalisation errors were made more often for high items than for low items. Separate regression analyses on error proportions from the adult data with each of the three measures of prevalence showed that the first two – type frequency of word body segments and the number of words containing the word body segment found in children's reading materials – were significant predictors (explaining .14 and .13 of the variance in naming errors, respectively), whereas summed frequency of words with the word body segment just missed statistical significance and explained slightly less

variance in naming errors (.11). In terms of conclusions about the role of token frequency in nonword reading, the same issue mentioned above for Andrews and Scarratt's study, namely, that using summed token frequency makes the distinction between type and token frequency less clear, also applies to the study by Treiman et al. (1990).

As an extension to Treiman et al. (1990) study, Bowey and Hansen (1994) also investigated the rime frequency effect, that is, more accurate naming of nonwords for high than low items, in developing readers. However, the high items in this study also had higher type and token frequency, and thus cannot answer the question about the relative importance of type and token frequency in print-to-sound conversion.

Additionally, in a set of experiments by Jared et al. (1990) the role of type and token frequency was investigated in word naming, with carefully controlled stimuli characteristics. Most importantly, the consistency effect in word naming, that is, faster and more accurate naming of consistent words (words with no enemies) compared to inconsistent words (words with friends and enemies), was found to depend on the relative frequencies of the friends and enemies rather than the number of friends and enemies. For instance, in Experiment 2, four groups of inconsistent words with crossed factors of frequency of friends (low or high) and frequency of enemies (low or high) were matched with four groups of consistent words in terms of important factors such as mean frequency of friends, length and frequency. The consistency effect (naming latencies and accuracies of the inconsistent items versus the matched consistent items) was found to vary based on the relative frequencies of friends and enemies, such that the largest effect was found for inconsistent items with low frequency friends and high frequency enemies (significant in both by-items and by-subjects analyses) and the smallest effect for inconsistent items with high frequency friends and low frequency enemies (ns. both in by-items and by-subjects analyses).

Furthermore, in Experiment 3, naming latencies to two types of inconsistent words were compared: words with higher number of friends than enemies (more-friends items) and words with higher number of enemies than friends (more-enemies items). Importantly, the two groups of words were matched in terms of their mean summed frequency of friends and enemies (and the mean summed frequency of enemies was larger than the mean summed frequency of friends in both groups). The frequencies of the words in each group were also matched. With these characteristics of the stimuli, if type frequency plays a role in

consistency effects, the more-friends group of words should yield a smaller consistency effect than the more-enemies group of words. If token-frequency is driving the consistency effect, then no differences should be found between the consistency effects produced by these two groups of items. 24 participants gave speeded naming responses to these items. The results for the latency data showed an overall consistency effect – inconsistent words were named more slowly than consistent words. No effect of the inconsistent word type (more-friends or more-enemies) was found, although the more-friends items showed a numerically larger consistency effect (42 ms) than the more-enemies items (34 ms). The error data showed a significant effect of inconsistent word type, such that more errors were made for more-enemies items (7.1%) than for more-friends items (3.6%). Jared et al. conclude that the latency and error data showing opposite numerical trends is suggestive of no reliable effect of type frequency in consistency effects. While these results provide evidence for the importance of token frequency in word reading, more direct evidence regarding the type of naming responses assigned to nonwords is still needed.

Finally, Norris (1994) compared the performance of his multiple-levels model (see Chapter 1, Section 1.2.4) against Glushko's (1979) nonwords when the frequency of the model's PSC rules were either based on type or token counts. Compared to the incidence of 'correct' responses (defined by Glushko) to inconsistent nonwords in the human data (78%), the type-based rules produced a more human-like performance (74%) than the token-based rules (56%). However, it is not clear whether the same nonwords were pronounced correctly by humans and by either type or token-based version of the model.

Overall, the available empirical evidence for the relative importance of type and token frequency and particularly compelling evidence for the role of token frequency in nonword reading is scarce. Some of the limitations pointed out in the abovementioned studies suggest that more research is needed.

4.1.2 Frequency measures in computational models

In addition to the models considered in Chapter 3, namely, the dual-route cascaded (DRC) model, the dual process connectionist (CDP++) model and the WSP model, another model of reading, a PDP model by Plaut et al. (1996, Simulation 1) is also included in the model comparisons in this chapter. This model is included due to the nature of the stimuli used

(described below in Section 4.1.3), which may prove particularly problematic for models with strong emphasis on consistency of PSCs. Therefore, a larger variety of these types of models may be informative[27].

The DRC model (see description of the model in Chapter 1, Section 1.2.1) assigns pronunciations to nonwords based on grapheme-phoneme correspondence rules, that is, grapheme sized PSCs with the highest type frequency. Thus, the model's reading behaviour reflects type frequency in a categorical manner, such that the PSCs with the highest type frequency will always be employed. Token frequency, on the other hand, is not influential in the DRC model's nonword reading.

The CDP++ model (see description in Chapter 1, Section 1.2.2) names nonwords based on PSCs of varying sizes, which are learnt via training on existing words. As the learning rate of the model during training is weighted by a token frequency value of each word, the correspondences in more frequent words should have a greater influence on the model's reading behaviour than those in less frequent words. However, correspondences in larger number of items would also influence the final reading performance more than correspondences occurring in fewer words. As such, the CDP++ model should show sensitivity to both type and token frequency of PSCs.

The PDP model by Plaut et al. (1996, Simulation 1, see description in Chapter 1, Section 1.2.3), which I will refer to as Psim1, also names nonwords based on different sized PSCs, learnt from exposure to existing words and their pronunciations. This model is also sensitive to both type and token frequencies of PSCs.

The WSP model (Chapter 2) also produces nonword pronunciations based on PSCs of varying sizes. The statistical properties that determine which pronunciation the model assigns to a letter string can be chosen by the user. Two versions of the WSP evaluated in Chapter 3 were WSP-type, where type frequency of the PSCs influences the final pronunciation the model gives, and WSP-token, where token frequency is influential instead. Note, however, that these versions of the WSP do not offer the clearest distinction between type and token frequencies, because the token measures in the WSP-token are based on summed token frequencies, which, as pointed out above (Section 4.1.1), also contain information about type frequencies. Three versions of the WSP model were considered in the current study: the

[27] I thank David Plaut for providing me with the simulation output from his model for the stimuli used in this chapter and Chapter 5.

vocabulary-optimised type and token versions of the model and the WSP-token version optimised for a nonword data set by Treiman et al. (2003), referred to as WSP-token-T, as this version of the model had a particularly strong performance across all three data sets considered in Chapter 3 and as it is also sensitive to token frequency of PSCs.

4.1.3 The current study

In Chapter 3, I compared the type and token versions of the WSP model against human nonword reading data sets and concluded that the two versions of the model perform similarly. The differences found were not systematic – sometimes the WSP-type produced a closer match to human reading behaviour and sometimes the output from the WSP-token was more similar to the human naming responses. Thus, these comparisons did not inform us about the relative importance of type and token frequency in nonword reading. However, the available data sets do not allow direct comparison of the two properties. In the current study, I took the following approach to investigating the role of token frequency in nonword reading. The experimental stimuli consisted of nonwords that were based on words (base words) with unique and irregular word bodies. This allowed detecting the incidence of pronunciations that were congruent with the pronunciation of the base word – so as to infer when the particular PSC embodied by the unique base word influenced nonword naming in skilled readers. As the nonwords in this experiment had only one word body neighbour (the unique base word), the type frequency and consistency of the experimental items was kept constant, while the token frequency of the items was manipulated by choosing base words with either high or low token frequency. Thus, using this nonword set, I asked whether nonwords with high-frequency word bodies elicited more base word congruent responses (i.e., irregular vowel pronunciations) than nonwords with low-frequency word bodies. If so, this would suggest that token frequency influences nonword naming in skilled readers. The same question was also investigated with nonwords with regular base words, but the conclusions from these items may not be as clear due to two reasons: 1) the base word congruent pronunciations could be arrived at by using grapheme-phoneme sized PSCs rather than word body sized PSCs (or word body analogies), 2) the incidence of regular pronunciations for the regular items was likely at ceiling regardless of the token frequency of the base words. For these reasons, the regular items were not included in the rating task described below (Section 4.2.3.2).

Additionally, this set of irregular nonwords lends well to another question. Namely, how are nonwords with high consistency but low type frequency named? This is because the consistency of the word bodies in these nonwords were perfect (with proportion 1 for each word body sized PSC), but the type frequency of these items is low, at 1 (the word body sized PSC only occurs in one existing word). As such, the incidence of irregular pronunciations assigned to these singleton nonwords can reveal more about the relative influence of consistency and type frequency of PSCs in nonword naming.

While Andrews and Scarratt (1998, Experiment 2) showed that even unique irregular word bodies can elicit irregular pronunciations, a closer inspection of their stimuli raises a question about how accurate the reported incidence of irregular pronunciations is in this case. My inspection of these items revealed that the word bodies in the irregular-unique set were not all unique: 41.7% (5 out of 12) of them occurred in other monosyllabic or disyllabic words and 25% occurred in other monosyllabic words. Some of these nonwords had more than one irregularly pronounced neighbour (e.g., word body *ign* also occurs in *benign* and *design*, in addition to *sign*), some had regularly pronounced neighbours or both. This might explain some of the patterns in irregular pronunciations to these items. For instance, the word body *inth* is pronounced irregularly in one word (*ninth*)[28] and regularly in four other words it occurs in (*plinth, absinth, hyacinth, labyrinth*), most of which should be known by university students (the participants in the study). As a probable reflection of this, the proportion of regular pronunciations assigned to nonwords with this word body was the highest in the irregular-unique category (97.9% and 100%). Thus, the accuracy of the incidence of irregular pronunciations assigned to irregular singleton items in Andrews and Scarratt's second experiment may be compromised. The choice of the stimuli was presumably due to an assumption that the student participants would only know one of the words with a given word body. However, some of the additional neighbours are relatively frequent and thus likely to be known.

An additional complication for an accurate estimate of irregular pronunciations assigned to nonwords comes from the fact that participants differ in the number of words they know. Because irregular pronunciation of nonwords crucially depends on whether the words in which these irregular word bodies occur (i.e., base words) are known, a formal assessment of vocabulary knowledge of the base words would be beneficial. This way individual lexical

[28] The word body *inth* in *ninth* also has a morphological structure, derived from *nine*, which might make the influence of the irregular pronunciation of *inth* less likely.

knowledge can be taken into account when calculating the incidence of irregular nonword pronunciations. Therefore, the current study aimed to expand on Andrews and Scarratt's findings by using strictly unique irregular items (considering both mono- and multisyllabic words) and by taking the readers' knowledge of the unique base words into account. The latter point is particularly important given that some of the base words in the current study have a considerably low token frequency.

These two questions, the role of token frequency in nonword reading and the influence of consistency relative to type frequency in nonword processing were investigated by collecting two types of responses to the experimental nonword items – naming responses and rating responses. The rating responses were acceptability ratings the participants gave to the experimental items when these items were paired with regular or irregular pronunciations. In order to ensure that the preceding nonword naming task did not influence the rating behaviour in the rating task, two groups of participants were included – one group completed the naming task, followed by the rating task (Naming-Rating group) and a second group only completed the rating task (Rating-Only group).

Both questions were also addressed by comparing computational models of reading against the human responses to the experimental items. Firstly, I asked whether models that are sensitive to token frequency of PSCs (the CDP++, WSP-token, WSP-token-T and Psim1) simulated the human responses in this experiment better than models that are not (the DRC). Secondly, I asked how closely the overall incidence of irregular responses to the experimental items matched the output from the models. This question was particularly important for models where consistency plays a central role, such as the Psim1 model and the WSP model.

The hypotheses of the current study were as follows.

The role of token frequency in nonword processing.

1. The incidence of base word congruent pronunciations assigned to nonwords with high token frequency will be higher than that to items with lower token frequency. Irregularly pronounced nonwords with high token frequency will also receive higher acceptability ratings than irregularly pronounced nonwords with lower token frequency.

2. Computational models that are sensitive to token frequency will be a closer match to the human data than models that are not.

The influence of consistency and type frequency in nonword reading. In these exploratory investigations, the incidence of irregular pronunciations to the experimental items was expected to be higher than that found in Andrews and Scarratt's study (1998, Exp. 2). Additionally, comparisons of human responses and model output were made with focus on the role of consistency and type frequency of these items, regardless of their token frequency.

4.2 Methods

4.2.1 Participants

Participants were undergraduates in Psychological Science in the University of Bristol. They completed the study online as a course requirement. Inclusion criteria for the study were that participants be native speakers of British English with normal or corrected-to-normal vision and no diagnosed or experienced reading difficulties. Participants were randomly assigned to one of two versions of the study. The Naming-Rating version consisted of three tasks: nonword naming, nonword rating and vocabulary task; the Rating-Only version consisted of only the latter two tasks. A total of 144 participants completed the experiment. Three participants were excluded due to failed audio recordings in the vocabulary task (i.e., no recordings at all or under 20% usable responses). By comparison, the average percentage of failed recordings in the vocabulary task for the retained participants was 4.5%, with maximum at 35.7% of trials lost for a single participant. Additionally, three participants were excluded due to their age. This was not an exclusion criterion originally, but as all participants that were 40 years old or older fell into the same group (Rating-Only) and as reading experience and vocabulary size correlate positively with age (e.g., Brysbaert, et al., 2016 for vocabulary size), it was deemed necessary to exclude these participants to ensure the two groups were as closely matched as possible. After excluding participants due to eligibility or other issues described above, the final sample size was 138, with 69 participants in each group (see Table 4.1).

Ethics approval for the present study was granted by the School of Psychological Science Research Ethics Committee in University of Bristol (ethics approval code: 170220100902).

Table 4.1

Demographics of the Naming-Rating and Rating-Only groups

Group	Mean Age (SD)	Age range	Sample size	Females	Males
Naming-Rating	20 (2.1)	18-29	69	54	15
Rating-Only	19.8 (2.3)	18-30	69	60	9

4.2.2 Materials

4.2.2.1 Naming task. The experimental items in the naming task consisted of monosyllabic nonwords that each shared a word body with a single existing English word (i.e., base word). The nonwords were created by adding a single or two-letter onset to the word body of the base words. Two different onsets were combined with each word body in order to create two nonwords (e.g., *hauge* and *snauge*) from each base word (e.g., *gauge*). These onsets did not contain any of the letters in the onset of the base word and the number of two-letter onsets was 16 or 17 in each of the four nonword categories described below. The resulting nonwords were not homophones to any existing words when they were pronounced either regularly (i.e., according to the most frequent pronunciation associated with each onset, vowel and coda sized PSCs) or irregularly, i.e., as a word body analogy (e.g., d*wonge* pronounced either as d*wQn_* or d*wVn_*). Each nonword had only one orthographic neighbour: the base word of the nonword. The number of homophones and the number of orthographic neighbours[29] were assessed against the WebCelex database (retrieved from http://celex.mpi.nl/). These nonwords were categorised as regular or irregular and as high or low (token) frequency, according to the characteristics of their base words. This resulted in four categories: Irregular-low, Irregular-high, Regular-low and Regular-high, with 28 items in each category. A Zipf scale was used as a measure of token frequency of the base words (Van Heuven et al., 2014). The values on this logarithmic scale vary mostly from 1 to 6 with the boundary for low and high frequency words between 3 and 4. Reasonably well in line with this, the medians for regular and irregular items were 3.15 and 3.61, respectively. The mean frequency of the base words for Regular-low and Irregular-low items were 2.68 (*SD* = 0.36) and 2.69 (*SD* = 0.59), respectively. For the high frequency items, the mean frequency was 3.74 (*SD* = 0.44) for Regular items and 4.58 (*SD* = 0.71) for Irregular items. The uniqueness of the nonwords was assessed against all items in the WebCelex database, excluding inflected forms of words or the same letter string in a different position. Thus, *sponge* was unique, even though *prolonged* or *congest* contain the letter string *onge*. Although efforts were made to include only base words with unquestionably unique word bodies, some exceptions remained. These were all compound words that shared a stem with the base word (e.g., *kilowatt* and *unleash* share a stem with *watt* and *leash,* respectively). However, these "questionably unique" items were a clear minority, and the number of these items was spread across the item categories in a reasonably even fashion (see Appendix 3, Table 3A).

[29] Defined as any letter string that differed from the nonword by one letter

Another set of experimental items was a subset of nonwords from Treiman et al. (2007). These items allowed investigation of context sensitive pronunciations assigned to letter strings beginning with C or G, used in evaluating the rating method (see Chapter 6 for full description).

Finally, there was a total of 130 filler nonwords. Approximately half of the fillers were generated using the ARC nonword database (Rastle et al., 2002). These items were 2-5 letters long and consisted of only orthographically existing onsets and word bodies. The other half of the fillers were selected from Pritchard et al. (2012) nonwords, a data set that consisted of items that the DRC and the CDP+ disagreed on. Table 4.2 summarises the number and type of nonwords used in the naming task (see Appendix 3, Tables 3B and 3C for full list of stimuli).

Table 4.2

Types of stimuli used in the naming and rating tasks

Nonword Naming Task				Nonword Rating Task			
Item Type	Example	In a Block	Total	Example		In a Block	Total
				text	*audio*		
Irregular-low	lusque	14	28	lusque	lVsk/lusk	28	56
Irregular-high	hauge	14	28	hauge	h1_/h$_	28	56
Regular-low	heint	14	28	-	-	-	-
Regular-high	foathe	14	28	-	-	-	-
Error	-	-	-	dwal	jEsts	5	10
Odd	-	-	-	gloost	glEst	5	10
C-critical	cepth	3	6	cepth	kEpT/sEpT	6	12
C-control	cupth	2	4	cupth	kVpT/sVpT	4	8
G-critical	gilsh	3	6	gilsh	gIlS/_IlS	6	12
G-control	galsh	2	4	galsh	g{lS/_{lS	4	8
Fillers	wholt	65	130	wholt	wQlt/w5lt	65	130
Critical trials			132				152
Filler trials			130				150
Total no. of trials			262				302

4.2.2.2 Rating task. For the rating task, different pronunciations of the experimental nonwords and a subset of the fillers used in the naming task were recorded[30]. Additionally, 10 Error items (nonwords associated with an implausible pronunciation) and 10 Odd items (nonwords associated with an unusual pronunciation) were included; analysis of the responses to these items are reported in Chapter 6. The audio files of the pronunciations were recorded in a silent room, spoken by a male, native speaker of English. The intensity of the resulting recordings was equalized to 70 dB. With a 100 ms silence at the beginning and end of each sound file, the mean duration of the sound files was 807.5 ms (*SD* = 124.2). See Table 4.2 for a summary of the item types used in the rating task and Appendix 3 (Tables 3D and 3E) for a full list of stimuli.

4.2.2.3 Vocabulary task. Items used in the vocabulary task were all the base words for the Irregular items (28 words). These words were presented with four definitions, the correct one and three foils (see Appendix 3, Table 3F for full list of stimuli).

4.2.3 Procedure

The two versions of the experiment, the Naming-Rating version and the Rating-Only version were run using Gorilla Experiment Builder (Anwyl-Irvine, et al., 2019). Participants completed all tasks online, on a computer. The order of the tasks was the same for each participant, starting with the nonword naming task (for Naming-Rating group only), followed by the nonword rating task and the vocabulary task. Each participant was randomly assigned to a version of the experiment. The stimuli in the nonword naming task and the nonword rating task were presented in two blocks, separated by a break, whereas the vocabulary task consisted of only one block of stimuli. The order of the blocks was randomised across participants and the presentation of the stimuli within a block in all three tasks was randomised for each participant. There were five practice trials for each task before the actual task began. The nonwords or words in each task were presented centrally on the computer screen, in capital letters and black font, against a white background.

4.2.3.1 Naming task. Participants were instructed to read aloud nonwords presented on the screen and they were told they had 3 seconds to pronounce each nonword before the next trial began. They were also instructed to pronounce the nonword again if they thought they had

[30] The experimental items in the rating task had otherwise identical orthographic form to the experimental items in the naming task, except for DWONGE and PHOUTE in the naming task corresponded to PHONGE and DWOUTE in the rating task, due to an error in stimuli creation (i.e., the onsets were swapped between these items).

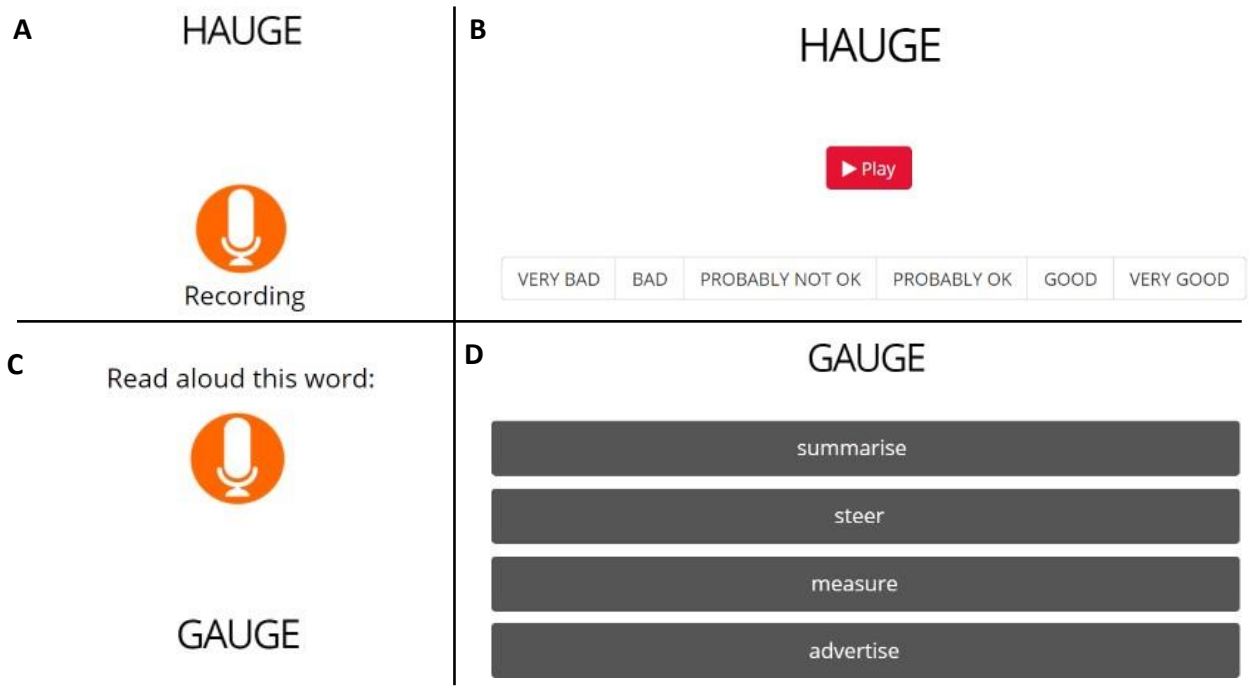
mispronounced the item. Each nonword was shown for 3000 ms, followed by a 200 ms blank screen before the next trial started. Each experimental nonword (132 items) and filler (130 items) was presented once. Nonwords that shared a base word (e.g., *hauge* and *snauge*, based on *gauge*) were presented in different blocks.

4.2.3.2 Rating task. Participants were instructed to assess how well the written forms of nonwords presented on the screen matched the pronunciation assigned to them. In each trial, below the nonword, there was a Play/Replay button, which participants would click to listen to the pronunciation for the nonword. Participants could listen to the same pronunciation up to five times. Under the Play/Replay button, there was a 6-point Likert scale with options VERY BAD, BAD, PROBABLY NOT OK, PROBABLY OK, GOOD, VERY GOOD. Participants would click one of the options and advance to the next trial (See Figure 4.1 for an example trial). The Irregular-low and Irregular-high nonwords were presented twice during the task, once associated with a regular and once with an irregular pronunciation (e.g., *snauge* would be pronounced as /sn$_/ in block 1 and as /sn1_/ in block 2 or vice versa). Approximately half of each item category was presented with a regular pronunciation in the first block and with irregular pronunciation in the second block. For nonwords that shared a word body (e.g., *hauge* and *snauge*), both items were presented in each block, so that the vowel pronunciation associated with each nonword was different than that associated with the other nonword within a block (e.g., pronunciations /h1_/ and /sn$_/ in block 1 and pronunciations /h$_/ and /sn1_/ in block 2 or vice versa).

4.2.3.3 Vocabulary task. Participants were instructed to read aloud words that appeared on the screen and choose the best definition for each word out of four options. They had three seconds to read aloud each word, before a 1000 ms blank screen appeared, followed by four definitions below the word. Participants clicked the definition that best corresponded to the meaning of the word on top of the screen (see Figure 4.1). The next trial started automatically after a 1000 ms blank screen. The items in the vocabulary task were presented in a randomised order and the correct definition was the 1st, 2nd, 3rd and the 4th option seven times each. The length of the definitions varied from one to 12 words, with each definition for a given word being approximately the same length.

Figure 4.1

Example trials of the naming, rating and vocabulary tasks



Note. Example trial of the nonword naming task (A), the nonword rating task (B), vocabulary task (C: a word to be read aloud; D: definition options for the word).

4.2.4 Data processing

Pre-processing of the data and analyses were conducted using R 4.0.3 (R Core Team, 2020).

4.2.4.1 Naming data. The verbal nonword naming responses were transcribed independently by four native English speakers, who had received training in the task but were naive to the critical manipulations of the study. The transcription happened in two waves such that two of the transcribers processed all of the items for each participant (first wave), re-processing items with discrepant transcriptions. Items for which the first wave did not reach a consensus on were transcribed by the other two transcribers (second wave). The second wave also re-processed any items with remaining discrepancies, that is, if the four transcribers each produced a different transcription for an item or if there was a tie between two different transcriptions for the same item. Finally, in order to maximise the number of items included in the analyses, any remaining items with discrepant transcriptions were included if the critical phoneme in the transcriptions for these items was agreed on amongst the transcribers

– i.e., the vowel in the Irregular items and the onset in the C/G items. If a participant gave more than one response to an item, the last complete response was transcribed. If the only response available was incomplete, this was transcribed. The first wave agreed on 90 % of the items and the second wave resolved the remaining discrepancies, reaching consensus for 96% of the items, at least regarding the critical phoneme. Taking into account other sources of data loss (see Exclusion of trials below), 0.3% of the items were lost due to inconsistencies in transcription.

For the analyses, the naming responses for the Irregular and Regular items were categorised as regular, irregular or other based on the critical vowel pronunciation (only regular or other categories were used for Regular items). Proportions of irregular and regular pronunciations were calculated for each participant and for each item. The by-subjects proportions were used in the analyses investigating the role of token frequency in nonword naming (Section 4.3.1), and the by-items proportions were used in comparisons to those found in Andrews and Scarratt (1998, Exp. 2, see Section 4.3.2). Two Irregular-low items (*crauche* and *jauche*) were excluded from the analyses as no participant had correctly defined or pronounced the base word (*gauche*) for these items. For comparisons of human data with the output from computational models, a human modal response was extracted from the naming data for the Irregular items as the most popular human naming response that was either regular or irregular (there were five items for which 'other' was the modal response, in which case the second most popular response (regular or irregular) was used). After excluding two items with no valid responses (*crauche* and *jauche*) and four items with a tie for a modal response, the remaining sample size was 23 for Irregular-low items and 27 for Irregular-high items. The number of items for each computational model depended on how many items were pronounced regularly or irregularly (i.e., 'other' responses were removed). The final number of items included in the analyses for each model can be seen in Table 4.4 (Section 4.3.3).

4.2.4.2 Rating data. The labels of the rating scale were re-coded as follows: 1 = VERY BAD, 2 = BAD, 3 = PROBABLY NOT OK, 4 = PROBABLY OK, 5 = GOOD and 6 = VERY GOOD. Mean ratings from each participant were then calculated, for each item group (e.g., irregular and regular pronunciations paired with Irregular-low and Irregular-high items), thus allowing the use of parametric hypothesis tests (Carifio & Perla, 2007).

4.2.4.3 Vocabulary data. Verbal responses in the vocabulary task were transcribed by two transcribers (the second wave transcribers). Overall, the transcribers agreed on 87%

(Naming-Rating group) and 86% (Rating-Only group) of the items. When discrepancies occurred, a strict inclusion criterion was used, such that an item was considered incorrect if even one of the transcribers deemed it incorrect. If both transcribers deemed a pronunciation of an item correct but did not agree on how this item was pronounced, it was excluded.

4.2.4.4 Exclusion of trials. Following the pre-registered data processing plan (https://osf.io/znpyf), for each participant, trials in the nonword naming and rating tasks for Irregular items were excluded from the analyses if a nonword in a given trial was based on a word the participant mispronounced or chose an incorrect definition for in the vocabulary task (e.g., *hauge* and *snauge* would be excluded if the participant pronounced or defined the word *gauge* incorrectly). As stated in the data processing plan, both strict and lenient scoring criteria for pronunciations given in the vocabulary task were used. As such, pronunciations with all the phonemes (strict) or only the vowel pronunciation (lenient) matching the pronunciation of the vocabulary item were deemed correct. However, due to already large number of lost trials, the analyses reported below were all based on the lenient criterion. Other sources of data loss were audio recording issues in the vocabulary and naming tasks, the former affecting both naming and rating data and the latter affecting only the naming data. (See Table 4.3 for summary of lost data).

Table 4.3

Percentage of lost trials in the Naming and Rating tasks

Data set	Vocab. Semantic	Vocab. Pronunciation	Vocab. Recording	Naming Recording	Unresolved transcription	Total
Naming data	16.82	14.39 (16.56)	2.28	0.91	0.28	34.68 (36.85)
Rating (NR)	16.82	14.39 (16.56)	2.28	-	-	33.49 (35.66)
Rating (RO)	18.37	15.32 (16.68)	2.23	-	-	35.92 (37.58)

Note. NR = Naming-Rating group, RO = Rating-Only group. The percentage lost trials based on the strict criterion in parenthesis for pronunciations of the vocabulary items (Vocab. Pronunciation) and the total percentage of lost trials (Total). The percentage of data loss is based on maximum total number of trials, 3864 (69 participants * 56 trials) for the naming data and 7728 (69 participants * 112 trials) for the rating data.

4.2.4.5 Statistical power. Sensitivity power analyses were computed using GPower (Faul, et al., 2007) for each hypothesis test and the resulting minimum, reliably detectable effect sizes for each analysis are reported along with the observed effect sizes from the analyses.

4.2.4.6 Transcription of output from computational models. The phonemic transcription of the output from the DRC, CDP++ and WSP models is based on the DISC phonemic alphabet. However, as described in Chapter 3 (Section 3.1), two changes were made to the output from the DRC model (namely, yod-pronunciations were changed from /W/ to /ju/ and phoneme /9/ was changed to the phoneme /$/). Unifying the transcription between the aforementioned models and the Psim1 model was more difficult, because the dialect of this model is based on North American English. The phonemic transcription used in Plaut et al. (1996) was changed into DISC, which was relatively straightforward for most phonemes. However, the vowel phonemes /Q/ (as in *pot* in British English) and /$/ (as in *door* in British English) were not perfectly matched with the vowel phonemes of the transcription used for the Psim1 model. In Plaut et al. (1996, Appendix C, p. 115), phoneme /o/ was linked with words *dog, broad* and *wash* and phoneme /a/ with words *pot, want* and *watch*. I thus equated the phoneme /o/ with /$/ and phoneme /a/ with /Q/. Additionally, transcription of /Or/ was equated with /$/ (as this phoneme was used for transcribing words like *swarm*). However, I acknowledge that the dialect differences between the British participants in the current study and the output from the Psim1 model may result in discrepancies that should not be considered a weakness of the Psim1 model (see Appendix 4).

4.3 Results

4.3.1 Effects of token frequency in nonword processing

I expected higher proportion of base word congruent responses to items with high token frequency compared to items with low token frequency. I also predicted higher acceptability ratings for irregularly pronounced Irregular-high items compared to Irregular-low items. Due to these directional hypotheses, the comparisons are conducted as one-tailed tests.

4.3.1.1 Naming responses. Paired samples t-tests were conducted for the Irregular and Regular items separately. For the Irregular items, the proportion of irregular pronunciations was higher for the Irregular-high items (M = .29, SD = .01) than for Irregular-low items (M = .25, SD = .16), and this difference was statistically significant ($t(68) = 1.98$, $p = .03$, dz = 0.24). For the Regular items, the proportion of regular pronunciations was higher for the Regular-high items (M = .91, SD = .08) than for the Regular-low items (M = .83, SD = .08), and this difference was also statistically reliable ($t(68) = 7.43$, $p < .001$, $dz = 0.89$). [31] See

[31] Sensitivity analyses was computed for a 1-tailed, paired samples t-test with an alpha level of .05, power of .8 and sample of 69, which yielded a minimum, reliably detectable effect size as Cohen's dz = 0.3.

Further investigation of human naming responses (Section 4.3.4) for additional considerations of these findings.

4.3.1.2 Rating responses. The mean acceptability ratings for irregularly pronounced Irregular-low and Irregular-high items were compared with a paired sample t-test (1-tailed) for Naming-Rating and the Rating-Only groups. The Naming-Rating group gave slightly higher ratings to Irregular-low items (M = 4.84, SD = 0.51) than to Irregular-high items (M = 4.79, SD = 0.52), but this difference was not statistically reliable ($t(68) = -1.24$, $p = .89$, $dz = -0.15$). By contrast, the Rating-Only group gave reliably higher ratings to the irregularly pronounced Irregular-high items (M = 4.49, SD = 0.53) compared to the Irregular-low items (M = 4.32, SD = 0.53, $t(68) = 3.18$, $p = .001$, $dz = 0.38$). [32]

Due to a considerable data loss described in Data processing (Section 4.2.4.4), the number of valid trials within a condition remained low for some participants. As such, the analyses of both naming and rating data were also run with more reliable individual means, where only participants with at least 10 valid trials within each condition were included. These analyses resulted in a comparable pattern of results (See Appendix 5), although the analyses were underpowered due to reduced sample sizes.

Importantly, as a demonstration of the importance of taking the vocabulary knowledge of the base words into account in the analyses reported thus far, when the rating analyses were run with the full set of data, that is, ignoring the vocabulary knowledge of the base words, both groups showed a statistically significant effect of token frequency (Naming-Rating group: $t(68) = 2.19$, $p = .02$, $dz = 0.26$; Rating-Only group: $t(68) = 3.7$, $p < .001$, $dz = 0.45$, 1-tailed). However, this difference in acceptability ratings was very likely the result of fewer items in the Irregular-low group for which the irregular word body PSC was known by the participants, which lead to lower acceptability ratings for irregularly pronounced Irregular-low items.

4.3.2 Influence of consistency and type frequency in nonword reading

I had predicted that the incidence of irregular pronunciations to the experimental items would be higher in my study compared to the study by Andrews and Scarratt (1998, Exp. 2), because the vocabulary knowledge of the base words was taken into account in the current

[32] Sensitivity analysis was computed for a 1-tailed, paired samples t-test with an alpha level of .05, power of .8 and sample of 69, which yielded a minimum, reliably detectable effect size as Cohen's dz = 0.3. This applies for both Naming-Rating and Rating-Only groups.

study, thus avoiding the potential underestimation of the proportion of irregular pronunciations. However, this was not the case, as the incidence of irregular pronunciations in my study was much lower (.28) than that reported by Andrews and Scarratt (.4). Both of these proportions are based on by-items calculations. The lower proportion found in my study was not due to the Irregular-low items dragging the average down, as the proportions for both types of items were very close to the overall average (Irregular-low: .26, Irregular-high: .28). These by-items proportions of irregular pronunciations in the current study were predicted most closely by the CDP++ (.25) and Psim1 (.3) models, followed by the WSP-token-T (.50), DRC (.04), WSP-type (.91) and WSP-token (.93).

4.3.3 Comparison of computational models in reading irregular singleton items

A chi-squared test for independence (with Yate's correction where required) was conducted to test whether there was an association between the type of response given (regular or irregular pronunciation) and the type of item (Irregular-low and Irregular-high) in the human modal responses and in the output from each computational model (Table 4.4). No reliable associations were found. To put this finding in terms of my research question, the proportion of irregular pronunciations for low and high Irregular items did not differ from one another in human responses or in the output from the computational models. Note that the proportions of irregular vowel pronunciations depicted in Table 4.4 for the human data are based on human modal responses relative to all the items with a clear irregular or regular modal response in the human data (i.e., ties and other responses were excluded). Therefore, these proportions differ from the ones reported in Sections 4.3.2 and from the by-subjects based proportions in Section 4.3.1.

Table 4.4

Comparison of the proportion of irregular pronunciations assigned to Irregular-low and Irregular-high items by humans and computational models

Source	Item type			Chi-squared test for Irregular-low vs. Irregular-high				
	Irregular-low	*Irregular-high*	*Irregular total*	X^2	*df*	*n*	*p-value*	*Yate's correction*
Humans	.30	.44	.38	1.03	1	50	.31	no
DRC	.08	.00	.04	0.52	1	52	.47	yes
CDP++	.27	.22	.25	0.16	1	53	.69	no
Psim1	.28	.32	.30	0.08	1	47	.78	no
WSP-type	.93	.89	.91	0.00	1	56	1.00	yes
WSP-token	.93	.93	.93	< 0.001	1	55	1.00	yes
WSP-token-T	.50	.50	.50	0.00	1	56	1.00	no

Considering the absolute values of proportions of irregular pronunciations, the Psim1 model's output was the closest match to the human responses, followed by that of CDP++ and the WSP-token-T. The vocabulary-optimised versions of the WSP model clearly overestimated the incidence of irregular pronunciations to both types of items, which suggest that the consistency of PSCs has too strong of an effect in the competition of pronunciation options. By contrast, the DRC model underestimated the incidence of irregular pronunciations, which was largely due to the model's inability to produce irregular pronunciations (two items with the word body *ugue* were pronounced irregularly, presumably due to a word body rule for this orthographic segment in the newer versions of the model). The pattern of the proportions of irregular pronunciations to Irregular-low and Irregular-high items in the human data was best reflected in the output from Psim1, which, like humans, produced a numerically larger proportion of irregular pronunciations for the Irregular-high items compared to the Irregular-low items. Surprisingly, the versions of the WSP sensitive to token frequency did not show this pattern, which was likely due to the consistency of the word body sized PSCs dominating in the competition between parsing styles: for the WSP-token, the proportion of irregular pronunciations was at ceiling, regardless of token frequency and 91% of the items were pronounced based on the word body parsing style. However, for the WSP-token-T, the word body PSCs were not as dominant – 48% of the items were pronounced according to this parsing style. It is unclear why this version of the model also failed to produce the expected pattern of naming responses (further discussion on a similar finding is discussed in Chapter 5,

Section 5.4). Furthermore, the CDP++ produced an opposite pattern to that in the human data, even though this model should be sensitive to token frequency of PSCs.

The comparisons described above do not tell us whether the models with similar proportions of irregular responses to those found in human data named the same items irregularly or regularly as humans did. To answer this question, I calculated the proportion of items for which each model matched the type of human modal response (regular or irregular, based on vowel pronunciation), across all items and for items with irregular or regular human modal response separately (see Table 4.5).

Table 4.5

Proportion of matching pronunciation types between human modal responses and output from computational models for Irregular items

Human pron. type	DRC	CDP++	Psim1	WSP-type	WSP-token	WSP-token-T
total (n = 50)	.66	.56	.46	.44	.42	.52
regular (n = 31)	1.00	.74	.68	.13	.10	.55
irregular (n = 19)	.11	.26	.11	.95	.95	.47

Note. Human pron. type = subsets of Irregular items based on the type of human modal response (regular or irregular).

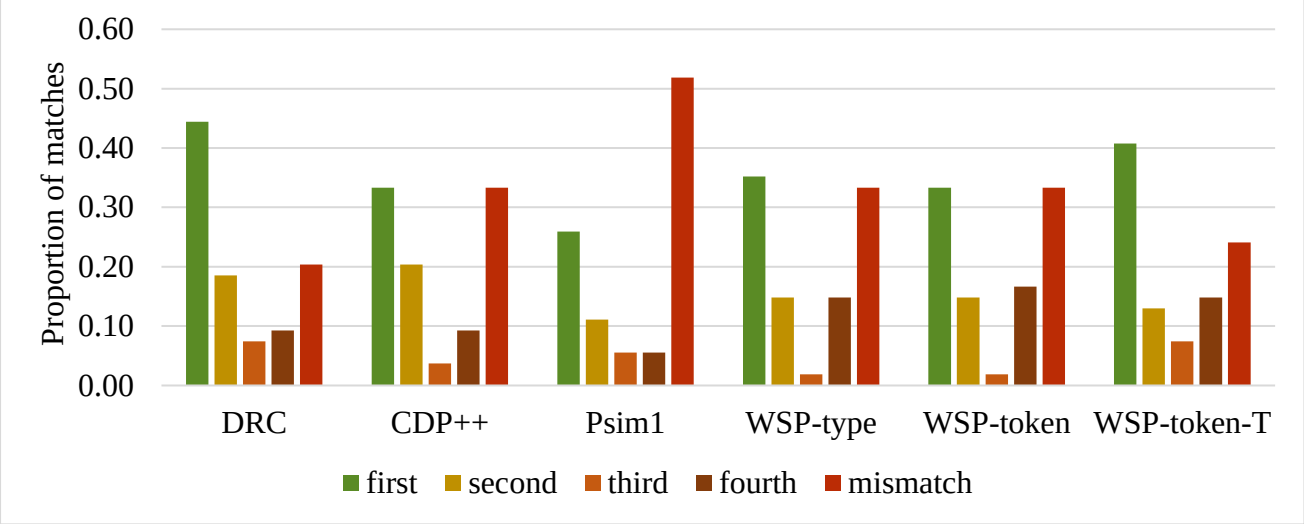
As is evident in Table 4.5, the highest proportion of matches to the human modal responses was produced by the DRC model, followed by CDP++, WSP-token-T, Psim1 and the two vocabulary-optimised versions of the WSP model. Looking at the items for which most participants assigned a regular pronunciation, it is clear that the high proportion of overall matches by the DRC, CDP++ and Psim1 models was mostly due to a high incidence of regular pronunciations assigned to the Irregular items – most items pronounced regularly by the participants were also pronounced regularly by these models. By contrast, the total proportion of matches produced by the vocabulary-optimised versions of the WSP model are mostly driven by a high incidence of irregular pronunciations – most items pronounced irregularly by participants were also pronounced irregularly by the versions of WSP model. The models that produce the most balanced output, i.e., relatively high proportion of matches in both regular and irregular categories, are the WSP-token-T, followed by the CDP++.

Finally, a finer grained inspection of the models' performance was carried out, in which each model's output was compared to the human naming responses by considering the whole

nonword pronunciation, rather than just the vowel. The participants' naming responses to 54 irregular nonwords for which at least one valid response was available were categorised by response frequency and any matching responses from each model were categorised accordingly into the first, second, third and fourth or lower most popular human responses. Number of mismatches, that is, items where a model's output was not produced by any participant, were also calculated (see Figure 4.2 for summary and Appendix 6 for full list of output to the stimuli and mismatches by the models).

Figure 4.2

Proportions of human-model matches to irregular nonwords arranged by human response frequency



Note. Proportion of matches was calculated for all 54 Irregular nonwords that had at least one valid response in the human data.

The DRC model matched the largest proportion of human modal responses (.44), and it also produced the lowest proportion of mismatches (.2). The second best performing model was the WSP-token-T, with .41 of the items matching human modal responses and .24 as mismatches. The output from the CDP++ and the vocabulary-optimised versions of the WSP model were similar regarding the proportion of first matches (from .33 to .35) and mismatches (.33) to the human data. However, the Psim1 model did not match any participants' response for over half of the items and the proportion of human modal responses from this model was also the lowest (.26).

Inspection of the mismatches produced by each model was carried out next. For the DRC model, the 11 mismatches were due to either vowel pronunciations not found in the human data (e.g., /6si/ for word body *ousse* and /V1v/ for word body *uave*) or minor differences such as yod pronunciation for *ghuede* and *shugue*, which were not produced by the participants. For the CDP++ model, the largest number of the 18 mismatches were due to an unusual or illegal coda (e.g., /z{tt/ for *zacht* or /fl3zJ/ for *flirsch*). There were also three items for which the model assigned a pronunciation /p/ to the onset *ps*, unlike any of the participants. Finally, two unusual vowel pronunciations were identified (e.g., /br$lt/ for *broult* and /dwuv/ for *duave*), along with two instances of irregular vowel pronunciations not produced by any of the participants (e.g., /stVnT/ for *stonth*). For the Psim1 model, most of the 28 mismatches were due to an unusual vowel pronunciation (e.g., *duave* as /dQv/ and *gleart* as /gl$t/) or a vowel pronounced regularly or irregularly when no participant did this. Seven nonwords also received an unusual coda or no coda at all (e.g., *shoung* as /SV/ and *lusque* as /lus/) and five mismatches were a result of a missing or unusual onset (e.g., *phealm* as /ilm/ and *zacht* as /QJ/). Turning to the three versions of the WSP model, nearly all of the 18 mismatches of WSP-type and WSP-token and the 13 mismatches of the WSP-token-T, were due to an irregular vowel pronunciation not produced by any of the participants (e.g., *neanse* as /nEnz/ and *snauge* as /sn1_/). Finally, all six models failed to match three to four out of four items that were named by only one participant (*flirsch*, *dwurgh*, *mirsch* and *wurgh*). Although the responses to these items are highly varied, based on a single participant, they were nevertheless retained in the analysis so that no more data would be lost, and because each model would be equally disadvantaged at trying to match these single responses.

Overall, no model performed particularly well on this data set – even the best performing model (the DRC) still failed to match 20% of the items, and only produced the most common human response for 44% of the items. The general issue for all the models compared above was producing a regular or irregular vowel pronunciation when they were not produced by participants – this was the main weakness of the DRC and especially the two vocabulary-optimised versions of the WSP. Apart from the vowel pronunciations, the CDP++ and Psim1 models also struggled with onsets and codas in this data set, producing illegal pronunciations or omitting these segments altogether. Note, that while the mismatches in vowel pronunciations between human responses and the Psim1 output may be partly due to dialect differences (or, indeed, the choices made in converting the original phonemic transcription

into DISC), the same explanation does not apply to issues with onsets and codas produced by Psim1.

4.3.4 Further investigation of human naming responses

Overall, skilled readers tended to pronounce most of the Irregular items regularly. Yet, some items received a relatively high proportion of irregular pronunciations (e.g., *fousse:* .78, *meird:* .75). We are thus left with the question, what is it about the irregularly pronounced items that elicit an irregular pronunciation? How are they different from the items that do not? As consistency and type frequency of the word body segment in these items is the same, but the difference was not solely based on token frequency of the body segments, the answer to this question must be elsewhere, such as in the properties of the vowel segment. I therefore investigated the properties of the items that received a regular human modal response (regular-modal items) and items with irregular human modal response (irregular-modal items) with the aid of the statistical properties in the WSP model's PSC knowledge.

4.3.4.1 Embedded regularly pronounced words. A potential influence on nonword pronunciations are existing words that are embedded within the nonword. Several studies have demonstrated the influence of embedded words in visual word recognition (e.g., Bowers et al., 2005; Nation & Cocksey, 2009; Snell et al., 2018). Since activation of the embedded words appears highly automatic during reading, access to the phonological forms of the words may affect the pronunciations assigned to nonwords. Inspection of the experimental nonwords revealed that some of them did contain existing words. Due to the limited number of items available for the experimental manipulations of the study, this issue was not taken into account during stimuli construction. Out of the 31 regular-modal items, 15 had an embedded, regularly pronounced existing word, such as *or* in nonwords based on *world*, *worl* or *worst*. Furthermore, all of these words had a higher token frequency than the base words they were embedded in. By contrast, out of the 19 irregular-modal items, seven contained regularly pronounced existing words, such as *ear* in nonwords based on *heart*. Here the token frequencies of the embedded words showed a less clear pattern, as some of them had higher and some lower token frequencies than the base words they were embedded in. There were four nonwords that contained a regular embedded word with a higher token frequency than the base words (e.g., *out* in nonwords based on *route*). Although this is a small number of items, the fact that these four items were still pronounced irregularly by majority of the participants suggests that embedded, regularly pronounced words do not solely dictate the

type of pronunciations nonwords receive. As a comparison, there were only two nonwords that had an irregularly pronounced embedded word in regular-modal items (*won* in *dwonge* and *wonge*) and only one such nonword in irregular-modal items (*we* in *tweize*)[33].

4.3.4.2 Properties of vowel segments. Another potential source of influence on pronunciation choices made by skilled readers is the statistical properties of the vowel segments in the nonwords. For instance, if the vowel alone is highly consistent and the pronunciation assigned to it occurs in several, highly frequent words, skilled readers may be less likely to pronounce these vowels based on larger units of PSCs, such as word bodies. By contrast, if there is more uncertainty about how to pronounce the vowel, skilled readers may be more likely to rely on contextual information and thus base their pronunciations on larger units of PSCs. This type of explanation has been suggested (Kessler, 2009) for the observation that skilled readers do not utilise context sensitive correspondences as often as would be expected based on the statistics of the writing system (Kessler, 2009; Steacy et al., 2019; Treiman et al., 2003).

Thus, I tested this idea with the assumption that if the reliability of the vowel-sized PSCs influences the unit size adopted in nonword reading, it should be seen in correlations between the statistical properties of the vowel sized PSCs and the proportions of irregular and regular pronunciations assigned to nonwords that contain these vowel segments. For each nonword with at least 10 valid responses (for both Irregular and Regular items), values for the vowel consistency, type frequency, summed token frequency and maximum token frequency were extracted from the PSC knowledge of the WSP. These properties are based on monosyllabic, mostly monomorphemic words. Split vowels were considered for any item where the first and the second part of the split vowel were separated by one letter (e.g., vowel for *bleize* was *ei_e*) and otherwise only the first part of the split vowel was considered (e.g., vowel for *steanse* was *ea*). The proportion of base word congruent responses (irregular vowel pronunciation for Irregular items and regular pronunciations for Regular items) for each item was correlated with each of these properties (see Table 4.6). For the Irregular items, the proportion of regular pronunciations was also correlated with the properties of the vowel segment. Table 4.6 also shows the mean values for each of the vowel segment properties.

[33] Note, the nonword *tweize* is in fact an orthographic neighbour and a homophone when pronounced irregularly to the word *tweeze.* This word is not found in WebCelex, and such it was missed during stimuli construction. While the high proportion of irregular responses to this item (.81) may be due to lexicalisation, the other nonword with the same word body, *bleize,* also received more irregular pronunciations (proportion of .45) than regular ones (proportion of .34).

Table 4.6

Correlations between proportions of base word congruent responses and statistical properties of the vowel segment of Irregular and Regular nonwords

	Regular items			Irregular items			
Vowel property	*BWC prop*	*Mean low*	*Mean high*	*BWC prop*	*reg. prop*	*Mean low*	*Mean high*
Consistency	.18	.83	.92	-.13	.24	.81	.71
Type Freq	.29 *	150.07	111.57	-.35 *	.5 ***	142.00	115.93
Sum Token Freq	.32 *	392.28	311.10	-.40 **	.55 ***	363.85	315.70
Max Token Freq	.38 **	6.19	6.23	-.43 **	.50 ***	5.90	6.26
number of items	56	28	28	50	50	22	28

Note. BWC prop = proportion of base word congruent responses. Mean low and Mean high columns depict mean values for each property of the vowel segments for low and high token frequency items, respectively. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05.

As seen in Table 4.6, each property of the vowel segments was positively associated with the proportion of regular pronunciations for both Irregular and Regular items and negatively associated with the proportion of base word congruent responses for the Irregular items. In other words, the more certainty there was about the context insensitive pronunciation of the vowel, based on any of the frequency measures, the higher the proportion of regular pronunciations assigned to the nonwords containing this vowel. The summed token frequency of the vowel-sized PSC had the strongest correlation with the proportion of regular and irregular pronunciations for the Irregular items and the maximum token frequency of the vowel-sized PSC was most strongly associated with the proportion of regular pronunciations assigned to the Regular items.

Considering the role of token frequency in nonword naming in light of these additional findings (i.e., the potential influence of vowel properties in base word congruent responses), the difference of the proportion of base word congruent responses found between low and high frequency items could be explained with the properties of the vowel segments if these properties differ between low and high items. Indeed, both Irregular-low and Regular-low items had numerically more reliable vowel sized PSCs based on two out of three frequency measures than the corresponding high items (Table 4.6). However, none of the differences in the frequency measures between low and high items were statistically significant (see Appendix 7 for the analyses). Furthermore, the numeric trend of higher frequency vowel

segments for Regular-low items would still not explain the higher proportion of base word congruent responses for Regular-high items than for Regular-low items. As such, even though the properties of the vowel segments were not considered during construction of the experimental stimuli, the difference in base word congruent pronunciations between low and high items in Irregular and Regular item groups cannot be attributed solely to the differences in the properties of the vowel segments.

4.4 Discussion

The current study aimed to answer two questions, using irregular singleton nonwords that were named and rated by skilled readers: 1) whether token frequency plays a role in nonword processing, 2) whether computational models sensitive to token frequency produce more human-like output for irregular singleton nonwords than models that are not. Exploratory investigation on the influence of consistency and type frequency in nonword reading was also conducted, where the incidence of irregular pronunciations of the experimental items in the current study were compared with those from a previous study (Andrews and Scarratt, 1998) and to the output from computational models. Finally, additional investigations of the pattern of naming responses were conducted to clarify why certain nonwords elicit irregular vowel pronunciations while others do not.

To answer the first research question of the current study, the role of token frequency in nonword processing was investigated with both nonword naming and nonword rating methods. The results from the analyses using both methods pointed to the same conclusion – token frequency has a small effect in nonword processing. However, there were issues of power in the analysis of the naming data for Irregular items: the observed effect size (dz = 0.24) was smaller than what the analysis with 69 participants was sensitive to detect (namely, effects of dz = 0.3 or larger). The naming responses for the Regular items, by contrast, showed a clear effect of token frequency: a higher proportion of regular pronunciations assigned to nonwords with high token frequency compared to nonwords with lower token frequency (however, see limitations discussed below). The results from the rating data were somewhat mixed, as the Rating-Only group showed the expected pattern of rating behaviour (higher ratings to irregularly pronounced Irregular-high items than to Irregular-low items, dz = 0.38), but the Naming-Rating group showed a statistically non-significant pattern in the opposite direction (higher ratings for Irregular-low items). As such, the question of whether the rating behaviour of the two groups is comparable, given the differing exposure to the

critical nonwords, remains relevant. This question will be addressed in Chapter 6 (Experiment 2). Nevertheless, considering the parts of the analyses that should not be influenced by excessive or different exposure to the critical items (namely, the naming responses from the Naming-Rating group and the rating responses from the Rating-Only group), the general pattern of the results from the naming and rating data point to the same conclusion – token frequency has a small effect on nonword processing.

My predictions for the second research question, namely, that models sensitive to token frequency produce a closer match to human data than models that are not, were partly supported: when proportions of pronunciation types were considered, the absolute and relative proportions of irregular pronunciations in the human data were matched best by models that are sensitive to token frequency (Psim1, and to a lesser extent CDP++ and WSP-token-T). However, the highest proportion of matching pronunciations for each item, whether the match is based on vowel pronunciation or the whole nonword, was produced by a model that does not take token frequency or even larger units of PSCs into account – the DRC model. The success of the DRC in simulating the current naming data was clearly due to the high incidence of regular pronunciations assigned to the experimental items, which is what the DRC model does almost exclusively.

Turning to the exploratory investigations, the influence of consistency and type frequency of PSCs in nonword reading was inspected as an incidence of irregular pronunciations to the Irregular items. This incidence was expected to be larger in the current study than in the study by Andrews and Scarratt (1998), because the vocabulary knowledge of the words the experimental items overlapped with was taken into account in the former, but not in the latter. However, the incidence of irregular pronunciations for the Irregular nonwords in the current study was considerably lower (.28) than that reported previously by Andrews and Scarratt (.4). There are several factors which might explain this difference.

Firstly, in the current study, only items with unique word bodies were chosen (see Section 4.1.3), based on mono- and disyllabic words, whereas Andrews and Scarratt's nonwords were not all unique. However, as the potential influence of the additional base words in Andrews and Scarartt's study was not always in the same direction (i.e., some items had regular, others irregular or both types of base words), this is unlikely the only basis for the difference found.

Secondly, the frequency of some of the base words in the current study was quite low, as the mean of Irregular-low items was 2.69 on Zipf scale, while the mean for Irregular-high items

was at 4.6. By comparison, the Zipf frequency of the Andrews and Scarratt's items was 4.23, which is fairly similar to the mean of Irregular-high items in the current study. Yet, even the Irregular-high items in the current study did not receive irregular pronunciations as often as the items in Andrews and Scarratt's study. Thus, this explanation is not sufficient.

Thirdly, Andrews and Scarratt's nonwords all had a single letter onset, whereas 61% of the Irregular items in the current study had a two-letter onset (importantly, as described in Section 4.2.2, the low and high items, for both regular and irregular items, had the same number of one-letter and two-letter onsets). The large number of more complex onsets in the current study was due to avoiding the construction of nonwords with orthographic neighbours, which often required using more complex onsets. It is possible that irregular nonwords with complex onsets elicit more regular pronunciations than nonwords with simpler onsets, for example because single-letter onsets are easier to process and thus reserve cognitive resources to searching for available lexical analogies for the remaining word body. Currently available nonword reading data sets do not contain enough items with the same irregular word body and varied onset complexity to test this idea, but it would be worth investigating in the future.

Regardless of the potential reason for the difference, there are now two nonword reading data sets that provide different estimates for the incidence of irregular pronunciations to nonwords with highly consistent PSCs but low type frequency. Based on these data sets, it appears that the consistency of PSCs has less of an effect on pronunciation choices when type frequency of PSCs is low. This is a particularly important finding for models with strong emphasis on the influence of consistency – the current data create strong pressure for making changes in the WSP model, which clearly overestimated the influence of consistency in print-to-sound conversion. However, this is a complex problem, given that the modified WSP model (or any computational model of reading) should simulate reading behaviour where certain items with these properties are pronounced irregularly, even when majority of these items are pronounced regularly. This type of behaviour cannot be achieved in the current form of the WSP model. Other models of reading compared in the current study were more successful in simulating the human data due to their high incidence of regular pronunciations for the experimental items. However, for a considerable proportion of the items, these models also failed to produce the same pronunciation type that the participants did. In other words, while the need for finding a better balance between regular and irregular pronunciations for high consistency-low type frequency items is particularly acute for the WSP model, the current

data set poses a problem that none of the current computational models have resolved successfully. My preliminary investigations into what other properties of the PSCs might influence the choice of pronunciations for the experimental items suggested that the reliability of the vowel segment might be important. This idea has been suggested previously (Kessler, 2009; Steacy et al., 2019) and modifications to the WSP model based on this idea may be beneficial (see Chapter 7, Section 7.5.3).

I now turn to the limitations of the current study. Firstly, there was considerable loss of data in the naming task (35% of trials) and the rating task (34% for Naming-Rating group and 36% for Rating-Only group), which was mostly due to some of the base words with low token frequency being not known well enough by the participants. However, it is important to bear in mind that this exclusion of trials was planned before the data processing, and was deemed necessary to ensure an accurate comparison between low and high Irregular items. If items with unknown base words had been included in the analysis, the difference between the Irregular-low and Irregular-high items would have been exaggerated. This is because majority of the unknown items were in the Irregular-low group, thus increasing the preference of regular pronunciations for these items, when the word body sized PSCs – dependent on lexical knowledge – are not available. In other words, the exclusion of items with unknown base words makes it more difficult to find support for my hypothesis that token frequency plays a role in nonword reading. Importantly, an analysis of the rating data supports this argument – when ratings for all items were included, the difference between acceptability ratings for irregularly pronounced Irregular-low items and Irregular-high items was larger than in the analysis where items with unknown base words were excluded (Section 4.3.1.2). Nevertheless, it is possible that the requirement for correct definition *and* pronunciation of the base words was too strict, and resulted in some data loss that could have been avoided, had only one of these criteria been used. This consideration is particularly important given that the vocabulary task was the last task participants completed, and thus lapses of concentration may have been more likely.

Another limitation of the current study became evident when potential other influences on the pronunciations for the experimental items were investigated. It was found that the frequency measures of the vowel-sized PSCs were negatively associated with the proportion of irregular pronunciations assigned to the items and positively associated with the proportion of regular pronunciations given to the items. Numerically, the vowel segments in the Irregular-low and Regular-low items had higher type frequency and summed token frequency than the

corresponding high items, but these differences were not statistically reliable. Nevertheless, the finding in support of the role of token frequency, namely, higher proportion of base word congruent responses for Irregular-high and Regular-high items compared to their low frequency counterparts may need to be considered with this limitation in mind. However, while this is the case for the Irregular items (as the critical difference in base word congruent responses between low and high items would be increased with higher frequency of the vowel segments in the Irregular-low items), the same issue does not apply for the Regular items. Here, the base word congruent responses were higher for the high group, even though the items in the low group had higher frequency vowel sized PSCs. I thus conclude that the findings regarding the properties of the vowel segments do not completely undermine the interpretation of the results from the analyses of token frequency effects in nonword reading. However, the findings regarding the vowel segment properties are something to consider in future studies where the critical naming responses rely on word body sized PSCs.

Finally, the evidence for the effect of token frequency in nonword naming based on the Regular items needs to be re-considered. This is because the base word knowledge for these items was not tested. As such, if the higher proportion of regular pronunciations assigned to Regular-high items than Regular-low items is interpreted as a result of two influences: the GPCs and the word body sized PSCs, both of which lead to or strongly encourage regular pronunciations, then unknown base words in the Regular-low group might reduce the likelihood of regular pronunciations, as only GPCs could be used in pronunciation assignment. This possibility was not fully considered when designing the experiment: the vocabulary knowledge of the base words of the Regular items was not tested because the proportion of regular pronunciations for both low and high items was expected to be at ceiling as GPCs were likely to be enough to elicit regular pronunciations to these items. The current study thus focused on the Irregular items. Although there was relatively high consensus regarding the vowel pronunciations for the Regular items, some items did receive alternative pronunciations, produced by a considerable proportion of the participants. For instance, the Regular-low item *breint* was pronounced as /br1nt/ by 33% of participants, as /br2nt/ by 19% and as /brint/ by 19% of participants. With this limitation, I cannot rule out the possibility that the difference found between Regular-low and Regular-high items actually reflects insufficient knowledge of the base words for Regular-low items, rather than effects of token frequency in nonword pronunciation assignment.

4.4.1 Conclusion

In this chapter, I reported a study investigating the role of token frequency of PSCs in nonword processing. Previous research has mainly suggested type frequency as a more influential property in nonword reading, but limitations in the studies providing this evidence leave it open whether token frequency has an effect as well. I therefore tested whether skilled readers read aloud nonwords and rate acceptability of nonword pronunciations such that PSCs occurring in frequent words are favoured over PSCs in less frequent words. The current study provides some evidence in favour of the effect of token frequency in nonword processing. However, this effect appears to be small. Furthermore, computational models that include token frequency in their print-to-sound conversion did not always outperform models without this property, which was likely related to the high incidence of regular pronunciations assigned to these items in the human data. Even though these experimental nonwords, with highly consistent PSCs exemplified by a single lexical item (i.e., with low type frequency), tended to receive regular pronunciations, approximately quarter of the items were still pronounced irregularly by majority of the participants. This suggests that the statistical properties of PSCs considered in the current study (i.e., consistency, type frequency and token frequency of the word body sized segments) are not sufficient for explaining the pattern of naming responses found for these types of items. None of the computational models of reading considered in the current study simulated the naming responses to these items adequately. The investigations in the current study also highlight additional, important factors to consider in future research, such as the lexical knowledge of the base words and vowel properties of the nonword stimuli.

Chapter 5 : Type frequency in nonword processing

5.1 Introduction

5.1.1 Type frequency in human nonword reading and computational models

Type frequency of print-to-sound correspondences (PSCs) refers to the number of words in which a particular letter cluster is pronounced the same way. When skilled readers assign pronunciations to nonwords, they are likely to use pronunciations corresponding to a given letter cluster in several existing words, rather than pronunciations associated with a given letter cluster in only a single word (e.g., Andrews and Scarratt, 1998, Exp. 2). This measure of frequency is different from token frequency, which refers to the frequency at which a given word occurs in the language. The idea that type frequency of PSCs influences nonword reading is widely accepted, which is demonstrated by the fact that this property is incorporated in several computational models of reading (e.g., Coltheart et al., 2001; Norris, 1994; Perry et al., 2010).

As outlined in Chapter 4, empirical studies investigating the relative importance of type and token frequency of PSCs in nonword reading have not always allowed strong conclusions to be drawn. This has mostly been due to insufficient contrast between the two measures. For instance, Andrews and Scarratt (1998, Exp. 2) demonstrated that irregular pronunciations are more common for items that share a word body with several, irregularly pronounced words (Irregular-many items, e.g., nonword *yight*, based on *might, night, right* etc.) than for nonwords sharing a word body with a single irregular word (Irregular-single items, e.g., *sonth* based on *month*). This finding can be interpreted as an effect of type frequency in nonword naming. However, the token frequency of the base words for these items was not controlled in this study. I calculated the mean token frequencies using a Zipf value (Van Heuven et al., 2014)[34] for Andrews and Scarratt's items; as the maximum token frequency of the base words for the Irregular-many items and as the token frequency of the base words for the Irregular-single items. The Irregular-many items had a higher mean token frequency (M = 5.1, SD = 0.72) than the Irregular-single items (M = 4.23, SD = 0.89) and this difference was

[34] I acknowledge that the Zipf values are based on British English, but as the base words are all relatively frequent, I believe the general pattern of the token frequencies would also apply to Australian English, and thus to the lexical and PSC knowledge of the participants in the Andrews and Scarratt's study

statistically reliable ($t(43.96) = 3.72$, $p < .001$, two-tailed). Therefore, Andrews and Scaratt's Irregular-many items may encourage irregular pronunciations not only because of the higher type frequency, but also due to higher token frequency compared to the Irregular-single items. As demonstrated in Chapter 4, token frequency seems to have a small effect in nonword naming, and as such, the effect of type frequency found in Andrews and Scarratt's study cannot be attributed to type frequency alone.

However, Andrews and Scarratt also provide correlational evidence for the relative importance of type frequency over token frequency – the proportion of regular pronunciations assigned to a nonword were better predicted by properties of the nonwords that were based on type rather than token counts (see Chapter 1, Section 1.1.3). This finding does not suffer from the fact that the frequency measures used were insufficiently distinguishable, because even though the summed token frequency measure contains information about type frequency, the pure type frequency measure was still a better predictor. As such, Andrews and Scarratt provide evidence for both the role of type frequency in nonword reading and the relative importance of type frequency over token frequency in nonword reading.

Similar correlational evidence was provided by Treiman et al. (1990, Exp. 3), where type-based measures of frequency of PSCs accounted more variance in the proportion of correct pronunciations to regular and consistent nonwords, compared to summed token frequency. However, these analyses were based on small sample of skilled readers (n = 15).

Other, less direct evidence for the influence of type frequency in nonword reading comes from studies focusing on naming latencies: Ziegler and colleagues (2001) report that skilled readers are faster at naming words and nonwords with more word body neighbours compared to words and nonwords with fewer word body neighbours. By contrast, a large-scale analysis of nonword naming data (Schmalz et al., 2017), using linear mixed effects regression models and Bayesian analyses, revealed no compelling evidence for effects of word body neighbourhood size (i.e., type frequency of word body sized PSCs for consistent bodies) on nonword naming latencies in skilled readers. Schmalz and colleagues' interpretation of the findings (the study was designed to test the psycholinguistic grain size theory) was that the word body neighbourhood size is not a sensitive measure for reliance on word body sized PSCs. They further suggest that whether an item has a word body neighbour or not could serve as a better measure. In line with this, the nonwords in these analyses had mostly regular

bodies, which I confirmed using the WSP model's PSC knowledge – only seven out of 218 items would result in a different pronunciation based on word body sized segment rather than the vowel segment alone. Naming latencies for mostly regular nonwords might not differ from one another enough to detect a difference as a function of word body neighbourhood size. This could be because converging support for a pronunciation from both GPC and body sized PSCs is sufficient for facilitating pronunciation decisions, such that additional support from more word body friends may be redundant in this context. By contrast, nonwords with irregular word bodies would allow a clearer distinction between the different sub-lexical unit sizes relied on in nonword reading, as the type of responses (regular or irregular vowel pronunciation) indicate whether word body or smaller PSCs were applied (see Chapter 1, Section 1.1.4).

Overall, conclusive empirical evidence for the importance of type frequency of PSCs in nonword reading is scarce (cf. Andrews and Scarratt, 1998), and thus more direct evidence for the importance of this property is needed.

As described in Chapter 4 (Section 4.1.2), all the computational models compared in the current PhD project are sensitive to type frequency of PSCs. In the Dual-Route Cascaded (DRC) model (Coltheart et al., 2001), this property is categorical, as only grapheme-sized PSCs with the highest type frequency are used. In the Connectionist Dual Process model (CDP++, Perry et al., 2010) and the connectionist model used in simulation 1 of Plaut et al. (1996) (Psim1), type frequency of PSCs has a more graded influence, as the likelihood of the model utilising a given PSC increases with the number of words in the training set containing this PSC. However, other influences, such as token frequency of the words in the training set, are also at play, making it difficult to predict precisely when the PSCs with the higher type frequencies would be utilised by these models. In the versions of the Weighted Segments Pronunciation (WSP) model considered thus far, type frequency has been a part of the competition criterion, in which case its influence is combined with the influence of consistency of PSCs (the likelihood of a given pronunciation to be used by the model is the product of type frequency and consistency of each PSC segment of a letter string). Three versions of the WSP model were included in the current study: two versions where the competition of the pronunciation options is based on the product of type frequency and consistency of PSCs, one optimised for the WSP model's vocabulary (WSP-type) and one for a nonword naming data set by Treiman and colleagues (2003) (WSP-type-T, see Chapter 3,

Section 3.2.2 for details about the data set). The latter version of the model was included due to its relatively strong performance in all the data sets it was tested on in Chapter 3 (Section 2.4.3). A version of the model with competition based on the product of token frequency and consistency of PSCs, optimised for the WSP model's vocabulary (WSP-token) was also included, for comparison. As the WSP-token is based on summed token frequencies, it should be also sensitive to type frequencies to some extent, given that the type and summed token frequency measures are highly correlated.

5.1.2 The current study

The study reported in Chapter 4 investigated the role of token frequency in nonword processing. This study provided some evidence for a small effect of token frequency in pronunciation assignment to nonwords. Given this finding, the current study set out to investigate the role of type frequency in nonword processing, while controlling for the effect of token frequency and consistency of PSCs. A group of participants named and then rated nonwords that were matched in token frequency and consistency but differed in type frequency of the existing words they resembled. This is an important extension to previous studies, in which the two frequency measures have not been controlled sufficiently. Sensitivity to the influence of type frequency was measured as the incidence of base word congruent responses, that is, irregular vowel pronunciations for the nonwords (e.g., pronouncing the vowel in a nonword *donth* as it is pronounced in its base word *month*). Both nonword naming responses and acceptability ratings to base word congruent pronunciations were collected.

Furthermore, output from computational models of reading were compared to the pronunciations from human participants. All the models considered in the current study, apart from the DRC model, should exhibit graded influence of type frequency in nonword naming, that is, the incidence of base word congruent pronunciations should be higher for items with several base words (Irregular-Many) than for items with a single base word (Irregular-Single).

Finally, two additional analyses were carried out to assess potential other influences in the main results of the current study. Firstly, the properties of the vowel segments within the experimental nonwords were compared between the Irregular-Single and Irregular-Many items, as the vowel segment properties were found to be associated with the type of responses given to the nonwords in Chapter 4. Secondly, the results from the current study were

compared to the findings from the experiment reported in Chapter 4, namely, potential list context effects were investigated for items that occurred in both experiments. With relatively high token frequency of items and more word body neighbours for some of the items (items with higher type frequency in the current study), participants might become more aware of the word body segments of the stimuli and thus utilise word body sized PSCs in their responses more than participants in the previous experiment (Chapter 4) may have done, where all the critical stimuli had a type frequency of 1 and some of the items had a low token frequency. Finally, the effects of type and token frequency were compared.

The hypotheses of the current study were as follows.

1. The incidence of base word congruent pronunciations (irregular vowel pronunciations) assigned to nonwords with high type frequency will be higher than that to items with lower type frequency. Irregularly pronounced nonwords with high type frequency will also receive higher acceptability ratings than irregularly pronounced nonwords with lower type frequency.

2. Computational models with graded sensitivity to type frequency will be a closer match to the human data than models without this property.

Additionally, investigation on the potential influence of the nonwords' vowel segment properties and the influence of list context on nonword naming were considered, and the effects of type and token frequency were compared.

5.2 Methods

5.2.1 Participants

Participants were undergraduates in Psychological Science at the University of Bristol. They completed the study online as a course requirement. Inclusion criteria for the study were that participants were native speakers of British English with normal or corrected-to-normal vision, no diagnosed or experienced reading difficulties and who had not participated in the study reported in Chapter 4. After removing four participants due to corrupt audio files, the final sample size was 55 (13 males), with a mean age of 19.8 years (SD = 2.9), ranging from 18 to 30.

Ethics approval for the current study was granted by the School of Psychological Science Research Ethics Committee in University of Bristol (ethics approval code: 0229).

5.2.2 Materials

5.2.2.1 Naming task. The nonwords in the naming task consisted of items with word bodies that are always pronounced irregularly in existing words (base words). Two groups of items were created: nonwords with word bodies that occur in a single existing word (e.g., nonwords *donth* and *stonth* based on a word *month*), referred to as Irregular-Single items, and nonwords with word bodies that occur in several words (e.g., *zalk* and *glalk* based on *walk, talk, chalk,* etc.), referred to as Irregular-Many items. Two nonwords were created for each base word. The onsets for each nonword did not contain letters from the onsets of its base words. The resulting nonwords were not homophones to existing words when the vowel pronunciation of these items was regular or irregular. Whenever possible, the nonwords did not have orthographic neighbours other than their base words[35]. Three word bodies in the Irregular-Many items were not completely consistent, but all the words going against the irregular pronunciation of these items had low token frequency (maximum frequency of these items was 2.97 (range 1.30-2.97), except for *coup,* which was 3.59, but shared the critical, irregular vowel pronunciation with the rest of its item group). Thus, the exceptions to consistency were likely not very influential.

Both Irregular-Many and Irregular-Single item groups consisted of 30 nonwords, based on 15 word bodies. The mean token frequency of the item groups was matched as closely as possible, while the type frequency (i.e., the number of words in which the given word body occurs in) was naturally higher in the Irregular-Many group than in the Irregular-Single group. The token frequency of each Irregular-Many item was quantified as the highest token frequency amongst the base words of the Irregular-Many item. The Irregular-Single items were selected from the experimental items in the experiment reported in Chapter 4. This selection process was predominantly based on the token frequency of these items, as finding equally high token frequencies with those of the Irregular-Many items was a challenge. However, two other considerations regarding the Irregular-Single items were made to ensure fair comparison between the Irregular-Many and Irregular-Single items. Firstly, it was ensured that the base words for Irregular-Single items were all relatively well known by the student population recruited as participants in the previous experiment in Chapter 4 (the sample of the current experiment was drawn from the same student population). Secondly, it was ensured that at least half of the Irregular-Single items had received a reasonably high proportion of irregular pronunciations in the previous experiment (Chapter 4).

[35] *east* and *oast* were neighbours for a nonword *yast,* in addition to the original, base word neighbours

After the initial stimuli construction, two additional analyses were performed to investigate the token frequencies of the nonword items (i.e., the token frequencies of the base words of these nonwords) and the nonwords' vowel segment properties. These investigations were important to ensure the final stimuli set was optimal for the current purposes, as these aspects of the stimuli were not considered during the original stimuli construction due to the limited number of items that satisfied other criteria for the stimuli (e.g., consistent and irregular word bodies).

Comparison of the token frequencies of the Irregular-Many and Irregular-Single items revealed that the Irregular-Many items had higher mean token frequency than Irregular-Single items, and this difference was statistically reliable ($t(56.33) = 2.31$, $p = 0.01$, 1-tailed). Because matching token frequency between the critical item groups is important in the current study, the initial, full set of items was reduced to a subset with more comparable token frequencies. This subset of 26 nonwords in each item group was achieved by removing four nonwords from the Irregular-Many group with the highest token frequency (nonwords *bralf, plalf, snast* and *yast*) and four nonwords from the Irregular-Single group with the lowest token frequency (nonwords *bealm, phealm, hauge* and *snauge*). The difference in token frequency between Irregular-Many and Irregular-Single items was not reliable for this subset ($t(47.5) = 0.76$, $p = 0.23$). Thus, all subsequent analyses were conducted with this subset of items. However, the initial, full set of items was presented to the participants, to have the maximum number of experimental items available (see Appendix 8 for the full set of stimuli and Appendix 9 for the analyses performed with the initial, full set of nonwords).

In the final sample of nonwords (n = 26 in each item group), the token frequencies, quantified as the maximum Zipf value (Van Heuven et al., 2014), were on average (4.77, $SD = 0.53$)[36] for the Irregular-Many items and 4.65 ($SD = 0.67$) for the Irregular-Single items. The mean type frequency (i.e., number of base words) for the Irregular-Many items was 4.92 ($SD = 2.81$), whereas the mean type frequency for the Irregular-Single items was 1 ($SD = 0$). Finally, based on the data reported in Chapter 4, the mean proportion of base words for the Irregular-Single items that were known by the participants was .92 ($SD = 0.08$) and the

[36] Quantifying the Irregular-Many items' token frequency based on means for each word body group yields a significantly lower mean token frequency value (3.52). However, for the purposes of matching the mean token frequency of the base words between the Irregular-Many and Irregular-Single items, I deemed it fairer to calculate the mean frequency for the Irregular-Many items based on the maximum token frequencies of each word body group. This was to avoid underestimating the token frequencies of the Irregular-Many items, and thus ensuring that any potential differences between the two groups of items would be due to differences in type frequency, rather than a higher token frequency in the Irregular-Many items compared to Irregular-Single items.

proportion of irregular pronunciation assigned to the Irregular-Single items was on average .29 (*SD* = 0.29). As such, it is highly likely that most participants in the current study know the base words for the Irregular-Single items. Furthermore, based on the word prevalence measures from Brysbaert et al. (2019), that is, the proportion of participants that correctly identified a given word as a word, the minimum prevalence value for the two most prevalent base words for each Irregular-Many item was .98. As such, at least two base words were very likely known by the participants in the current study, thus ensuring that the Irregular-Many items were indeed different from the Irregular-Single items for each participant. This is why vocabulary knowledge of the base words was not tested in this study.

Due to the finding that frequency measures of the vowel segment are associated with the proportion of regular pronunciations of the nonwords in the previous experiment (Chapter 4), the properties of the vowel segments in the experimental stimuli of the current study were also compared between Irregular-Many and Irregular-Single items, with independent samples t-tests (1-tailed). The mean type frequency of the vowel segments was higher in the Irregular-Many items (*M* = 255.7, *SD* = 151.56) compared to the Irregular-Single items (*M* = 124.46, *SD* = 131.36), and this difference was statistically significant ($t(49) = 3.34$, $p < .001$). The mean summed token frequency of the Irregular-Many items (*M* = 620.64, *SD* = 334.64) was higher than that of the Irregular-Single items (*M* = 338.46, *SD* = 305.33), also a statistically reliable difference ($t(49.59) = 3.18$, $p = .001$). Finally, the mean maximum token frequency of the Irregular-Many items (*M* = 6.84, *SD* = 0.79) was also higher than that of the Irregular-Single items (*M* = 6.37, *SD* = 1.13), and this difference was statistically reliable ($t(44.77) = 1.77$, $p = .04$). Thus, the Irregular-Many items have significantly more reliable vowel segments, which might reduce the likelihood of irregular pronunciations assigned to these items. As such, the vowel properties of the Irregular-Many items work against my hypothesis, according to which the proportion of irregular pronunciations for the Irregular-Many items would be higher than that for Irregular-Single items.

In addition to the original 60 nonwords (30 Irregular-Many and 30 Irregular-Single items), the participants also named 202 filler nonwords. The results of the naming responses to some of these filler items (including items with C or G onset) will be reported in Chapter 6.

5.2.2.2 Rating task. The Irregular-Single and Irregular-Many items were also included in the rating task, where they were presented along with a regular or irregular pronunciation assigned to them (e.g., nonword *donth* pronounced irregularly as /dVnT/ or regularly as

/dQnT/). The Irregular-Single and Irregular-Many items were presented intermixed with 132 filler nonwords (including 20 C or G onset nonwords, 10 error nonwords and 10 odd nonwords, the rating responses to which will be reported in Chapter 6). See Appendix 8, Tables 8C and 8D for the full set of stimuli.

5.2.3 Procedure

Each participant completed a nonword naming task, followed by a nonword rating task. The procedure for these tasks was identical to that in the previous experiment reported in Chapter 4 (Section 4.2.3). As in this previous experiment, the nonwords sharing a word body were presented in a different block in the naming task, and each nonword was presented twice in the rating task, paired with a regular pronunciation in one block and irregular pronunciation in the other block.

5.2.4 Data processing

Pre-processing of the data and analyses were conducted using R 4.0.3 (R Core Team, 2020). Data processing in the current study was very similar to that in the study reported in Chapter 4.

5.2.4.1 Naming data. Failed audio recording resulted in loss of 1.8% of the nonword naming responses. The remaining responses were transcribed independently by two transcribers, who received training for the task, but were naive to the critical manipulations of the study. Any discrepancies in the transcriptions were processed again by the two transcribers. The transcription from the slightly more experienced transcriber were followed for the remaining 4% of items for which consensus could not be reached.

For investigations of the role of type frequency in nonword naming, the naming responses were categorised as regular or irregular (based on the vowel pronunciation) and proportions of irregular responses for each item group were calculated for each participant. For comparisons of human data with the output from computational models, a human modal response was extracted from the naming data as the most popular human naming response that was either regular or irregular (there were three items for which ‘other’ was the modal response, in which case the second most popular response (regular or irregular) was used). The final sample size for the models depended on the number of regular and irregular vowel pronunciations each model produced; these are depicted in Table 5.2.

5.2.4.2 Rating data. The labels of the rating scale were re-coded as follows: 1 = VERY BAD, 2 = BAD, 3 = PROBABLY NOT OK, 4 = PROBABLY OK, 5 = GOOD and 6 = VERY GOOD. Mean ratings from each participant were then calculated, for each item group (e.g., irregularly pronounced Irregular-Many and Irregular-Single items).

5.2.4.3 Statistical power. Sensitivity power analyses were computed using GPower (Faul, Erdfelder, Lang & Buchner, 2007) for each hypothesis test and the resulting minimum, reliably detectable effect sizes for each analysis will be reported along with the observed effect sizes from the analyses.

5.2.4.4 Transcription of output from computational models. The transcription of the output from computational models was identical to that reported in Chapter 4 (Section 4.2.4.6).

5.3 Results

5.3.1 Effects of type frequency in nonword processing

I expected higher proportion of base word congruent responses to items with higher type frequency (Irregular-Many items) compared to items with low type frequency (Irregular-Single items). I also predicted higher acceptability ratings for irregularly pronounced Irregular-Many items compared to Irregular-Single items. Due to these directional hypotheses, the comparisons are conducted as one-tailed tests. Both comparisons were conducted as by-subjects analyses.

5.3.1.1 Naming responses. The mean proportion of irregular vowel pronunciations for Irregular-Many items were compared to those of Irregular-Single items with a paired samples t-test. The Irregular-Many items were pronounced irregularly more often (M = .44, SD = .12) than the Irregular-Single items (M = .30, SD = .11), and this difference was statistically reliable: $t(54) = 7.96$, $p <$ .001, $dz = 1.07$[37].

To compare the proportions of the irregular responses for the two item groups in the current study with those reported in previous studies (i.e., Andrews & Scarratt, 1998, Exp. 2), by-items means were also computed. The proportion of irregular responses was .43 for the Irregular-Many items and .31 for the Irregular-Single items in the current data set. By contrast, Andrews and Scarratt report a higher incidence of irregular responses, namely, .65

[37] Sensitivity analysis with an alpha level of .05, power of .8 and sample size of 55 yielded a minimum, reliably detectable effect size of dz = 0.34

for Irregular-Many items and .40 for Irregular-Single items, when all the named items were nonwords (however, very similar proportions were obtained for the same items when they were named intermixed with words). This comparison will be considered further in the Discussion.

5.3.1.2 Rating responses. The mean acceptability ratings for irregularly pronounced Irregular-Many items and Irregular-Single items were compared with a paired samples t-test. The Irregular-Many items received higher acceptability ratings (M = 5.37, SD = 0.42) than the Irregular-Single items (M = 4.63, SD = 0.46), and this difference was confirmed statistically: $t(54) = 13.21$, $p <. 001$, $dz = 1.78$[38].

The same analyses of the naming and rating responses were also performed with the original, full set of items (30 Irregular-Many items and 30 Irregular-Single items). These analyses yielded comparable results (see Appendix 9).

5.3.1.3 Vowel segment properties. In Chapter 4, it was found that the proportion of regular and irregular pronunciations assigned to nonwords was associated with frequency measures of the vowel segments within these nonwords, that is, the higher the frequency and number of words in which a regular vowel-sized PSC occurred, the higher the likelihood of this vowel being pronounced regularly in nonwords and the lower the likelihood of this vowel being pronounced irregularly in nonwords. It was therefore investigated whether a similar relationship existed in the nonword set used in the current study. The proportion of regular and irregular responses (based on the vowel pronunciation) assigned to each nonword were correlated with the statistical properties of the vowel-sized PSC that corresponded to the orthographic vowel segment within each nonword. The results of the correlations are depicted in Table 5.1.

[38] Sensitivity analysis with an alpha level of .05, power of .8 and sample size of 55 yielded a minimum, reliably detectable effect size of dz = 0.34

Table 5.1

Correlations between proportions of regular and irregular responses and statistical properties of the vowel segment of Irregular-Single and Irregular-Many items

	Irregular-Single items			Irregular-Many items		
Vowel property	*reg. prop*	*irreg. prop*	*Mean*	*reg. prop*	*irreg. prop*	*Mean*
Consistency	.76 ***	-.56 **	.71	-.24	.19	.74
Type Freq	.74 ***	-.6 **	124.46	.5 **	-.52 **	255.77
Sum Token Freq	.78 ***	-.66 ***	338.46	.54 **	-.56 **	620.64
Max Token Freq	.69 ***	-.64 ***	6.37	.65 ***	-.61 ***	6.84

Note. reg. prop = proportion of regular vowel pronunciations, irreg. prop = proportion of irregular vowel pronunciations. Mean columns depict mean values for each property of the vowel segments for Irregular-Single and Irregular-Many items. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05.

Each property of the vowel segments was positively correlated with the proportion of regular pronunciations and negatively correlated with the proportion of irregular pronunciations assigned to the nonwords. This pattern was true for both Irregular-Single and Irregular-Many items, although the consistency of the vowel segments was not reliably associated with the type of responses given for the Irregular-Many items. Summed token frequency had the strongest correlation to the proportion of regular and irregular pronunciations assigned to the Irregular-Single items, whereas the maximum token frequency was the most strongly associated property with the proportion of regular and irregular pronunciations assigned to the Irregular-Many items. Thus, a similar relationship between the properties of the vowel segments of nonwords and the proportion of different types of pronunciations assigned to them found in Chapter 4 was also found in the current data set. When a vowel is pronounced regularly in highly frequent words, the same vowel is more likely to be pronounced regularly in nonwords than when the vowel is pronounced regularly in less frequent words. Table 5.1 also shows the mean values for the properties of the vowel segments, demonstrating that the difference between Irregular-Single and Irregular-Many items goes against the hypothesis regarding the role of type frequency in nonword reading: the properties of the vowel segments should decrease the likelihood of irregular pronunciations assigned to the Irregular-Many items compared to the Irregular-Single items. As such, the higher incidence of irregular pronunciations assigned to the Irregular-Many items than Irregular-Single items cannot be explained by a difference in the vowel segment properties.

5.3.1.4 List context effects. There were 28 nonwords in the Irregular-Single item group that were also named in the previous experiment, reported in Chapter 4. Using this subset of items that overlapped between the previous study and the current study, I tested whether the incidence of irregular responses for these items was higher in the current study, potentially indicating list context effects as the source of the difference. The incidence of irregular pronunciations assigned to these 28 items in each experiment were compared with an independent samples t-test (1-tailed) for both by-items and by-subjects means. In the by-items data, the mean proportion of irregular pronunciations was slightly higher in the current study (M = 0.30, SD = 0.30) than in the previous study (M = 0.28, SD = 0.28). However, this difference was not statistically reliable (t(53.63) = 0.26, p = .40). Considering the by-subjects means, the proportion of irregular pronunciations was also numerically higher in the current study (M = 0.30, SD = 0.10) compared to the previous study (M = 0.29, SD = 0.10), but this difference was not statistically reliable (t(114.88) = 0.63, p = .26). Thus, based on this subset of items, the nonword naming responses were not based on word body sized PSCs more often in the current study than in the previous study reported in Chapter 4.

5.3.1.5 Effect of type frequency compared to token frequency. To inspect the relative importance of type and token frequencies of PSCs in nonword processing, the effects of type frequency (current study) and token frequency (Chapter 4) were compared. The nonword naming and rating responses from the current study were compared to the nonword naming responses from the Naming-Rating group and nonword rating responses from the Rating-Only group from the study reported in Chapter 4. Cohen's dz was used as an estimate of effect size, which expresses the standardized mean difference between two measures (e.g., the difference in proportion of irregular vowel pronunciations for Irregular-Many and Irregular-Single items). In the naming responses, the effect of type frequency was 1.07 (Cohen's dz), and the effect of token frequency was 0.24 (Cohen's dz, but this analysis was underpowered). In the rating responses, the effect of type frequency was 1.78 (Cohen's dz) and the effect of token frequency was 0.38 (Cohen's dz). It thus appears that the type frequency has a larger effect in both nonword naming and rating. This will be considered further in the discussion (Section 5.4).

5.3.2 Comparison of computational models in reading irregular nonwords

The human modal responses to each experimental item were extracted and compared to the output from the computational models in several ways. Firstly, a Chi-squared test for

independence was carried out to test whether there was an association between the type of response given (regular or irregular vowel pronunciation) and the type of item (Irregular-Single and Irregular-Many) in the human modal responses and in the output from each computational model. These results are summarised in Table 5.2.

Table 5.2

Comparison of the proportion of irregular pronunciations assigned to Irregular-Single and Irregular-Many items by humans and computational models

Source	Item type			Chi-squared test for Irregular-Single vs Irregular-Many				
	Irregular-Single	*Irregular-Many*	*Total*	X^2	*df*	*n*	*p-value*	*Yate's correction*
Humans	.42	.54	.48	0.69	1	52	0.41	no
DRC	.00	.15	.08	1.95	1	48	0.16	yes
CDP++	.28	.76	.52	11.54	1	50	< .001	no
Psim1	.30	.59	.45	3.58	1	42	0.06	no
WSP-type	.88	.88	.88	0.00	1	52	1.00	yes
WSP-token	.92	.92	.92	0.00	1	52	1.00	yes
WSP-type-T	.42	.38	.40	0.08	1	52	0.78	no

The human data and output from most of the computational models did not show reliably more irregular responses for the Irregular-Many items than the Irregular-Single items, except for the CDP++ model and (marginally) Psim1 model. Numerically, the absolute values of the proportions of irregular pronunciations for Irregular-Single and Irregular-Many items in the human data were most closely matched by the output from the Psim1, WSP-type-T and CDP++. The DRC model produced noticeably lower proportions and the WSP-type and WSP-token models produced clearly higher proportions of irregular pronunciations to the nonwords than humans did. Like in the human data, the Irregular-Many items received a higher proportion of irregular pronunciations than the Irregular-Single items in the output from DRC[39], CDP++ and Psim1 models. However, the three versions of the WSP model did not produce this pattern, even though the WSP-type and WSP-type-T are sensitive to type frequency of PSCs. These findings will be considered further in the Discussion.

Next, to compare the types of responses given to each nonword between the model and the human data, I calculated the proportion of items for which each model matched the type of

[39] These were due to multi-letter graphemes *igh* and a split vowel *y_e* (in word body *yme*)

human modal response (regular or irregular, based on vowel pronunciation), across all items and for items with irregular or regular human modal response separately (see Table 5.3).

Table 5.3

Proportion of matching pronunciation types between human modal responses and output from computational models for nonwords

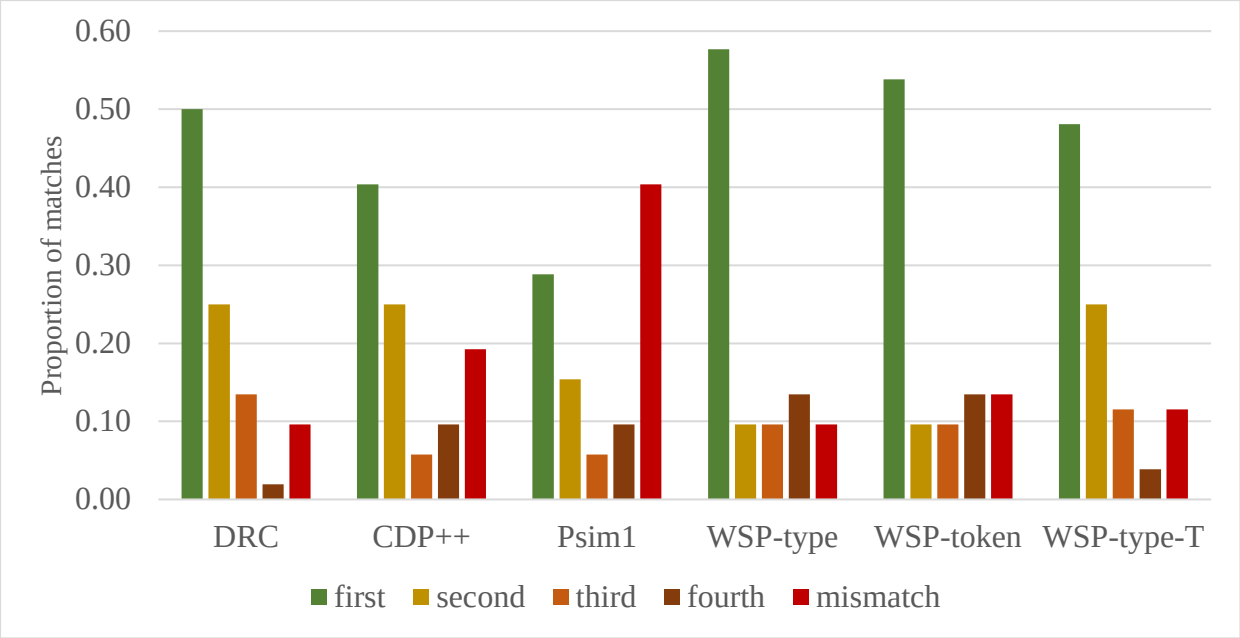
	DRC	CDP++	Psim1	WSP-type	WSP-token	WSP-type-T
total (n = 52)	.60	.54	.48	.60	.56	.54
regular (n = 27)	1.00	.52	.56	.22	.15	.63
irregular (n = 25)	.16	.56	.40	1.00	1.00	.44

As seen in Table 5.3, the highest proportion of matches to the human modal responses was produced by the DRC and the WSP-type models, followed by WSP-token, CDP++ and WSP-type-T models, and finally Psim1. The high proportion of overall matches of the WSP's type and token versions stemmed from a high incidence of irregular pronunciations – all items pronounced irregularly by the participants were also pronounced irregularly by these versions of the WSP model. By contrast, the overall performance of the DRC model was mostly driven by matching all the items the participants pronounced regularly. Finally, the CDP++, Psim1 and WSP-type-T had more balanced performance across the items receiving a regular or irregular human modal response.

Finally, for a more detailed comparison of the human and model naming responses, the human naming responses were arranged according to the frequency of different responses (first, second, third and fourth or lower most common responses). The proportion of items each model matched were then extracted (considering the whole nonword pronunciation, not only the vowel pronunciation), as well as the proportion of items for which the model output was not produced by any of the participants. Figure 5.1 summarises the performance of each model.

Figure 5.1

Proportions of human-model matches to nonwords arranged by human response frequency



Note. Proportion of matches was calculated for 52 nonwords that had at least one valid response in the human data.

The three versions of the WSP model and the DRC model produced similar proportions of matches to the human modal response – from .48 (WSP-type-T) to .58 (WSP-type). These models also produced comparable, low proportions of mismatches – from .10 (DRC and WSP-type) to .12 (WSP-token and WSP-type-T). the CDP++ model had somewhat poorer performance, with proportion of human modal responses at .40 and mismatches at .19. Finally, the Psim1 had the weakest performance in both the proportion of human modal responses (.29) and the proportion of mismatches (.40).

There were different types of mismatches produced by each model. For the WSP-type and WSP-token versions, all of the five (WSP-type) and seven (WSP-token) mismatches were a result of irregular vowel pronunciations (e.g., *tilst* as /t2lst/ or *gnolk* as /n5k/). In addition to irregular vowel pronunciations, the WSP-type-T also produced a handful of other types of mismatches (e.g., onset in *gealth* pronounced as /_/). For the DRC model, four of the five mismatches were due to an odd vowel pronunciation (e.g., *fousse* as /f6si/ or *meird* as /m1rd/). For the CDP++, half of the 10 mismatches were because of an irregular vowel pronunciation (e.g., *nalt* as /n$lt/) and the remaining mismatches were a result of odd or

phonotactically illegal onset or coda (e.g., *psorld* as /p$ld/ or *zalk* as /z$kk/). For the Psim1 model, 12 of the 21 mismatches were because of odd or missing onset or coda (e.g., *phoup* as /up/ or *shoung* as /SV/), along with eight odd or irregular vowel pronunciations (e.g., *gleart* as /gl$t/ or *nalm* as /nQm/). See Appendix 10 for full list of model output and mismatches produced by each model.

5.4 Discussion

The current study investigated the role of type frequency of PSCs in nonword processing. Naming responses to nonwords and acceptability ratings for pronunciations assigned to nonwords were collected for two types of items – nonwords that overlapped with a single, irregularly pronounced base word (Irregular-Single items) and nonwords that overlapped with several, irregularly pronounced base words (Irregular-Many items).

The first hypothesis of the current study was that the Irregular-Many items would receive more base word congruent naming responses (i.e., irregular vowel pronunciations) than the Irregular-Single items and that irregularly pronounced Irregular-Many items would be rated as more acceptable than irregularly pronounced Irregular-Single items. Both naming and rating data clearly indicated that this was the case – the Irregular-Many items received more base word congruent responses than the Irregular-Single items and the Irregular-Many items were rated as more acceptable when they received a base word congruent pronunciation than the Irregular-Single items. The same conclusion can be drawn from the analyses with the original, full set of items for both naming and rating data (Appendix 9). However, the reduced set reported in the main text benefits from more comparable mean token frequencies of the Irregular-Many and Irregular-Single items, thus ruling out the possibility that the token frequency drove the differences in the naming and rating responses.

Comparing the mean proportions of the base word congruent responses for the Irregular-Many and Irregular-Single items to those reported in previous studies (i.e., Andrews & Scarratt, 1998, Exp. 2), the proportions for both item groups were lower in the present study compared to Andrews and Scarratt's results. A similar observation was made in Chapter 4, and one potential explanation of this difference may be the onset complexity of the nonwords. Whereas Andrews and Scarratt used nonwords with a single letter onset, 62% of the experimental items in the present study contained two-letter onsets. These more complex onsets were distributed evenly between the Irregular-Many and Irregular-Single items. As suggested in the Discussion of Chapter 4, irregular word bodies may be more difficult to

recognise after a complex onset compared to a simple one and thus using word body analogies in nonword naming may become less likely, an idea left for future research to test. Additionally, most of the nonwords in Andrews and Scarratt's study had orthographic neighbours – the words they were based on (unless the base word had a complex onset, such as *chalk* or *sleigh*). A potential reason for more base word congruent pronunciations for items with orthographic neighbours is that when a nonword differs from an existing word by only one letter, the existing words (or base words) may be more salient and their pronunciations more influential in assigning pronunciations to the nonwords they resemble. The current study does not answer the question of whether the mechanism behind base word congruent pronunciations is best characterised as a lexical or word body analogy (i.e., recognising the base word of a nonword and imitating the pronunciation of the base word in nonword naming, particularly regarding the overlapping segment – the word body) or as easier access to PSCs of different sizes from memory (i.e., easier retrieval of PSCs with high consistency, frequency or other statistical properties that would advantage retrieval of pronunciations for word body sized segments). Nevertheless, the current study demonstrates that the incidence of base word congruent responses for nonwords is considerable even when these items do not have many or any orthographic neighbours.

The second hypothesis of the current study was that computational models in which the influence of type frequency of PSCs is graded would simulate human naming responses more accurately than models without this property. Considering the different criteria for model performance, there was some evidence supporting this hypothesis, but equally some evidence against it. Absolute values of proportions of irregular pronunciations to the critical items in the human data were best matched by models with graded influence of type frequency (Psim1, WSP-type-T and CDP++), but the pattern of these proportions was most accurately reflected in the output from the DRC model, a model without this property. Proportions of matching pronunciation types to the human modal responses, considering the critical vowel pronunciation only, were best simulated by the WSP-type and the DRC, each with a particular strength in matching either regular or irregular vowel pronunciations. Finally, the highest number of matches to the human modal responses, considering the whole nonword pronunciation, were produced by the WSP-type and WSP-token, yet, the WSP-type's and the DRC's output matched the highest number of items altogether (i.e., the models produced the smallest number of pronunciations not produced by a single participant). Taken together, although graded influence of type frequency appears to be beneficial in simulating the current

human naming data, it does not appear to be absolutely necessary, as a model without this property (the DRC) had a consistently strong performance on this data set. It should also be noted that the performance of each model considered in these comparisons remained at a modest level, for instance, as the highest percentage of matching pronunciations to the human modal responses was as best 58% for whole nonwords and 60% for matching vowel pronunciation types.

The comparisons of the output from the computational models and the human modal responses also revealed that although the CDP++ and Psim1 models showed the same general pattern of proportions of irregular pronunciations assigned to the Irregular-Single and Irregular-Many items, the difference was more pronounced in the output from these models compared to the human data, as seen in the Chi-squared test outcomes. Namely, the association between pronunciation type (irregular or regular) and item type (Irregular-Single or Irregular-Many items) was statistically reliable for the CDP++ and marginally significant for the Psim1 model's output, unlike for the human data. Given the clear results from the human naming and rating data regarding the role of type frequency in nonword processing when mean proportions of base word congruent responses were considered (Section 5.3.1), it is somewhat surprising that the same outcome was not found in the human data using the Chi-squared test. However, this can be understood in terms of differences in statistical power between the two analyses. Firstly, the Chi-squared test is non-parametric and as such has less statistical power to detect relationships between variables. Secondly, the comparison of proportion of irregular responses to the two item types were by-items comparisons in the Chi-squared tests (and based on human-modal responses only), whereas the comparisons reported in Section 5.3.1 were based on by-subjects means, and as such, the larger sample size in the by-subjects analyses corresponds to higher statistical power.

Another surprising finding in the model output comparisons was that none of the versions of the WSP model produced the expected numeric trend (higher proportion of irregular pronunciations for Irregular-Many than Irregular-Single items). For WSP-type and WSP-token versions of the model, this is likely because the proportions of irregular pronunciations in the model's output were very high, regardless of the item type. Because the word bodies of the experimental items were all perfectly consistent but varied in type frequency, and because the small segment parsing style is considerably disadvantaged in these versions of the model, it appears that the consistency of the word body sized PSCs was too influential in the WSP model's pronunciations. A similar issue was seen in the model's performance in the

experiment investigating the role of token frequency in nonword processing (Chapter 4). As such, the current study further confirms that these versions of the WSP model suffer from too great an influence of the consistency of PSCs, particularly in the larger segment parsing styles. This issue will be considered in Chapter 7.

However, this explanation is not sufficient for the performance of the WSP-type-T, as the proportions of irregular pronunciations produced by this version of the model were not at ceiling. It is therefore unclear why the WSP-type-T did not produce more irregular pronunciation to the Irregular-Many items compared to the Irregular-Single items. There appear to be several factors that contribute to this pattern of output from the WSP-type-T. Most importantly, the strength of the onset and the coda segments in the competition between the three parsing styles can have a considerable influence on the final output from the model. For instance, the item *psask* was pronounced regularly by the WSP-type-T, mostly because the strength of the word body parsing style relied on a relatively uncommon onset *ps* (segment strength = 0.10) and a moderately frequent and entirely consistent word body *ask* (segment strength = 0.85). By contrast, the small segment parsing style included the onset *ps* (segment strength = 0.10), a highly frequent vowel *a* (segment strength = 1.74) and a considerably frequent and entirely consistent coda *sk* (segment strength = 1.26). The vowel and the coda in the small segment parsing style thus resulted in a mean strength that could not be exceeded even with a slightly higher weight applied to the mean strength of the word body parsing style. To demonstrate that this difference is mostly related to the weak onset and strong coda, rather than a considerably strong vowel segment, the model's output is based on the word body parsing style when this word body is combined with any single-letter onset. Thus, there were instances where the small segment parsing style won based on onset or coda segments, even though each parsing style produced the same pronunciations for these segments. This explanation is likely relevant also for the similar findings in Chapter 4 (section 4.3.3), where WSP-token-T did not show sensitivity to token frequency of nonwords. The issue of the strength of the onset and coda segments will also be considered in Chapter 7.

Moving on to the two additional analyses, it was confirmed that the same type of relationship between vowel segment properties of the nonwords and the type of pronunciations assigned to the nonwords found in the previous study (Chapter 4) was also present in the current study – vowel segments pronounced regularly in several or highly frequent words were more likely pronounced regularly when these vowels occurred in nonwords. Importantly, the frequency measures of the vowel segments were higher in the Irregular-Many items than in the

Irregular-Single items, thus ruling out the possibility that the difference in the proportion of base word congruent pronunciations between Irregular-Many and Irregular-Single items could be explained by the vowel segment properties of these items.

Additionally, potential list context effects were investigated by comparing the naming responses to a subset of items that were used in both the current study and a previous study investigating the role of token frequency in nonword processing (Chapter 4). The comparison of interest here was whether the incidence of base word congruent pronunciations would be higher for this subset of items in the current study compared to the previous study. Both by-items and by-subjects analyses pointed to the same conclusion – the incidence of base word congruent pronunciations for the items did not differ between the two studies. As such, it seems unlikely that the clear effect of type frequency found in the current study would be attributable to list context effects, such as salience of the base words in the current study leading to stronger reliance on word body sized reading approach, compared to those in the previous study.

Turning to the limitations of the current study, the clearest shortcoming for the empirical findings in the present study was that the rating responses were collected from the same participant group that had already named the same nonwords. As such, it is possible that the rating behaviour was influenced by the preceding exposure to the nonwords. Further consideration of the relationship of naming and rating responses will be covered in Chapter 6.

Finally, comparing the effect sizes from the study investigating the role of token frequency in nonword processing (Chapter 4) and the evidence for the role of type frequency in the current study, it was found that type frequency of PSCs appears to have a larger effect on nonword processing, both in nonword naming and rating responses. However, while the designs used to investigate the role of these two properties in nonword processing were identical, the strength of the experimental manipulations were likely different. I am not aware of a way to compare the strength of manipulation between these two measures. With this limitation, the evidence regarding the relative importance of type and token frequency in nonword processing should only be considered tentative.

5.4.1 Conclusion

In this chapter, I reported investigations of the role of type frequency of PSCs in nonword processing. Although some empirical evidence suggests that this property is influential in nonword reading, only a handful of studies provide compelling evidence for the role of type

frequency. Therefore, I aimed to discover whether skilled readers read aloud nonwords and rate acceptability of nonword pronunciations such that PSCs occurring in several words are favoured over PSCs occurring in a single word. Importantly, the mean token frequencies of the stimuli and consistency of the critical PSC segments in the nonwords were comparable in the current study, thus allowing stronger conclusions to be drawn regarding the role of type frequency alone. The current study provides clear evidence for the influence of type frequency of PSCs in nonword processing, both in patterns of nonword naming and rating responses. Several comparisons between the participants' nonword naming responses and output from computational models were also made. The computational models with graded influence of type frequency in their print-to-sound conversion (i.e., the connectionist models and the WSP model) were not clearly superior to the performance of the DRC model, which only considers grapheme sized PSCs with the highest type frequency. Although each model had their strengths in simulating particular aspects of the human naming responses, none of the models clearly outperformed the other models, and none of the models simulated the naming responses at a sufficiently high level of accuracy.

Chapter 6 : Evaluation of the Nonword rating method

6.1 Introduction

Nonword reading studies have traditionally relied on phonemic transcription to convert participants' verbal reading responses into a written form. These phonemic transcriptions are then used as the data for the subsequent analyses (e.g., Andrews & Scarratt, 1998; Pritchard et al., 2012; Mousikou et al., 2017). The inherent assumption in this approach is that these transcriptions reflect the print-to-sound correspondence (PSC) knowledge of skilled readers. However, this approach has two obvious sources of human error – pronunciation errors made by the participants, that is, mispronouncing what one intended to say (participants often name hundreds of nonwords (e.g., Pritchard et al., 2012; Mousikou et al., 2017)), and transcription errors made by the transcribers. Transcription of nonword naming responses is also time consuming and particularly prone to errors when transcribing large data sets. One approach for reducing the number of errors in phonemic transcriptions is having at least two individuals transcribe the verbal responses and comparing the similarity of the transcriptions produced. However, the inter-rater agreement for nonword transcriptions may remain relatively low, as exemplified by a recent study (De Simone et al., 2021, Experiment 1), where the agreement between two scorers of English nonword responses was only moderate (Cohen's kappa = 0.57).

As an alternative to transcribing nonword reading responses, a handful of studies have employed a multiple-choice method, where each nonword is presented with a choice of existing words and the participants' task is to circle the option containing the (typically vowel) pronunciation they would assign to the given nonword. The nonword processing responses obtained using this method have been compared to nonword reading responses in developing readers (Johnson, 1970, Pilot Study B) and adult skilled readers (Ryder & Pearson, 1980; Treiman, Kessler & Bick, 2003). Johnson (1970) found that the nonword naming and multiple-choice responses by second, fourth and sixth grade students differed from one another, however, the responses from the two tasks were more similar amongst students with higher reading skills and students on higher grades. Ryder and Pearson, on the other hand, did not find a difference between the naming method and the multiple-choice method, using a between-subjects design. Treiman et al. (2003, Exp. 2) also compared

nonword naming responses to multiple choice responses as a between-subjects design. The response options in the multiple-choice task were two words, associated with different vowel pronunciations, and 'neither' for trials where neither of the provided options corresponded to the pronunciation a participant would give to the nonword. While the main pattern of results from both the nonword naming group and the multiple-choice group were similar, the multiple-choice group gave more 'neither' responses compared to the proportion of responses that were classified as 'other' in the nonword naming group. Treiman and colleagues conclude that ensuring participant engagement in the task and removing a neutral response option would improve the multiple-choice method. In its current form, it was deemed less reliable than the traditional nonword naming method.

While the multiple-choice method may indeed serve as a way to avoid collecting and transcribing nonword naming responses, it might not provide the benefits that other alternatives to a nonword naming task may offer. Namely, choosing a pronunciation out of options involves choosing the most acceptable option, compared to the alternatives. As such, the responses from both nonword naming and the multiple-choice methods may only reflect the human modal responses, or each participant's 'first choice', although participants may find other pronunciations almost as plausible as their first choice. Another way to avoid the issues associated with nonword naming transcriptions is to obtain acceptability ratings for a pronunciation assigned to a nonword. The same nonword can be paired with several alternative pronunciations, as separate trials, thus providing assessments of each alternative pronunciation for a given nonword in isolation, rather than relative to other pronunciations. This approach, the nonword rating method, may reveal more about the PSC knowledge that skilled readers have, beyond a single naming response per item and beyond the best option out of alternatives.

Treiman and Zukowski (1988) used a nonword rating method where participants judged the acceptability of pronunciations assigned to nonwords on a scale from 1 (not a possible pronunciation) to 4 (possible pronunciation). The same items rated in this task (Experiment 2) were also named by a different group of participants (Experiment 1), thus allowing comparison of the two methods. The nonword stimuli in both experiments overlapped with exception words in antibody, body or only the vowel segment (e.g., for an exception word *friend,* a nonword with overlapping antibody was *frieth,* a nonword with overlapping word body was *chiend* and a nonword with an overlapping vowel was *chieth*). In the naming responses, there were more vowel pronunciations congruent with the exception words (i.e.,

analogy pronunciations) for the nonwords overlapping in word body than the other two types of nonwords, i.e., *chiend* was pronounced more often as /JEnd/ than *frieth* or *chieth* were pronounced as /frET/ or /JET/. Similarly, in the rating task, when the experimenter pronounced these items as an analogy to the exception words they overlapped with, the word body nonwords received reliably higher acceptability ratings than the antibody or vowel nonwords did. However, no alternative pronunciations were provided for the same nonword (as separate trials), such as pronunciations following the GPC-rules, which might have provided further insights into the PSC knowledge of participants in this study.

The nonword rating method has recently been used in evaluating computational models of reading (Gubian, et al. 2022). Gubian and colleagues compared two methods of model evaluation – a naming method and a rating method. In the naming method, participants' nonword naming responses were collected and phonemically transcribed to produce a nonword naming data set – Gubian and colleagues used the nonword naming data set from Mousikou et al. (2017). The naming responses were then compared to the output from computational models, so that the output of the models was considered acceptable if at least one human participant produced the same pronunciation and unacceptable if no human participant produced the same pronunciation. In the nonword rating method, Gubian et al. used a speech synthesiser to produce aural versions of the computational models' output for the Mousikou and colleague's nonwords. These pronunciations were then paired with the written form of the nonwords, and skilled readers rated the acceptability of these pronunciation-nonword pairs on a scale from 1 (very bad) to 6 (very good). Using this method, a model's output for a given nonword was acceptable if it received a median rating of at least 4 ('probably ok') and unacceptable if it received a median rating of 3 ('probably not ok') or lower.

Gubian and colleagues compared the performance of three computational models or algorithms that can name the disyllabic nonwords. To investigate potential false positives and false negatives of the naming method, Gubian and colleagues included all the pronunciations from these models/algorithms that matched a single participant's naming response in the Mousikou et al. data set or that was not matched by any participant's naming response. This was because a single matching pronunciation may turn out to be a naming or transcription error rather than a deliberate naming response (i.e., a false positive) and a pronunciation not produced by a limited sample of 41 participants may still be a generally accepted pronunciation amongst skilled readers (i.e., a false negative). This comparison revealed that

19% of the items could be classified as false positives, as these items were deemed unacceptable by the rating method but acceptable by the naming method. Notably, 58% of the items could be classified as false negatives, as these items were deemed acceptable by the rating method but unacceptable by the naming method. That is, the rating method can identify naming responses that are deemed acceptable by skilled readers, despite not being the naming responses the readers might produce themselves.

Another important finding by Gubian et al. was that the ratings of nonword pronunciations also converged with the naming data for human modal responses. These results, among other evaluations of the two methods reported by Gubian et al., suggest that the rating method is a feasible alternative to the naming method in evaluations of computational models.

Taken together, attempts to explore alternatives to the traditional nonword naming method have been scarce. The similarities in the pattern of results between the naming method and the alternative multiple choice or rating methods (Ryder & Pearson, 1980; Treiman & Zukowski, 1988) are promising and suggest that more work is needed to discover whether and under which circumstances the nonword naming method could be replaced with an alternative method.

6.2 Experiment 1

The aim of the present experiment was to provide a detailed comparison of the nonword naming and nonword rating methods. All the studies outlined above, except for Johnson (1970), have used between-subjects designs. Therefore, one cannot rule out the possibility of between-group differences in nonword naming and rating responses. A more stringent comparison of the two methods should include both types of responses from the same group of participants. Thus, in the present experiment, a group of participants read aloud nonwords and subsequently rated the acceptability of different ways of pronouncing these nonwords on a six-point scale, ranging from ‘very bad’ to ‘very good’. This procedure allowed the collection of data about the participants’ knowledge of PSCs beyond the responses they arrived at in the naming task – the same participant might deem several alternative pronunciations for the same item as acceptable. However, exposure to the same nonwords in two different tasks may also result in atypical or non-representative nonword processing responses, which was also addressed in the current experiment by including a between-subjects comparison of the rating responses: another group of participants only rated the different pronunciations paired with the nonwords, without naming them beforehand. As

such, comparing the rating responses from these two groups of participants allowed determination of whether the previous nonword naming task for one of the groups influenced their rating behaviour in the rating task.

The rating method was assessed in three different ways: 1) inspecting rating responses to items that no participant should deem acceptable, based on the PSCs the relevant experimental items contained (e.g., *dwal* pronounced as /jEsts/), these items were used in evaluating the specificity of the rating method; 2) comparing rating responses to nonwords for which previous studies have shown a clear pattern of naming responses (Treiman et al., 2007, Experiment 1); and 3) comparing naming and rating responses to nonwords that shared a word body with irregularly pronounced words, with varying token frequencies. These items were used for evaluating the sensitivity of the rating method. The groups of experimental stimuli and their purpose in the experiment are described in more detail in Section 6.3.2.

The pattern of rating responses was generally expected to converge with the pattern of naming responses. This convergence means high acceptability ratings for pronunciations produced by the majority of the participants (i.e., human modal responses) and low acceptability ratings for pronunciations no participant would produce or very few or no participant produced in previous studies. As a six-point rating scale was used, a pronunciation for a nonword was considered accepted by a participant if the pronunciation received a rating 'probably ok', 'good' or 'very good', and a pronunciation was considered rejected if it received a rating 'probably not ok', 'bad' or 'very bad'. Thus, if the rating method has high sensitivity, human modal responses should be accepted in the rating task and if the rating method has high specificity, implausible pronunciations (e.g., *dwal* pronounced as /jEsts/) should be rejected in the rating task.

The hypotheses for the present experiment are listed next, for each of the three item types used in the evaluation of the rating method.

Mean ratings for 10 Error items (e.g., *dwal* pronounced as /jEsts/) and 10 Odd items (e.g., *gloost* pronounced as /glEst/) in the rating task were expected to show the following pattern:

1) The mean ratings for the Error items and for the Odd items would be clearly in the 'unacceptable' range, i.e. significantly below 4 ('probably ok').
2) The mean ratings for the Odd items would be significantly higher than the ratings for the Error items.

The second set of hypotheses are based on naming responses from previous experiments. These are experiments by Treiman et al. (2007) and Treiman and Kessler (2019, Exp. 1), in which participants read aloud nonwords that either had an onset *c* or *g*, followed by the vowel *e* or *i* (critical items, e.g., C-critical: *cepth* or G-critical: *gipth*) or items that had an onset *c* or *g*, followed by vowels *a, o* or *u* (control items, e.g., C-control: *capth* or G-control: *gupth*). In the English writing system, onset *c* can receive a soft pronunciation /s/ (as in *cell)* after vowels *e* and *i* and *g* can receive a soft pronunciation /_/ (as in *gene*) after these vowels. However, the standard or hard pronunciation of *c* as /k/ (as in *cat*) and *g* as /g/ (as in *game*) occurs when these onsets are followed by vowels *a, o* or *u*. Based on the pattern of naming responses reported in previous studies, the following predictions were made about the relative and absolute values of the ratings for the C and G-initial items:

1) **Relative ratings.** For control items, hard pronunciations would be favoured over soft pronunciations (i.e., hard pronunciations would receive higher ratings than soft pronunciations). For C-critical items, soft pronunciations would be favoured over hard pronunciations. For G-critical items, hard pronunciations would be favoured over soft pronunciations.
2) **Absolute ratings.** Soft pronunciations for control items would be rated as unacceptable, with mean ratings below 4 (‘probably ok’). Soft pronunciations for critical items would be rated as acceptable, with mean ratings above 3 (‘probably not ok’).

Finally, the aim of the analyses of the naming and rating responses to the Irregular items with varied token frequency was to directly compare the two types of responses as a within-subjects design. The excepted pattern of results from the comparison of the naming and rating data for the irregular items was as follows:

1) Irregularly pronounced items in the naming task would be rated as more acceptable when they are paired with an irregular pronunciation in the rating task compared to when they are paired with a regular pronunciation.
2) Regularly pronounced items in the naming task would be rated as more acceptable when they are paired with a regular pronunciation in the rating task compared to when they are paired with an irregular pronunciation.

6.3 Experiment 1 Method

6.3.1 Participants

Two groups of participants were included in the current experiment – a group that first named the experimental items and then rated the acceptability of different pronunciations assigned to the same items (Naming-Rating group), and a group that only rated the acceptability of the pronunciations assigned to the items (Rating-Only group). The participant characteristics, stimuli and procedure for these two groups are described in detail in Chapter 4 (Section 4.2).

6.3.2 Materials

The materials of the current experiment were divided into three categories, each serving a distinct function in the evaluation of the rating method.

6.3.2.1 Error and Odd items. These items served a dual-purpose in the evaluation of the rating method. Firstly, participants' engagement in the rating task was assessed via responses to nonwords with implausible pronunciations (Error items), such as *dwal* pronounced as /jEsts/. As the orthographic and phonological forms of these items did not contain any (or almost any) PSCs found in English, no participant should find these pronunciations acceptable. Secondly, an estimate of the method's specificity was calculated using the ratings for the Error items mentioned above and items with odd pronunciations (Odd items), such as *gloost* pronounced as /glEst/. As the vowel pronunciations for the Odd items represented PSCs typically not found in English, nearly no human participant should intentionally produce these pronunciations and as such should not accept these pronunciations either. As such, the Error and Odd items should not be accepted in the rating task. However, as the Odd items contained some PSCs found in English (e.g., the onsets and codas in each item), the acceptability ratings for these items should be higher than those for the Error items. See Appendix 11, Table 11A for full list of stimuli.

6.3.2.2 Items with context sensitive onset C or G. A subset of C-initial and G-initial items from a nonword naming experiment by Treiman and colleagues (2007, Exp. 1) and additional two items (*gerd* and *gord*) were used to evaluate the rating method. In principle, this method should show a converging pattern of results with that of the naming data from the previous studies. The expected pattern of results from the rating method was based on naming responses from adult skilled readers in experiments conducted by Treiman and colleagues (2007, Exp. 1) and Treiman and Kessler (2019, Exp. 1). The proportion of context sensitive pronunciations for the experimental items were calculated as the proportion of soft

pronunciations relative to the total of soft and hard pronunciations for each item group. The findings from the two studies show that the proportion of soft pronunciations for critical items with *c* or *g* onset were approximately .8 and .15, respectively[40]. The proportion of soft pronunciation for control items with *c* or *g* onset were .01 at most. The proportions of soft pronunciations thus reflect the statistics of the English language in the sense that the critical items received soft pronunciations, while the control items did not.

In the rating task, each item was presented twice – once paired with a soft pronunciation (e.g., *cilsh* pronounced as /sIlS/) and once paired with a hard pronunciation (e.g., *cilsh* pronounced as /kIlS/). In line with the findings from the nonword naming experiments described above, the rating responses for critical and control items, when they were paired with a soft or hard pronunciation, were expected to show a similar pattern, that is, rejecting pronunciations that were produced by nearly no participant in the naming task and accepting pronunciations that were produced often enough to be considered deliberate in the naming task (i.e., proportion of .15 is considered high enough to reflect deliberate naming responses). Additionally, the previous naming data suggests certain relative patterns of ratings, such as higher acceptability ratings for C-critical items paired with soft pronunciations than C-critical items paired with hard pronunciation, as a clear majority of the participants in the naming task produced soft pronunciations for these items. See Appendix 11, Table 11B for a full list of stimuli.

6.3.2.3 Irregular items with low and high token frequency. Naming and rating responses to nonwords with irregular word bodies were compared directly, as a within-subjects design. Each nonword was presented twice in the rating task – once paired with an irregular vowel pronunciation (e.g., *bealm* pronounced as /bElm/) and once paired with a regular pronunciation (e.g., *bealm* pronounced as /bilm/). The acceptability ratings in the rating task were expected to reflect the naming responses in the naming task for each subject, that is, the items a given participant pronounced irregularly should also receive higher ratings from that participant in the rating task, when these items are paired with an irregular compared to a regular pronunciation. These items were also used for evaluating sensitivity of the rating method, by extracting the human modal response for each item in the naming task. For a full list of stimuli, see Appendix 3, Tables 3B and 3D.

[40] These figures are based on the proportions of soft pronunciations for the critical items produced by university students (Treiman & Kessler, 2019, Exp. 1), which were reasonably similar to the ones reported in Treiman et al. (2007), separately for items beginning with *ce* (.84) and *ci* (.87) and for items beginning with *ge* (.16) and *gi* (.04).

6.3.3 Procedure

The procedure is described in detail in Chapter 4, Section 4.2.3. In summary, the Naming-Rating group completed a nonword naming task, nonword rating task and a vocabulary task, whereas the Rating-Only group completed the last two tasks. Each trial in the rating task consisted of presentation of the nonword's written form and pronunciation, the acceptability of which the participants rated on a six-point scale from VERY BAD to VERY GOOD. The vocabulary task consisted of pronouncing a word and choosing the best definition for it from four options.

6.3.4 Data processing

6.3.4.1 Exclusion of trials. The procedure for trial exclusion reported in Chapter 4 was also followed in the current study – i.e., any nonwords that were based on a word that was pronounced or defined incorrectly by a participant in the vocabulary task were removed from that participant's data. See Chapter 4 (Section 4.2.4) for further details about trial exclusion.

6.3.4.2 Naming data. Apart from the data loss for the Irregular nonwords with low and high token frequency (reported in detail in Chapter 4, Section 4.2.4.4), very few trials were lost for the C and G-initial items. Due to audio recording issues, transcriber disagreement or a participant not responding on time, 0.65 % of the trials for C and G-initial items were lost.

6.3.4.3 Rating data. The labels of the rating scale were re-coded as follows: 1 = VERY BAD, 2 = BAD, 3 = PROBABLY NOT OK, 4 = PROBABLY OK, 5 = GOOD and 6 = VERY GOOD. Mean ratings for each item group from each participant were then calculated. The means from each participant were treated as continuous data and thus parametric hypotheses tests were used except for the analyses of the Error and Odd items (see below). Despite ongoing debate on which statistical tests are appropriate for data from Likert-type response formats (Carifio & Perla, 2007), parametric tests have been shown to be remarkably robust against assumption violations, such as scale of measurement (Norman, 2010). Participants with extreme outliers (defined as values that were more than three times the interquartile range below the 1st quartile or above the 3rd quartile) in any of the relevant item groups for a given analysis were removed. For the C and G-initial items, only one participant in the Rating-Only group was excluded due to outliers (remaining n = 68). Due to notably skewed ratings for the Error items (the mean rating was the minimum value of 1 ('very bad') for 80% of the participants in the Naming-Rating group and 72% of the participants in the

Rating-Only group), a non-parametric binomial sign test was used for the analyses of the Error and Odd items, which required no outlier removal.

6.3.4.4 Statistical power. Sensitivity power analyses were computed using GPower (Faul et al., 2007) for each hypothesis test and the resulting minimum, reliably detectable effect sizes for each analysis will be reported along with the observed effect sizes from the analyses. For C and G-initial items, where several hypothesis tests were performed on the same data set, Bonferroni correction was applied so that each test after the global 2x2x2 ANOVA was included in what was considered a family of tests for each participant group (Naming-Rating or Rating-Only groups) separately. Therefore, the corrected alpha level for all the hypotheses tests for C and G-initial items was .05/14 = .004.

6.4 Experiment 1 Results

6.4.1 Error and Odd items

The Error and Odd items were expected to receive ratings in the 'unacceptable' range (i.e., below 4 ('probably ok') and the Odd items were expected to receive higher ratings than the Error items.

Additionally, a measure of specificity of the rating method was calculated as the percentage of trials that were rejected (i.e., that received a rating 3, 'probably not ok', or lower), out of all the trials in the rating task.

These predictions were tested for both the Naming-Rating and the Rating-Only group. A binomial sign test was used due to notably skewed ratings for the Error items, which violated assumptions for both t-tests and Wilcoxon signed-rank test.

The median rating of 1 for Error items and 2.6 for Odd items both differed reliably from the critical value 4 ('probably ok') in Naming-Rating group (both $p < .001$). Similarly, the median rating of 1 for Error items and 2.3 for Odd items were both significantly different from 4 (both $p < .001$) in the Rating-Only group. Furthermore, the median difference in acceptability ratings for Error and Odd items was 1.6 ($p < .001$) for the Naming-Rating group and 1.3 ($p < .001$) for the Rating-Only group.

The specificity of the rating method, calculated for both groups of participants separately, was 90.43% in the Naming-Rating group and 92.61% in the Rating-Only group.

The findings suggest that the participants in both groups paid attention to the task and that the rating method has a high level of specificity – unlikely pronunciations for nonwords were rated as unacceptable. Furthermore, items that embody some PSCs of English (Odd items) were deemed more acceptable than items with virtually no PSCs of English (Error items).

6.4.2 Items with context sensitive onset C or G

The acceptability ratings to the C and G-initial items were expected to reflect the general pattern of naming responses found in previous studies (Treiman et al., 2007; Treiman & Kessler, 2019). This resulted in a number of predictions, both in terms of the relative ratings (e.g., that hard pronunciations are favoured over soft pronunciations for control items) and the absolute ratings (e.g., soft pronunciations for control items would be rated as unacceptable). See the last paragraphs of Section 6.2 for the full list of predictions.

This pattern of results was expected to be found in both Naming-Rating and Rating-Only group. The predictions regarding the relative ratings were tested with a 2x2x2 repeated measures ANOVA, inspecting the effects of Onset (C, G), Condition (Control, Critical) and Pronunciation (Hard, Soft) on mean acceptability ratings, for each participant group separately.

Most importantly, the analyses revealed a significant three-way interaction between Onset, Condition and Pronunciation in both groups (see Appendix 12, Table 12A). All lower order interactions and main effects were also statistically significant, apart from Onset x Condition interaction (ns. in both groups) and the main effect of Onset in the Naming-Rating group only. In the presence of the three-way interaction, simple two-way interactions at each level of Onset were inspected next. For both groups, there was a reliable interaction between Condition and Pronunciation, along with significant main effects of the two (see Appendix 12, Table 12B). Finally, simple main effect of Pronunciation was computed for each Onset-Condition combination, for both groups[41]. These comparisons revealed that the simple main effects of Pronunciation were significant for all other Onset-Condition combinations except for the C-critical items, in both groups (Table 6.1). Descriptive statistics for both groups are presented in Table 6.2.

[41] The minimum detectable effect size for this effect of interest was f(V) = 0.47, for Bonferroni corrected alpha level of .004, power of .8 and the smallest sample size of 68 in these analyses, see Table 6.1.

Table 6.1

Effect of Pronunciation (hard or soft) on acceptability ratings of C and G-initial items at each level of Onset and Condition

Onset	Condition	df	F	p-value	Cohen's f
Naming-Rating group (n = 69)					
C	Control	1, 68	187	< .001	1.66
C	Critical	1, 68	5	0.03	0.27
G	Control	1, 68	329	< .001	2.20
G	Critical	1, 68	148	< .001	1.47
Rating-Only group (n = 68)					
C	Control	1, 67	264	< .001	1.98
C	Critical	1, 67	0.58	0.45	0.10
G	Control	1, 67	276	< .001	2.03
G	Critical	1, 67	135	< .001	1.42

Table 6.2

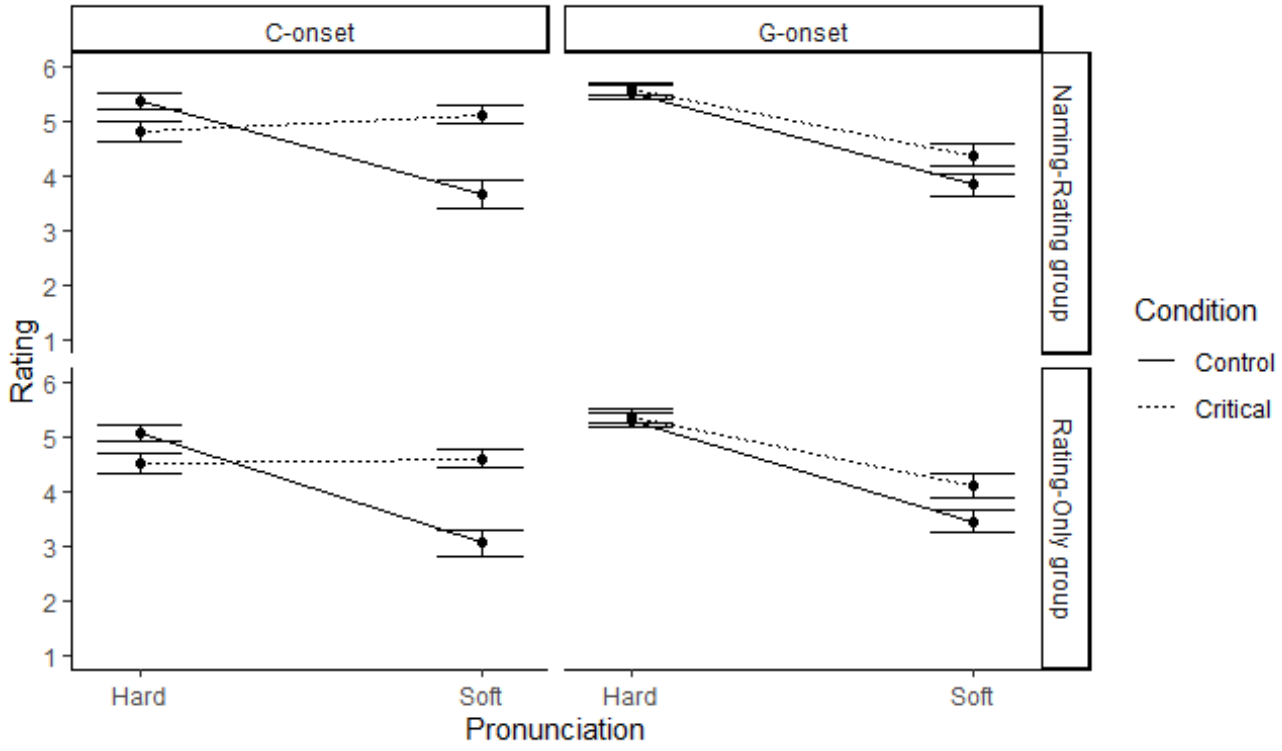
Mean ratings for C and G-initial items paired with soft and hard pronunciations

Condition	Pronunciation	Group			
		Naming-Rating (n = 69)		*Rating-Only (n = 68)*	
		M	*SD*	*M*	*SD*
C-control	*Hard*	5.37	0.60	5.08	0.58
	Soft	3.68	1.03	3.06	0.97
G-control	*Hard*	5.54	0.52	5.31	0.55
	Soft	3.85	0.86	3.46	0.87
C-critical	*Hard*	4.83	0.82	4.52	0.75
	Soft	5.12	0.71	4.62	0.71
G-critical	*Hard*	5.61	0.45	5.39	0.51
	Soft	4.39	0.85	4.12	0.89

Figure 6.1 depicts the pattern of mean ratings for the C and G-initial items. As can be seen, in both groups, hard pronunciations for control items received higher ratings than soft pronunciations for these items. The ratings for the C-critical items do not differ based on pronunciation type. Hard pronunciations for the G-critical items received higher ratings than soft pronunciations for the same items. The initial three-way interactions found between Onset, Condition and Pronunciation can be interpreted as Condition modulating the effect of Pronunciation on acceptability ratings only for the C-initial items, but not for G-initial items.

Figure 6.1

Mean ratings of C-onset and G-onset nonwords by Naming-Rating and Rating-Only groups



Note. Error bars are 95% confidence intervals of the mean ratings. Naming-Rating group (n = 69), Rating-Only group (n = 68).

Next, the predictions regarding the absolute ratings were tested with one-sample t-tests (see Table 6.3 for detailed results). These tests confirmed that ratings for critical items with soft pronunciations were reliably above 3 ('probably not ok') for both C and G-initial items, in both Naming-Rating and Rating-Only groups. However, C and G-initial control items with soft pronunciation were rated reliably below 4 ('probably ok') only by the Rating-Only group, whereas the ratings given by the Naming-Rating group were not significantly different from the critical value. [42]

[42] Bonferroni-adjusted alpha level of .004 was used for each group of participants. As such, the minimum detectable effect size for a two-tailed, one-sample t-test with power of .8 and n = 68 was Cohen's d = 0.47.

Table 6.3

Results of one-sample t-tests comparing mean ratings for C and G-initial items to critical values of 4 ('probably ok') and 3 ('probably not ok')

Item type	Mean vs crit. value	95 % CI of mean	df	t-value	p-value	Cohen's d
Naming-Rating group (n = 69)						
C-control soft	3.68 vs 4	3.43 – 3.93	68	-2.59	0.012	-0.31
G-control soft	3.85 vs 4	3.64 – 4.05	68	-1.48	0.14	-0.18
C-critical soft	5.12 vs 3	4.95 – 5.29	68	24.97	< .001	3.01
G-critical soft	4.39 vs 3	4.19 – 4.60	68	13.67	< .001	1.65
Rating-Only group (n = 68)						
C-control soft	3.06 vs 4	2.83 – 3.3	67	-8.00	< .001	-0.97
G-control soft	3.46 vs 4	3.25 – 3.67	67	-5.12	< .001	-0.62
C-critical soft	4.62 vs 3	4.44 – 4.79	67	18.89	< .001	2.29
G-critical soft	4.12 vs 3	3.90 – 4.34	67	10.34	< .001	1.25

In summary, both groups favoured hard pronunciations for control items over soft pronunciations, as expected. However, only the Rating-Only group judged control items with soft pronunciations as unacceptable, while the Naming-Rating group was more lenient for these items, with mean ratings approximately at 'probably ok'. Furthermore, against my predictions, C-critical items with soft pronunciations were not favoured over C-critical items with hard pronunciations by either group, although both groups rated C-critical items with soft pronunciations as acceptable. Finally, hard pronunciations assigned to G-critical items were favoured over their soft-pronunciation counter parts, and G-critical items with soft pronunciations were judged as acceptable by both groups, in line with my predictions.

The results from the rating method thus converge with the pattern of results previously found via the naming method in some ways (hard pronunciations for all control items and G-critical items were favoured, while soft pronunciations for G-critical items were still accepted), but not others (the preference for soft pronunciations for C-critical items was not evident in the rating data, nor was rejection of soft pronunciations assigned to control items in the Naming-Rating group).

6.4.2.1 Naming responses. The C and G items were also named by the Naming-Rating group. The proportion of soft pronunciations was .46 for the C-critical items, .02 for the G-critical items, .004 for the C-control items and 0.00 for the G-control items. The incidence of context sensitive pronunciations was thus considerably lower in this sample compared to the

previously reported findings of .8 for C-initial and .15 for G-initial items (Treiman & Kessler, 2019). This discrepancy will be considered further in the Discussion.

6.4.3 Irregular items with low and high token frequency

This direct comparison of the naming and rating responses to the Irregular items was expected to show a converging pattern of responses, such that the same pronunciations produced in the naming task would also be favoured in the rating task.

The irregular items were also used to arrive at a sensitivity measure for the rating method.

The naming responses to the Irregular items (regardless of their token frequency) from each participant in the Naming-Rating group were categorised as regular or irregular, based on the critical vowel pronunciation ('other' responses were excluded as these were not represented in the rating task). Out of the valid naming responses, 51% were pronounced regularly, 28% irregularly and 21% were lost as 'other' responses. The set of regularly and irregularly named items was unique for each participant, based on the types of responses they gave. The mean acceptability ratings were then calculated for both sets of items for each participant (see Table 6.4 for summary).

Table 6.4

Mean acceptability ratings to regularly and irregularly named Irregular items

	Regularly named		**Irregularly named**	
	Rating – Regular pronunciation	*Rating – Irregular pronunciation*	*Rating – Regular pronunciation*	*Rating – Irregular pronunciation*
Mean	5.35	4.42	4.66	5.38
SD	0.45	0.59	0.61	0.48

Note. Regularly/Irregularly named = items that received a regular/irregular pronunciation in the naming task. Rating – Regular/Irregular pronunciation = mean ratings for nonwords that were paired with a regular/irregular pronunciation in the rating task.

For the regularly named item set, when these items were paired with regular pronunciations in the rating task, they received reliably higher ratings than when they were paired with irregular pronunciations ($t(68) = 13.65$, $p < .001$, $dz = 1.64$). Similarly for the irregularly named item set, irregularly pronounced items in the rating task received higher acceptability

ratings than when they were pronounced regularly ($t(68) = 10.04$, $p < .001$, $dz = 1.22$).[43] Thus, the participants' naming behaviour converged with their subsequent rating behaviour of the same items.

Furthermore, mean ratings to irregularly and regularly pronounced Irregular items in the rating task (regardless of token frequency) given by both groups of participants also converged with the overall pattern of the naming responses from the naming task given by the Naming-Rating group. The naming responses from the Naming-Rating group showed overall preference for regular pronunciations: the mean proportion of regular pronunciations, out of all valid responses ($M = .51$, $SD = 0.1$) was higher than the proportion of irregular pronunciations ($M = .28$, $SD = 0.08$), $t(68) = 11.45$, $p < .001$, $dz = 1.38$. The overall ratings to Irregular items from the Naming-Rating group also showed preference for regular pronunciations ($M = 5.06$, $SD = 0.47$) over irregular ones ($M = 4.82$, $SD = 0.49$), $t(68) = 4.97$, $p < .001$, $dz = 0.6$. Similarly, the overall ratings from the Rating-Only group were higher for the regularly pronounced items ($M = 4.69$, $SD = 0.47$) than for the irregularly pronounced items ($M = 4.43$, $SD = 0.49$), $t(68) = 6.50$, $p < .001$, $dz = 0.78$[43].

Additionally, a sensitivity measure of the rating method was computed in the following way: only trials where a nonword was paired with the pronunciation corresponding to the human modal response (regular or irregular vowel pronunciation) for each item in the naming task were retained. Percentage of trials in the rating task that were higher than 4 ('probably ok') out of all the trials were then calculated. There were 48 items with a clear human modal response (i.e., no ties between the proportion of regular and irregular responses[44]). The total number of trials was thus (48 items * 69 participants) minus missing trials (trials in the rating task were classified as missing if a participant did not know the word that a given nonword was based on). The sensitivity of the rating method was 94.05% in the Naming-Rating group and 87.65% in the Rating-Only group.

However, the human modal responses in this full set of data were sometimes arbitrary, as the percentage of participants giving the modal response was low for some items (e.g., only 21% pronounced the nonword *choung* regularly as /J6N/). Furthermore, due to the considerable

[43] Minimum effect size of 0.34 was computed for a 2-tailed, paired samples t-test with alpha level of .05, power of .8 and sample size of 69.
[44] Additional two items were excluded due to an error in stimuli construction: the onsets for the items were swapped between the naming and the rating tasks (*phoute* and *dwonge* in the naming task, but *dwoute* and *phonge* in the rating task).

data loss (see Chapter 4, Section 4.2.4), the number of participants with valid responses in the naming task was low for some items (e.g., only one participant correctly defined the word *kirsch,* and as such, only one participant's naming response determined the human modal response for the nonwords *flirsch* and *mirsch,* and only one participant provided a rating for these items). These factors cast doubt on whether some of the human modal responses should be considered common enough to expect clear acceptance of these pronunciations in the rating task. To consider this possibility, a stricter criterion for a human modal response was also applied, such that only items with at least 35 valid responses (half of the original sample size of 69) and at least 50% of the participants giving this response were retained. This resulted in a sample of 25 items. This stricter analysis showed that the sensitivity of the rating method was 95.46% in the Naming-Rating group and 90.13% in the Rating-Only group.

6.5 Experiment 1 Discussion

The evaluation of the nonword rating method revealed that overall, this method seems to capture nonword processing in the expected way. Firstly, implausible pronunciations assigned to nonwords (Error and Odd items) were rated as unacceptable. These acceptability judgements were also fine-grained enough to differentiate between items with no PSCs of English (Error items) and those with some PSCs of English (Odd items). The specificity of the rating method was also high, ranging from 90% to 93%.

Secondly, the pattern of results from the rating task matched reasonably well with the pattern found in previous naming studies (Treiman et al., 2007; Treiman & Kessler, 2019): soft pronunciations assigned to C and G-critical items were rated as acceptable, hard pronunciations to C and G-control items were rated more acceptable than soft pronunciations to these items, and ratings for G-critical items with soft pronunciations were rated less acceptable than G-critical items with hard pronunciations by both groups of participants. However, there were two findings that were not expected. The first one was that neither group preferred soft pronunciations assigned to C-critical items over hard pronunciations for these items. However, the Naming-Rating group's naming responses to these items in the naming task should give some context to this finding: the proportion of soft pronunciations produced by the Naming-Rating group was .46 for C-critical items and .02 for G-critical items, both of which are lower than the proportions reported in previous studies (.8 and .15, respectively). This might be because only a small subset of the items was used in the current study. Nevertheless, this lower incidence of context sensitive pronunciations for C-critical items bridges the gap between the results from the naming and rating methods – if only less

than half of the C-critical items were assigned a soft pronunciation, then not favouring the soft pronunciation for these items in the rating task may not be that surprising. The second unexpected finding was that, unlike the Rating-Only group, the Naming-Rating group did not reject soft pronunciations assigned to C and G control items. However, the mean ratings for these items remained at the border of acceptable and unacceptable. A potential reason for this lenient rating, evident throughout the Naming-Rating group's responses, is fatigue or practice effects for rating the nonwords after the preceding naming task. These potential explanations will be addressed in Experiment 2.

Finally, direct comparison of the Naming-Rating group's naming responses to their subsequent rating responses of the same items revealed that the naming behaviour converged with the rating behaviour: the type of pronunciation assigned to a nonword in the naming task (regular or irregular) was also favoured in the rating task. The overall ratings for regularly and irregularly pronounced items by both groups were also in line with the overall naming responses from the Naming-Rating group: higher proportion of regularly named Irregular items was reflected in the higher acceptability ratings for regularly named Irregular items in the rating task. The sensitivity measures of the rating task were also high, ranging from 88% to 94%.

Limitations of the current study will be covered in the General Discussion (Section 6.9).

To conclude, the assessment of the nonword rating method revealed high sensitivity and specificity of the method and promising convergence between the pattern of results from the naming and rating methods. However, the current study raises questions about the preceding exposure to the items that will be rated, and some differences found between the pattern of ratings from the two groups of participants in the present experiment will be addressed in the next experiment.

6.6 Experiment 2

The comparisons between nonword naming and rating responses have so far been promising, as demonstrated by the mostly expected pattern of acceptability ratings in Experiment 1 of the current chapter. However, some unexpected results were also obtained, as the group of participants that had previously named the same nonwords (Naming-Rating group), produced a different pattern of acceptability ratings for these nonwords than another group of participants that had only rated the nonwords (Rating-Only group). More specifically, the Naming-Rating group did not show the expected pattern of acceptability ratings for Irregular

nonwords, when these nonwords were divided based on the token frequency of the words they shared a word body with (reported in Chapter 4, Section 4.3.1) or for the C and G-initial items (Experiment 1, current chapter), while the Rating-Only group did.

Experiment 2 set out to clarify whether this difference in the groups' rating behaviour was due to fatigue or practice effects. To test this, a new group of participants first named unrelated nonwords that were comparable in difficulty to the nonwords in Experiment 1 and then rated the same nonwords that were used in Experiment 1. This way, if fatigue was the main reason for the unexpected rating responses from the Naming-Rating group, the new group of participants (Unrelated-Rating group) should show similar pattern of ratings with the Naming-Rating group, for the Irregular nonwords with low and high token frequency and C and G-initial nonwords. If, on the other hand, the pattern of rating responses from the Naming-Rating group were a result of practice effects due to more extensive exposure to experimental items, the new Unrelated-Rating group should show more similar rating behaviour to that of the Rating-Only group.

Additionally, to complement the available rating data for the C and G-onset items from Experiment 1, the current experiment added rating responses for these items from two more groups of participants. Together these four groups of participants formed a combination of conditions that allowed a closer inspection of the potential influence of previous tasks on the subsequent rating behaviour – when the same items have been named before rating them (Naming-Rating group in Experiment 1 and Naming-Rating-type group in the current experiment), when unrelated nonwords have been named beforehand (Unrelated-Rating group in the current experiment) and when no nonwords were named before the rating task (Rating-Only group in Experiment 1).

A secondary goal of the current experiment was to complement the findings from Experiment 1 by directly comparing naming and rating responses to another set of irregular nonwords, from another group of participants (Naming-Rating-type group, reported in Chapter 5). This comparison served as a conceptual replication of the findings reported in Experiment 1 (Section 6.4.3) – namely, that skilled readers tended to favour the type of pronunciations in the rating task that they themselves had produced previously in the naming task. The naming and rating responses from this group also served as another measure of sensitivity of the rating method.

Finally, a replication of the pattern of ratings for the Error and Odd items in Experiment 1 was carried out with two additional groups of participants (the Unrelated-Rating and the Naming-Rating-type groups). None of these groups named these items before rating them.

6.7 Experiment 2 Method

6.7.1 Participants

The first group of participants named nonwords that were unrelated to the upcoming nonwords in the rating task (Unrelated-Rating group), the second group of participants named nonwords and subsequently rated the same nonwords (Naming-Rating-type group, Chapter 5, see Section 5.2.1 for participant characteristics). The sample size for the Unrelated-Rating group was 64 (13 males), with mean age of 19.86 years (*SD* = 3.23). The two groups of participants in the current experiment were from the same student population as the participants in Experiment 1, recruited based on the same eligibility criteria and an additional requirement that they had not participated in the Experiment 1. See Table 6.5 for a summary of the tasks completed by each of the four groups of participants.

Table 6.5

Tasks completed by groups of participants in Experiment 1 and Experiment 2

Group	Experiment	Naming task items	Rating task items
Naming-Rating	Experiment 1	C & G initial, Irregular-token	Error & Odd, C & G initial, Irregular-token
Rating-Only	Experiment 1	-	Error & Odd, C & G initial, Irregular-token
Unrelated-Rating	Experiment 2	Unrelated items	Error & Odd, C & G initial, Irregular-token
Naming-Rating-type	Experiment 2	C & G initial, Irregular-type	Error & Odd, C & G initial, Irregular-type

Note. Irregular-token = nonwords with irregular word bodies that are shared with words with low or high token frequency (items from Chapter 4). Irregular-type = nonwords with irregular word bodies that are shared with a single or several existing words (items from Chapter 5).

Ethics approval for the current experiments was granted by the School of Psychological Science Research Ethics Committee in University of Bristol (ethics approval code: 0229).

6.7.2 Materials

The nonwords named by the Unrelated-Rating group should be different but comparable in difficulty to the items in Experiment 1, so that the naming task would be equally taxing for

both groups of participants – the Naming-Rating group in Experiment 1 and the Unrelated-Rating group in the present experiment. To this end, it was ensured that the number of nonwords that are likely more difficult to name, that is, nonwords with several plausible pronunciations (nonwords sharing a word body with an irregularly pronounced word/words and nonwords with context sensitive onsets C or G) was the same for both groups of participants. The rest of the items were fillers. See Appendix 3, Tables 3B and 3C for the full list of nonwords named by the Naming-Rating group (Experiment 1) and Appendix 11, Table 11C for the nonwords named by the Unrelated-Rating group. The nonwords rated by the Unrelated-Rating group were identical to the ones rated by the two groups in Experiment 1.

The nonwords named by the Naming-Rating-type group included the C and G-initial items also named by the Naming-Rating group in Experiment 1, nonwords with irregular word bodies that either occurred in several existing words (Irregular-Many items) or only in a single word (Irregular-Single items) and filler items. These items were used in the experiment reported in Chapter 5, which includes some of the naming and rating results from the Naming-Rating-type group (see Appendix 8, Tables 8A and 8B for full list of items named by the Naming-Rating-type group). The nonwords rated by the Naming-Rating-type group were the Error and Odd items and C and G-initial items, identical to the ones used in Experiment 1. Additionally, the Naming-Rating-type group rated the Irregular-Many and Irregular-Single items (see Appendix 8, Tables 8C and 8D for the full list of items rated by the Naming-Rating-type group).

6.7.3 Data processing

6.7.3.1 Exclusion of participants. The pre-registered data processing plan (https://osf.io/znpyf) for the experiment reported in Chapter 4 was also followed in the current experiment. According to this plan, any participant who rated four or more of the Error items as ‘probably ok’ or higher, would be excluded. This is because such lenient rating behaviour is suggestive of insufficient attention for the task. One participant in the Unrelated-Rating group was excluded due to this criterion, leaving the sample size for this group at 63.

6.7.3.2 Exclusion of trials. The procedure for trial exclusion reported in Chapter 4 was also followed in the current study. Two native speakers of English transcribed the Unrelated-Rating group’s naming responses in the vocabulary task, agreeing on 95% of the items. Due to an error in pre-processing of the vocabulary responses, the item *heart* was marked as unknown for all the participants. This error was later corrected, but the correctness of the

pronunciations of this item was assessed by myself, a non-native English speaker with full awareness of the purpose this study. However, as the word *heart* is a highly frequent word and there was little variation in how this item was pronounced, I believe this had little impact on the assessment accuracy of the participants' vocabulary knowledge for this item. Overall, 29.59% of the rating trials were lost due to insufficient vocabulary knowledge. See Table 6.6 for a summary of the data loss for all three groups that rated the Irregular items with low and high token frequency (the Unrelated-Rating group reported in the present experiment and the Naming-Rating and Rating-Only groups from Experiment 1).

Table 6.6

Percentage of items excluded in the rating task of Irregular nonwords

Group	Semantic	Pronunciation	Total
Rating-Only (n = 69)	18.37	17.55	35.92
Naming-Rating (n = 69)	16.82	16.67	33.49
Unrelated-Rating (n = 63)	17.23	12.36	29.59

Note. Semantic = percentage of items lost due to incorrect definition of the base word; Pronunciation = percentage of items lost due to incorrect vowel pronunciation of the base word, lost recordings and discrepancies in the phonemic transcription.

6.7.3.3 Naming data. Only naming responses from the Naming-Rating-type group will be reported. The percentage of lost trials due to failed audio recording was 1.8% for the Irregular-Many and Irregular-Single nonwords and 0.64% for the C and G-initial items.

6.7.3.4 Rating data. As in Experiment 1, the labels of the rating scale were re-coded as numeric, mean ratings for each item group from each participant were then calculated and participants with extreme outliers in any of the relevant item groups for a given analysis were removed. For the analyses of the C and G onset items, one participant was excluded from the Unrelated-Rating group, the remaining sample size for this group was thus 62. Due to skewed ratings for the Error items (the mean rating was the minimum value of 1 ('very bad') for 81% of the participants in the Unrelated-Rating group and 89% of the participants in the Naming-Rating-type group), a non-parametric binomial sign test was used for the analyses of the Error and Odd ratings data and therefore no outliers were removed.

6.8 Experiment 2 Results and Discussion

6.8.1 Irregular items with low and high token frequency

The purpose of the current analysis was to clarify the source of discrepant rating responses found in Chapter 4, namely, that ratings for Irregular nonwords with low and high token frequency did not differ from one another in the Naming-Rating group, but the ratings from the Rating-Only group showed the expected pattern (i.e., higher ratings for Irregular-high than for Irregular-low items). The Unrelated-Rating group rated the Irregular nonwords, but had not named these items before. Additionally, to supplement the findings from Experiment 1, the general preference for regular or irregular pronunciations for the Irregular items was investigated in the Unrelated-Rating group's rating responses. Finally, another measure of sensitivity of the rating method was computed from the ratings of the Unrelated-Rating group.

The Unrelated-Rating group gave slightly higher ratings to Irregular-high items (M = 4.73, SD = 0.58) than to Irregular-low items (M = 4.66, SD = 0.61), but this difference was not statistically reliable: $t(62) = 1.42$, $p = 0.08$, $dz = 0.18$. [45]

Due to a considerable loss of trials (because of insufficient vocabulary knowledge of the words the nonwords were based on), the same analysis was also run excluding any participants with less than 10 valid responses in each condition. The outcome of this analysis was similar: the Irregular-high items received higher ratings (M = 4.76, SD = 0.53) than the Irregular-low items (M = 4.67, SD = 0.59), a difference that was confirmed statistically ($t(53) = 1.86$, $p = .03$, $dz = 0.25$).[46]

The rating responses to the Irregular-high and Irregular-low items showed the expected pattern (i.e., higher ratings for Irregular-high than for Irregular-low items), but this difference did not reach statistical significance when the minimum effect size for the analyses was taken into account (observed effect size at most dz = 0.25 vs minimum effect size dz = 0.34). Thus, the current experiment showed a numerical trend in the expected direction and a difference between acceptability ratings for the Irregular items that approached significance. By contrast, the difference in the acceptability ratings from the Naming-Rating group in Chapter 4 was numerically in the opposite direction and far from statistical significance (see the results in detail in Chapter 4, Section 4.3.1). The Rating-Only group in Chapter 4 was

[45] The minimum detectable effect size with alpha level of .05, power of .8 and sample size of 64 was $dz = 0.31$
[46] The minimum detectable effect size with alpha level of .05, power of .8 and sample size of 55 was $dz = 0.34$

therefore the only group that showed a reliable difference in the acceptability ratings in the expected direction. With these patterns of rating responses from the three groups, it is concluded that the two groups who did not name the critical items before rating them showed the expected rating behaviour at least as a numerical trend. As such, the question of why the rating behaviour of the Naming-Rating group differed from that of the Rating-Only group in Chapter 4 appears to be related to practice effects rather than fatigue.

Another comparison between the naming and rating responses between mean ratings for the Irregular items when they were paired with an irregular pronunciation in the rating task and when they were paired with a regular pronunciation was conducted. As the naming responses indicated an overall preference for regular pronunciations in the Naming-Rating group in Experiment 1 (Section 6.4.3), the same preference was expected to be found in the ratings of the Unrelated-Rating group for the same items. The overall ratings to Irregular items from the Unrelated-Rating group showed higher acceptability ratings for regular pronunciations (M = 5.02, SD = 0.48) over irregular ones (M = 4.70, SD = 0.56), $t(62) = 6.63$, $p < .001$, dz = 0.84[47].

Finally, a sensitivity measure of the rating method was computed using the rating trials that included a clear human modal pronunciation for a given nonword (48 items). The sensitivity of the rating method was 91.59%. Thus, the high level of sensitivity of the rating method found in Experiment 1 (Section 6.4.3) was replicated with a group of participants that had not named the critical items beforehand (the Unrelated-Rating group).

6.8.2 Items with context sensitive onset C or G

The aim of the present analyses was to clarify why the pattern of ratings for these items differed between the Naming-Rating and Rating-Only groups in Experiment 1.

The C and G-initial nonwords were rated by the Unrelated-Rating and Naming-Rating-type groups. A 2x2x2 repeated measures ANOVA, inspecting the effects of Onset (C, G), Condition (Control, Critical) and Pronunciation (Hard, Soft) on mean acceptability ratings were run, for each group separately. These analyses were conducted to test the same hypotheses as in Experiment 1, namely, that hard pronunciations are favoured over soft pronunciations for control items, that soft pronunciations are favoured over hard pronunciations for C-critical items and that hard pronunciations are favoured over soft pronunciations for G-critical items.

[47] The minimum detectable effect size with alpha level of .05, power of .8 and sample size of 63 was dz = 0.36

The most important findings were a significant three-way interaction between Onset, Condition and Pronunciation in both groups (Appendix 13, Table 13A). All lower order interactions and main effects were also statistically significant, apart from Onset x Condition interaction (ns. in both groups) and the main effect of Onset for Naming-Rating-type group. Due to the significant three-way interaction, simple two-way interactions at each level of Onset were inspected next. This analysis showed a reliable interaction between Condition and Pronunciation, along with significant main effects of the two for both groups (Appendix 13, Table 13B). Finally, simple main effect of Pronunciation was computed for each Onset-Condition combination[48]. These comparisons revealed that the simple main effects of Pronunciation were significant for all other Onset-Condition combinations except for C-critical items (Table 6.7). See Table 6.8 for descriptive statistics.

Table 6.7

Effect of Pronunciation (hard or soft) on acceptability ratings of C and G-initial items at each level of Onset and Condition

Onset	Condition	df	F	p-value	Cohen's f
Unrelated-Rating group (n = 62)					
C	Control	1,61	278.00	< .001	1.43
C	Critical	1,61	2.14	0.15	0.03
G	Control	1,61	188.00	< .001	1.15
G	Critical	1,61	92.00	< .001	0.75
Naming-Rating-type group (n = 55)					
C	Control	1,54	210.00	< .001	1.32
C	Critical	1,54	1.27	0.27	0.02
G	Control	1,54	207.00	< .001	1.30
G	Critical	1,54	202.00	< .001	1.28

[48] The minimum detectable effect size for this effect of interest was f(V) = 0.52 for Bonferroni corrected alpha level of .004, power of .8 and the smallest sample size of 55 in this analysis, see Table 8.

Table 6.8

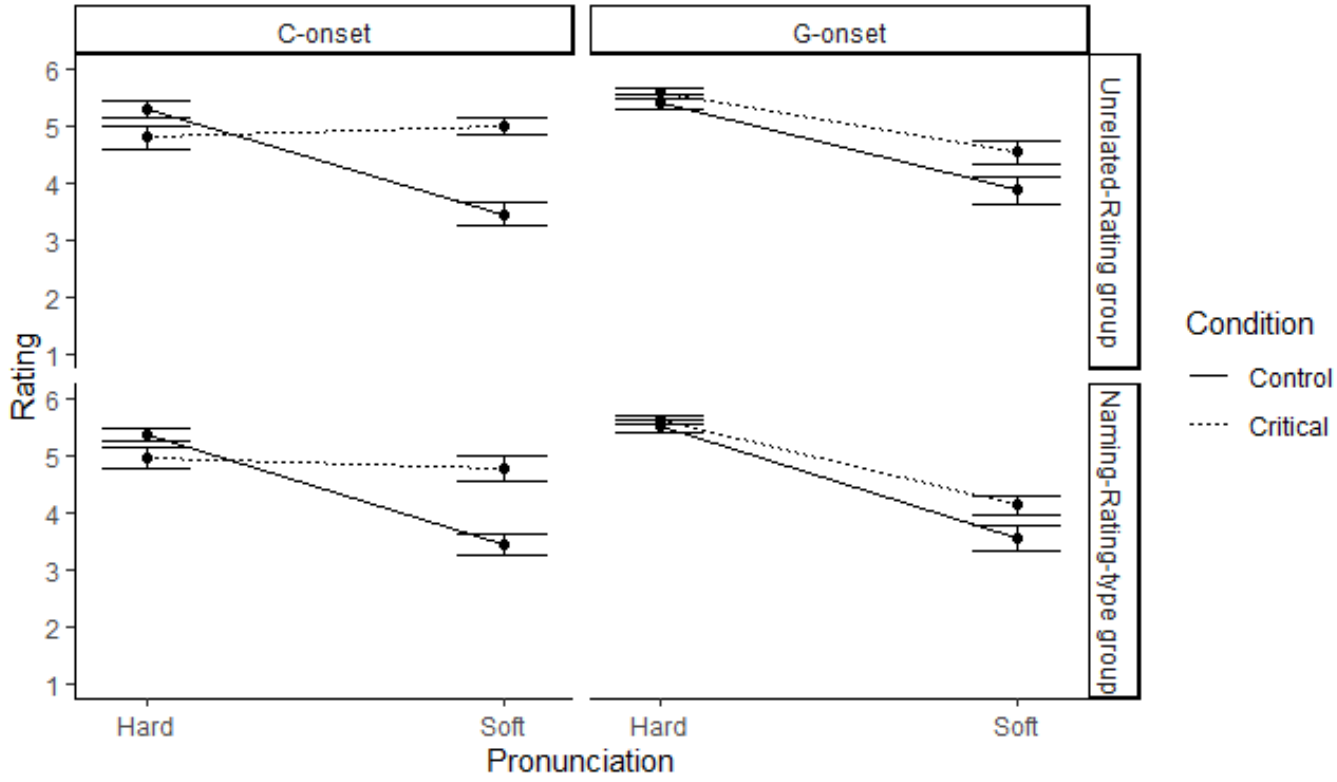
Mean ratings for C and G-initial items paired with soft and hard pronunciations

Condition	Pronunciation	Group			
		Unrelated-Rating (n = 62)		*Naming-Rating-type (n = 55)*	
		M	*SD*	*M*	*SD*
C-control	*Hard*	5.30	0.59	5.37	0.53
	Soft	3.47	0.86	3.45	0.83
G-control	*Hard*	5.42	0.55	5.52	0.50
	Soft	3.89	0.99	3.56	0.98
C-critical	*Hard*	4.82	0.88	4.97	0.76
	Soft	5.01	0.60	4.78	0.90
G-critical	*Hard*	5.59	0.42	5.63	0.37
	Soft	4.56	0.83	4.14	0.73

The results from these analyses are also depicted in Figure 6.2, which shows that soft pronunciations for the G-critical, G-control and C-control items received reliably lower acceptability ratings from both groups compared to ratings for these items when they were assigned a hard pronunciation. The initial three-way interactions found between Onset, Condition and Pronunciation can be interpreted as Condition modulating the effect of Pronunciation on acceptability ratings only for the C-initial items, but not for G-initial items. For C-initial items, effects of Pronunciation are only seen in control items, where items with hard pronunciation receive higher ratings than items with soft pronunciation. There is no reliable effect of Pronunciation on C-initial critical items. For G-initial items, items with hard pronunciation are rated as more acceptable than items with soft pronunciation, regardless of the Condition.

Figure 6.2

Mean ratings of C-onset and G-onset nonwords by Unrelated-Rating and Naming-Rating-type groups



Note. Error bars are 95% confidence intervals of the mean ratings. Unrelated-Rating group (n = 62), Naming-Rating-type group (n = 55).

To summarise these findings in terms of the three hypotheses, the pattern of results from both groups of participants supported two out of the three hypotheses:

Hard pronunciations were preferred for control items and G-critical items. However, C-critical items with soft pronunciation did not receive reliably higher ratings than C-critical items with hard pronunciation, which goes against my prediction. Nevertheless, the Unrelated-Rating group showed a numerical trend in line with this prediction.

Next, the predictions regarding the absolute values of the acceptability ratings were tested with one sample t-tests. The C and G-critical items with soft pronunciations were expected to

receive reliably higher ratings than 3 ('probably not ok'), and the C and G-control items with soft pronunciations were expected to receive reliably lower ratings than 4 ('probably ok'). These predictions were confirmed (see Table 6.9)[49], except that G-control items with soft pronunciation did not receive ratings reliably below 4 from the Unrelated-Rating group (when no correction for multiple comparisons was applied) or by either of the groups (when Bonferroni-correction and the corresponding smallest, reliably detectable effect size was considered). Due to different interpretations associated with these two possible results (that either Unrelated-Rating group or both groups failed to reject soft pronunciations assigned to G-control items), both results are considered in the Discussion.

Table 6.9

Results of one-sample t-tests comparing mean ratings for C and G-initial items to critical values of 4 ('probably ok') and 3 ('probably not ok')

Item type	Mean vs crit. value	95 % CI of mean	df	t-value	p-value	Cohen's d
Unrelated-Rating group (n = 62)						
C-control soft	3.47 vs 4	3.25 – 3.69	61	-4.83	< .001	-0.61
G-control soft	3.89 vs 4	3.63 – 4.14	61	-0.90	0.37	-0.12
C-critical soft	5.01 vs 3	4.86 – 5.16	61	26.30	< .001	3.34
G-critical soft	4.56 vs 3	4.35 – 4.77	61	14.79	< .001	1.88
Naming-Rating-type group (n = 55)						
C-control soft	3.45 vs 4	3.23 – 3.68	54	-4.86	< .001	-0.66
G-control soft	3.56 vs 4	3.30 – 3.82	54	-3.35	0.001	-0.45
C-critical soft	4.78 vs 3	4.53 – 5.02	54	14.58	< .001	1.97
G-critical soft	4.14 vs 3	3.94 – 4.34	54	11.50	< .001	1.55

Finally, the Naming-Rating-type group also named the C and G-onset items before rating them. The proportion of soft pronunciations was .35 for the C-critical items, .01 for the G-critical items, .01 for the C-control items and 0.00 for the G-control items. These proportions are similar to the ones found in Experiment 1 for the Naming-Rating group, although the incidence of soft pronunciations for C-critical items was somewhat lower in the current study (.35 compared to .45 in Experiment 1). The pattern of these naming responses also help understanding some of the findings in the rating responses: as the C-critical items were pronounced with a soft *c* less than half the time, the ratings for these items with a soft

[49] The smallest detectable effect size for a 2-tailed, 1-sample t-test with alpha level of .004, power of .8 was Cohen's d = 0.49 (for a sample of 62) and 0.52 (for a sample of 55)

pronunciation were not favoured over the same items with a hard pronunciation (e.g., Figure 6.2).

Taken together, the pattern of ratings for the C and G-initial items in the current experiment was very similar to the one reported in Experiment 1. Most of the predictions were confirmed, and the unexpected findings in the current experiment revolved around the same item groups as those in Experiment 1, namely, that neither group favoured soft pronunciations for C-critical items over hard pronunciations and that one or both groups failed to reject soft pronunciations assigned to G-control items. See General Discussion (Section 6.9) for further consideration of these findings.

6.8.3 Irregular items with low and high type frequency

The Naming-Rating-type group named and then rated nonword items with irregular word bodies that varied in type frequency (i.e., the number of words the nonwords shared a word body with, see Chapter 5). This allows another direct comparison between the naming and rating responses to the same items, from the same participants. I tested whether the type of naming responses given in the naming task would also be favoured in the rating task, e.g., whether items that were named regularly in the naming task would receive higher acceptability ratings in the rating task when they are paired with regular pronunciations compared to when they are paired with irregular pronunciations (and vice versa for the irregular items). First, naming responses to each item (n = 60) were categorised as regular, irregular or other based on the vowel pronunciation. Out of all the trials in the naming task, 49% resulted in a regular vowel pronunciation, 34% in an irregular vowel pronunciation and 17% of the trials were lost as ‘other’ responses. For each item retained for a given participant in the naming data, the ratings for the corresponding items were also retained from the rating task. Paired-samples t-tests were then run on mean ratings for the items that were named regularly in the naming task and separately for the items that were named irregularly in the naming task. Table 6.10 summarises the mean ratings for the regularly and irregularly named items, when they are paired with regular and irregular pronunciations in the rating task.

Table 6.10

Mean acceptability ratings to regularly and irregularly named Irregular items by Naming-Rating-type group

	Regularly named		Irregularly named	
	Rating – Regular pronunciation	*Rating – Irregular pronunciation*	*Rating – Regular pronunciation*	*Rating – Irregular pronunciation*
Mean	5.37	4.57	4.38	5.58
SD	0.39	0.43	0.54	0.37

Note. Regularly/Irregularly named = items that received a regular/irregular pronunciation in the naming task. Rating – Regular/Irregular pronunciation = mean ratings for nonwords that were paired with a regular/irregular pronunciation in the rating task.

For the regularly named item set, these items received reliably higher ratings in the rating task when they were paired with regular pronunciations compared to when they were paired with irregular pronunciations ($t(54) = 12.42$, $p < .001$, $dz = 1.67$). Similarly for the irregularly named item set, irregularly pronounced items in the rating task received higher acceptability ratings than when they were pronounced regularly ($t(54) = 15.92$, $p < .001$, $dz = 2.15$).[50] These results show that the participants' naming behaviour converged with their subsequent rating behaviour of the same items.

These results thus replicate the findings reported in Experiment 1 (Section 6.4.3), where a different group of participants named and rated a different set of nonwords. These findings are considered further in the General Discussion (Section 6.9).

Finally, a measure of sensitivity was calculated as the number of items paired with the human modal naming response that were rated as acceptable in the rating task. For 57 items with a clear human modal response, and rating responses from 55 participants, the sensitivity measure was 94.23%.

[50] The sensitivity analyses for a two-tailed, paired samples t-test with alpha level of .05, power of .8 and sample size of 55 yielded a minimum effect size of 0.38.

6.8.4 Error and Odd items

To replicate the findings reported in Experiment 1 (Section 6.4.1), the Error and Odd items were rated by the Unrelated-Rating and Naming-Rating-type groups. The predicted pattern of results was that these items would receive ratings below 4 ('probably ok') and that the Odd items would receive higher ratings than the Error items.

A binomial sign test was used due to notably skewed ratings for the Error items. The median rating of 1 for Error items and 2.5 for Odd items both differed reliably from the critical value 4 ('probably ok') in Unrelated-Rating group (both $p < .001$). Similarly, the median rating of 1 for Error items and 2.2 for Odd items were both significantly different from 4 (both $p < .001$) in the Naming-Rating-type group. Furthermore, the median difference in acceptability ratings for Error and Odd items was 1.5 ($p < .001$) for the Unrelated-Rating group and 1.2 ($p < .001$) for the Naming-Rating-type group.

The specificity of the rating method was calculated for both groups of participants separately, as a percentage of trials that were rejected (i.e., that received a rating 3, 'probably not ok', or lower), out of all the Error and Odd item trials in the rating task. The specificity of the rating method was at 91.75% in the Unrelated-Rating group and 94.45% in the Naming-Rating-type group.

A pooled measure of specificity was also computed, from all four groups of participants (from Experiment 1 and Experiment 2) combined, which showed that out of 5120 trials (20 items * 256 participants), 92.21% were rated as unacceptable, demonstrating a high level of specificity for the rating method.

Thus, a total of four groups of participants rated these items similarly, and in line with the predictions. This was the case regardless of whether these participants have named nonwords before completing the rating task (three of the groups) or not (one group).

6.9 General Discussion

In two Experiments, four different groups of participants gave acceptability ratings for pronunciations assigned to nonwords. The first experiment aimed to answer the following questions 1) whether the ratings show clear rejection of implausible pronunciations of nonwords and a preference for pronunciations containing some PSCs of English (Odd items) rather than none (Error items), 2) whether the ratings show a comparable pattern to that from previously reported naming responses (C and G-initial items), and 3) whether naming

responses are comparable to the rating responses from the same participants (nonwords with irregular word bodies). Due to unexpected findings in Experiment 1, namely, different patterns of ratings from two groups of participants, Experiment 2 aimed to discover the source of these differences, with fatigue or practice effects as potential reasons for the differences. Additionally, Experiment 2 was also designed to replicate findings from Experiment 1.

My first aim of the study was to inspect the acceptability ratings to Error items (e.g., *dwal* pronounced as /jEsts/) and Odd items (e.g., *gloost* pronounced as /glEst/). As expected, both types of items were rejected and the Odd items, which contained some PSCs of English, were rated as more acceptable than the Error items, which contained virtually no PSCs of English. This pattern of ratings was obtained from four different groups of participants, none of which had named these items beforehand. However, three of the groups had named other nonwords before the rating task, while one of the groups only completed the rating task. As such, the preceding nonword naming task does not seem to influence the pattern of rating behaviour for these items. The specificity of the rating method was high (92%), quantified as the percentage of Error and Odd item trials receiving 'probably not ok' or worse ratings from the four groups of participants combined. These findings speak to the usefulness of the rating method, as it provides fine-grained information about pronunciation preferences in skilled readers.

The second aim of the study was to compare rating responses to the pattern of naming responses obtained from previous studies (Treiman, et al., 2007, Exp. 1; Treiman & Kessler, 2019, Exp. 1). As expected, all four groups of participants favoured hard pronunciations over soft pronunciations for control items (regardless of onset) and for G-critical items. All groups of participants also deemed soft pronunciations for critical items as acceptable (regardless of onset) and rejected soft pronunciations assigned to C-control items (except for Naming-Rating group, Exp. 1). These findings fit well with the pattern of naming responses to a larger set of C and G-initial items reported in previous studies (Treiman et al., 2007; Treiman & Kessler, 2019), where virtually no participant produced a soft pronunciation for the control items and the proportion of soft pronunciations for the C-critical and G-critical items were approximately .8 and .15, respectively.

However, there were some unexpected findings. Firstly, the soft pronunciation for C-critical items was not favoured over the hard pronunciation for C-critical items in the rating task, by any of the four groups of participants. However, the proportion of soft pronunciations for the

C-critical items in the naming task was not above .50 from either of the two groups of participants that named these items. As such, it appears that the small subset of C-critical items used in the present experiments did not elicit soft pronunciations as often as in previous studies with larger sets of items (Treiman, et al., 2007; Treiman & Kessler, 2019). As such, no reliable preference of soft pronunciations for C-critical items in the rating task agrees with the naming responses for the subset of items used in the current experiments.

Secondly, the Naming-Rating group (Experiment 1), the Unrelated-Rating and the Naming-Rating-type groups (Experiment 2) failed to reject G-control items with soft pronunciation. Therefore, naming the C and G-initial items before rating them does not seem to explain the unexpected rating behaviour of the Naming-Rating group, because the Unrelated-Rating group did not name these items but still showed the same, unexpected pattern of ratings. Only one group of participants in the current study (the Rating-Only group, Experiment 1) showed the expected pattern of ratings for these items, suggesting that fatigue may explain the difference – the expected pattern of ratings was only obtained from a group of participants that had not named any nonwords before rating them. However, without Bonferroni correction for multiple comparisons, the Naming-Rating-type group would show the expected pattern, in which case, fatigue would not serve as a likely explanation for the differences in ratings found in the different groups, since a group that had named nonwords before rating them also showed the expected pattern of ratings. If this is the case, both fatigue and practice effects can be ruled out as explanations. Another potential reason for the pattern of results is that many skilled readers are not completely clear about which contexts are associated with a soft pronunciation of *g*. This could be because the relationship between preceding vowels and a soft pronunciation of *g* is not highly consistent: for instance, in the word initial position, the proportion of soft pronunciations for *g* followed by *e* is .67 in monosyllabic words and *g* followed by *i* is .25 in monosyllabic words (Treiman et al., 2007)[51]. Thus, the unexpected pattern of ratings in some of the groups in the current study may simply be due to a sample of participants who are not particularly familiar with the nature of the context-dependent pronunciations in words beginning with a *g.* Alternatively, in an attempt to combine the two explanations suggested so far, having completed another task before the rating task may be enough to impair the tenuous knowledge of when context sensitive pronunciation of *g* is appropriate in many skilled readers.

[51] However, Treiman and colleagues show that some groups of words, such as G-initial polysyllabic words with a Latinate suffix, followed by *e* or *i* have very high proportions of soft pronunciations – at 1.00.

Finally, only the Naming-Rating group in Experiment 1 failed to reject C-control items with soft pronunciation as well. While the mean ratings for these items (3.68) were below the critical value of 4 ('probably ok'), it was not sufficiently low to be statistically reliable (even without Bonferroni correction, the observed effect size of 0.31 would still be below the minimum, reliably detectable effect size of 0.34). Given that three groups successfully rejected soft pronunciations assigned to the C-critical items, one of which had also named these items before rating them (the Naming-Rating-type group), it appears that the Naming-Rating group in Experiment 1 happened to consist of participants that either did not know the nature of context-sensitivity in pronouncing *c* onsets well enough or tended to give particularly lenient ratings overall. The latter idea is supported by the numerical trend of the Naming-Rating group accepting the largest percentage of pronunciations and rejecting the lowest percentage of pronunciations, as reflected in the highest sensitivity value and the lowest specificity value of the rating method from the Naming-Rating group, compared to the other participant groups.

Overall, considering the results from the C and G-initial items, the data from the four groups of participants suggest that the results from the rating task are relatively robust. Importantly, the deviations from the expected results may reveal something valuable about the two methods: the rating method might tap into the 'certainty' of participants' PSC knowledge in a more detailed manner than the naming method does. Most participants' first choice of pronunciation (i.e., the naming response they gave) was consistent with the regularities regarding context sensitive *c* and *g* pronunciations, as virtually no participant assigned soft *c* and *g* pronunciations to the control items. Therefore, the results from the naming method suggest that skilled readers know in which contexts the soft pronunciation of *c* and *g* do and do not occur. However, when participants were given options (i.e., alternative pronunciations in the rating task), they appeared to not be as certain about which pronunciations are acceptable, as pronunciations not produced by virtually any participant in the naming task[52] were still rated as acceptable in the rating task. Thus, the results from the rating method suggest there is more uncertainty about the contexts in which the soft pronunciation of *c* and *g* occur in existing words. Of course, an alternative interpretation of the findings is that the rating method is not reliable (see also limitations below). However, the few instances in which discrepancies between the naming and the rating methods were found are outweighed

[52] Note that even though the current study only used a subset of C and G-initial items from Treiman et al. (2007), the naming responses from the two groups in the current experiments also showed a very low proportion of soft pronunciations assigned to the control items – at most at .01.

by the amount of evidence supporting the idea that the two methods capture something similar about the PSC knowledge of skilled readers.

The third aim of the current study was to compare nonword naming and rating responses using a within-subjects design. These direct comparisons of the naming and rating responses in Experiment 1 (Section 6.4.3) and in Experiment 2 (Section 6.8.4) both suggest that skilled readers give higher acceptability ratings to pronunciations they have themselves produced earlier for the same items. The current results thus demonstrate convergence between nonword naming and rating responses, while ruling out the possibility of between-group differences. However, the reason for this convergence could be a result of pronunciation preference for a given nonword, or memory-based fidelity to the pronunciation one has produced previously. Given the number of nonwords each participant named in the naming task (262 nonwords), the latter, memory-based explanation is unlikely to be sufficient. Therefore, it appears that both the naming and the rating method capture something similar about the pronunciation preferences for nonwords, likely reflecting the PSC knowledge skilled readers apply in these tasks. Furthermore, participants with no previous exposure to the critical nonwords (the Rating-Only group and Unrelated-Rating group) still showed a similar pattern of ratings as the participants who had named the same items before rating them, that is, an overall preference for regular pronunciations over irregular pronunciations for the Irregular items with low and high token frequency (Sections 6.4.3 and 6.8.1).

The sensitivity of the rating method, as an average of the sensitivity scores from the four groups of participants was 92%. Thus, the rating method is a reliable source of information about common pronunciations to nonwords.

Turning to the limitations of the present experiments, the most concerning issue was the small sample of C and G-initial items (i.e., six critical items and four control items for each onset). This is problematic as this sample of C and G-initial items is hardly representative of these spelling patterns in the English language. Thus, generalisation of the results from this small subset of items is limited. Furthermore, although the participant sample sizes for these by-subjects analyses were relatively large, some of the between-group differences in the patterns of ratings for the items may be sporadic. As such, interpretation of these group differences should be done with caution. Nevertheless, the rating method was also used with larger sets of items (in particular, the Irregular items reported in Chapter 5, where data loss was not an issue in the same way as it was for the Irregular items reported in Chapter 4), thus

demonstrating that the findings regarding the acceptability ratings cannot be completely disregarded based on these limitations. This issue of small sample of items was a concern already when designing the experiment, but due to practical limitations, that is, the already sizeable sample of items in each experiment (one of the criteria being approximately 50% of fillers in the full set of items presented to the participants), adding more items would have increased the duration of the testing sessions excessively. This is especially the case because any included item would be presented twice in the rating task – once paired with soft and once with hard pronunciation of the onset. Overall, the data for the analyses reported in this chapter come from experiments with a primary goal in investigating the role of type and token frequency in nonword reading (reported in Chapters 4 and 5) and obtaining data for the analyses reported in the current chapter was a secondary goal.

Another potential limitation of the current experiments is that all the data was collected online. As such, even with the Error items that also served as a check for attentiveness, one cannot be sure how much effort participants put into the experimental tasks. Admittedly, collecting data of this sort in a laboratory setting is likely to produce higher quality data, as participants would typically engage and try harder in such a setting compared to online participation. However, evidence for the difference in data quality between laboratory and online studies is mixed (e.g., Chmielewski & Kucker, 2020; Clifford & Jerit, 2014; Kim et al., 2019). Although one of the benefits of the rating method is that it allows collecting data from large numbers of participants with ease, it would still be beneficial for future studies to complement the findings reported here with laboratory-based experiments.

6.9.1 Conclusion

To conclude, in two experiments and total of four participant groups, verbal nonword naming responses were compared to acceptability ratings for nonword pronunciations. These comparisons allowed evaluation of the traditional nonword naming method (verbal naming responses) and a relatively understudied nonword rating method (acceptability ratings for nonword pronunciations). The sensitivity and specificity of the rating method was high (both 90% or above), as assessed against nonword naming responses. The pattern of naming and rating responses to different types of nonwords was mostly comparable. The few instances where the pattern of rating responses diverged from that of the naming responses could mostly be explained, for instance, with different procedures for different groups of participants or with a sample of items used in the current study compared to previous studies.

Importantly, some of the diverging patterns of results may be interpreted as the rating method revealing finer-grained information about skilled readers' PSC knowledge – such as whether and how reliably skilled readers are aware of the contextual cues for soft and hard pronunciations of C and G onsets in words and nonwords. To my knowledge, these evaluations are the first to include such detailed and varied comparisons between the two methods, both as within and between subjects. Overall, the findings reported in this chapter are very promising and suggest that the nonword rating method is a feasible alternative for the nonword naming method in investigations of the PSC knowledge of skilled readers.

Chapter 7 : General Discussion

In this chapter, the findings from the present PhD project are summarised, relative to the aims of the project (outlined in Chapter 1, Section 1.3) and previous literature. Implications of these findings are discussed briefly, before turning to the strengths and limitations of the current PhD work. Finally, I discuss directions for future research, especially regarding further development of the Weighted Segments Pronunciation (WSP) model, which I developed as a part of the current PhD project.

7.1 Summary of main findings

7.1.1 Can the WSP model simulate central tendencies of nonword reading in skilled readers?

The WSP model simulates reading aloud as a process where choice of a pronunciation for a letter string is determined by statistical properties of print-to-sound correspondences (PSCs) of varying sizes. The model's tendency to produce pronunciations based on smaller or larger PSCs is also influenced by weights for the three different parsing styles that are available for the model. The weights can be set by the user, or they can be chosen based on optimisation of the model for a set of words or nonwords. The statistical properties determining the choice of pronunciation can also be set by the user, but in the simulations reported in the current PhD work these properties were consistency and frequency of PSCs, calculated either based on types (WSP-type version of the model) or tokens (WSP-token version of the model). These two versions of the model were tested against human nonword reading data in parallel, to uncover whether a model sensitive to type-based statistical properties (i.e., number of words in which a given PSC occurs) is a better fit to human data than a model sensitive to token-based statistical properties (i.e., the frequency of words and the number of words in which a given PSC occurs).

In Chapter 3, the performance of the WSP-type and WSP-token were assessed against three nonword reading data sets and compared to the performance of the dual-route cascaded model (DRC) and the connectionist dual process model (CDP++) on the same data sets. In these comparisons, the two versions of the WSP model were optimised for the model's vocabulary, which consists of monosyllabic words.

The focus of the comparisons between the computational models was on the vowel pronunciations. Successful simulation of the naming responses in the three data sets required a model to produce regular, standard vowel pronunciations (e.g., the vowel *a* in the nonword *namp* pronounced as in *cat*) and irregular, that is, context-sensitive vowel pronunciations, both regarding the preceding consonantal context, such as an item *wabs* pronounced as in *watch,* and the following consonantal context, such as an item *blange* pronounced as in *strange*.

Most importantly, the WSP-type and WSP-token versions of the model produced reasonably similar naming responses to human participants across the three data sets, demonstrating the model's ability to produce regular, as well as both types of irregular pronunciations to the nonwords, often for the items for which the majority of the human participants had also produced these types of responses. However, the similarity to the most common responses in the human data (human modal responses) was not perfect, as the maximum percentage of matching pronunciations between the WSP model and the human modal responses across the three data sets was 78% (WSP-token model's performance on a data set by Treiman et al., 2003). The lowest performance was seen in a small data set by Andrews and Scarratt (1998, Exp. 1), where the WSP-type only matched 38% of the human modal responses. By comparison, the highest and lowest performance across the three data sets, quantified as the percentage of matches between model output and human modal responses for the DRC model were 74% and 50%, respectively. These percentages for the CDP++ model were 75% and 38%, respectively. These and other comparisons to the human naming responses demonstrated that the two versions of the WSP model fared well against the DRC and the CDP++ models.

Investigations reported in Chapter 3 also revealed that the WSP's vocabulary was not optimal for nonword reading performance, as it led to excessively strong influence of one of the parsing styles (the antibody-coda parsing style), which empirical work has demonstrated to be less influential, compared to other parsing styles (e.g., Andrews & Scarratt, 1998, Exp. 1; Kessler & Treiman, 2001; Treiman et al. 1995). Furthermore, optimising the WSP model with one of the three data sets the model was tested on tended to produce better performance on all of the data sets. The WSP model optimised for the data set by Treiman et al. (2003) was particularly well suited for simulating the naming responses in all three data sets.

The comparison of the WSP-type and WSP-token versions of the model did not provide a definitive answer to the question of whether a model using type or token-based measures of statistical properties better capture human nonword reading, as both versions of the model performed similarly, or slightly better than the other on some of the data sets.

Furthermore, the performance of the WSP model and other computational models on new nonword reading data sets (Chapters 4 and 5) revealed similar levels of overall performance, where the highest percentages of matches to the human modal responses were 44% (the DRC model on the Chapter 4 items with varied token frequencies) and 58% (the WSP-type model on the Chapter 5 items with varied type frequencies). The performance of different versions of the WSP model on both data sets indicated that compared to the human reading responses, the WSP model produced too many irregular pronunciations (word body analogies) to nonwords with word bodies that are always pronounced irregularly in existing words. This was interpreted as consistency of the PSCs having too strong an effect on the competition of different parsing styles in the model.

However, evaluating the performance of the computational models as if the models were individual participants, the DRC and the WSP models compared to the human naming responses as well as an average participant did in the data set by Pritchard et al. (2012). This finding paints a considerably optimistic picture of the performance of these models. Yet, the results reported above suggest that the models are not performing as well as would be expected, if they are to predict the human modal responses.

To conclude, the performance of the WSP model is comparable to that of other computational models. There are several ways to evaluate a model's ability to simulate human nonword reading. Comparing the range of matching pronunciations human participants share with each other to the matches between human participants and computational models showed that the DRC and the WSP perform comparably to skilled readers. However, high accuracy in predicting human modal responses as a criterion of success showed that none of the models performed adequately on any of the testing sets. As the focus on the model evaluations was primarily on the vowel pronunciations, each model failed to perfectly predict which items would receive an irregular or regular vowel pronunciation in the human data. These findings are in concordance with recent evaluations of computational models (e.g., Pritchard et al., 2012; Mousikou et al., 2017; Treiman et al., 2003), which conclude that the current

computational models do not fully capture the patterns of nonword reading responses from skilled readers.

7.1.2 Can the WSP model simulate variability in skilled nonword reading?

The versions of the WSP model described above aimed to simulate the central tendencies in skilled nonword reading. The WSP model can also operate in a variable mode, which aims to simulate variability in nonword reading. Ambiguous nonwords typically receive several different pronunciations from a group of participants. For instance, the nonword *salm*, was pronounced as /s{lm/, /s#m/ and /s#lm/ by 52%, 32% and 5% of participants, respectively, in the Pritchard et al. (2012) data set. The WSP model's variable mode also produces different pronunciation options for nonwords, as well as probabilities for each pronunciation option, which are based on the strength of the competing pronunciations resulting from the three different parsing styles. As with the WSP model described in the previous section, the strength of the different pronunciation options in the variable mode of the WSP model is also based on consistency and frequency of PSCs and weights for the three parsing styles.

The variable output from the WSP model can be extracted using two methods: the raw probabilities method, where the probabilities for the different pronunciation options represent the proportions of participants producing different pronunciation options in the human data, and the multiple simulation runs method, where each simulation run of the same data set represents responses from a single participant. In the latter method, the pronunciation of each item is random, but based on the probabilities for each pronunciation option for a given item. Thus, even though a given simulation run (representing a single participant) might pronounce a segment of one nonword regularly, such as the word body *alm* in *salm*, the same segment in a different nonword might be pronounced irregularly. This type of within-subjects variability is also seen in responses from human participants (e.g., in Pritchard et al., 2012 data set, see also Ulicheva et al., 2021), and it is this variability, the pronunciation choices skilled readers make when encountering an ambiguous letter string, that the multiple simulation runs method aims to simulate. As a set of simulation runs using this method can yield notably different output, five sets of simulation runs were generated to gauge the range of performance the model can achieve, using this method.

The three data sets used to evaluate the WSP model's performance in the previous section were also used to evaluate the performance of the model in the variable mode (Chapter 3). The similarity of the proportions for different pronunciation options from the WSP model's

variable mode and those in the human data were compared by correlating the two proportions for each item in a given data set. When the model's output was extracted using the raw probabilities method, these correlations primarily ranged from moderate to strong, with the weakest correlation at .21 (the third most common pronunciation options in the Pritchard et al. 2012 data set) and the strongest correlations at .77 (WSP-type, for the context sensitivity scores in Treiman et al., 2003 data set). Using the multiple simulation runs method, the performance of the model fluctuated, such that the best performing set of simulation runs for each data set showed an increase in the correlations compared to the raw probabilities method (from .25 to .86), while the worst performing set of simulation runs produced at best moderate correlations in each data set (from .11 to .49).

However, the WSP model's variable output was also evaluated against how well individual participants match human modal responses and other response categories arranged by frequency in the Pritchard et al. (2012) data set. These comparisons suggested that the individual simulation runs of the WSP model perform, on average, comparably well to individual skilled readers, although the WSP model produces somewhat more uncommon responses and responses not produced by any of the participants. Nevertheless, the average performance of the WSP's simulation runs was mostly within the range of the performance of individual skilled readers.

Output from the WSP multiple simulation runs was also compared to another model that produces variable output, namely, a connectionist model by Zevin and Seidenberg (2006), in which the variable output is based on slightly different sets of words the versions of the model were exposed to during training. The output for a set of nonwords from the WSP and Zevin and Seidenberg's model were compared to human naming responses for the same nonwords (Andrews & Scarratt, 1998, Exp. 2). Comparing the pattern of proportions of regular responses given to different types of nonwords in this data set, as well as the pattern of pronunciation variability for the different types of nonwords revealed that both models captured the general pattern found in the human data. However, the Zevin and Seidenberg's model simulated the proportions of regular responses better than the WSP model, whereas the WSP model had a stronger performance in simulating the variability of nonword responses than the Zevin and Seidenberg's model.

Overall, these results demonstrate that some of the variability in human naming responses can be captured by the WSP model and that the individual simulation runs of the model produce typical naming responses, on average, nearly as often as individual participants do.

However, when the model's variable mode is evaluated based on how similar the pattern of proportions for different pronunciation options are between the model output and human data, the level of performance remained relatively low for some data sets. This was the case especially for the Pritchard et al. (2012) data set, for which the human-model correlations were at best .37 (raw probabilities method) for the human modal responses, whereas the corresponding human-human correlations for the modal responses was .83.

Apart from Zevin and Seidenberg (2006), previous modelling work of variability in nonword reading is scarce and has focused on individual differences in developing and dyslexic readers (e.g., Perry et al., 2019; Ziegler et al., 2008). Importantly, the individualised modelling approach taken in these studies as well as the work by Zevin and Seidenberg does not account for within-subjects variability in nonword reading. This is something that, to my knowledge, only the WSP model does.

7.1.3 Does token frequency of PSCs influence nonword processing in skilled readers?

This question was addressed primarily in Chapter 4, although findings from an additional participant group reported in Chapter 6 also bear relevance to this question. I aimed to answer this question by testing whether PSCs occurring in highly frequent words are favoured over PSCs in less frequent words, both in pronunciations assigned to nonwords and in acceptability ratings given to pronunciations assigned to nonwords. The experimental stimuli consisted of nonwords sharing a word body with frequent words (high-items, e.g., *breird,* based on the word *weird*) or less frequent words (low-items, e.g., *bealm,* based on the word *realm*). If token frequency is influential in nonword processing, then more base word congruent pronunciations (i.e., a nonword pronounced to rhyme with its base word) should be assigned to high-items than to low-items, and the acceptability ratings should be higher for high-items paired with their base word congruent pronunciation compared to low-items.

In the naming responses from one group of participants, the incidence of base word congruent pronunciations was higher for high-items than for low-items, and this difference was significant, showing a small effect of token frequency in nonword naming ($dz = 0.24$). However, the analysis was underpowered. The pattern of rating responses was mixed, but two groups of participants, neither of which had named the critical nonwords before rating them,

gave higher acceptability ratings for the high-items than the low-items. One of the groups showed a small, statistically reliable effect of token frequency (*dz* = 0.38), while the other group's pattern of ratings remained marginally significant. A third group of participants showed an opposite pattern of rating responses, which was not statistically reliable.

Taken together, token frequency of print-to-sound correspondences (PSCs) appears to have a small effect on nonword processing: when several plausible pronunciations can be assigned to a letter string, skilled readers are more likely to use pronunciations corresponding to PSCs in frequent words than pronunciations corresponding to PSCs in less frequent words. Although this finding was not obtained consistently, evidence from both nonword naming responses and acceptability ratings given to nonword pronunciations supported this conclusion, at least as a numerical trend from several different groups of participants.

This conclusion is in line with results from previous research (Andrews & Scarratt, 1998; Johnson, 1970), which report some influence of token frequency, but conclude that type frequency is more influential in nonword reading. Due to the limitations in previous research, outlined in Chapter 4 (Section 4.1.1), such as inadequate consideration of consonantal context in vowel pronunciations (Johnson, 1970) or insufficiently separable measures of token and type frequency (Andrews and Scarratt, 1998), the findings from the current PhD project provide, to my knowledge, the strongest evidence to date for the (small) role of token frequency in nonword reading.

7.1.4 Does type frequency of PSCs influence nonword processing in skilled readers?

I aimed to answer this question by testing whether PSCs occurring in several words are favoured over PSCs occurring in a single word, both in pronunciations assigned to nonwords and in acceptability ratings given to pronunciations assigned to nonwords. The nonword stimuli consisted of nonwords that either shared a word body with several existing words (Irregular-Many items) or with a single existing word (Irregular-Single items), while the token frequency of the word body segments was comparable between the two item groups. If type frequency is influential in nonword processing, then Irregular-Many items should receive more base word congruent pronunciations than Irregular-Single items, and the acceptability ratings should be higher for the Irregular-Many items paired with their base word congruent pronunciation compared to Irregular-Single items.

These predicted differences were confirmed statistically, both for pronunciations assigned to the nonwords (*dz* = 1.07) and for acceptability ratings given to the nonword pronunciations

(*dz* = 1.78). One group of participants was tested and showed this clear pattern of naming and rating responses. Thus, the investigations in Chapter 5 show that type frequency of PSCs has a large effect on nonword processing – when a letter string has several plausible pronunciations, skilled readers are more likely to pronounce it according to PSCs in several words than according to PSCs in only a single word.

Considering the influence of type frequency and token frequency together, it appears that extraction of PSCs from experience with reading is mostly driven by type frequency, but may be slightly enhanced or hindered by converging or diverging PSCs occurring in highly frequent words. That is, pronunciations associated with a particular spelling pattern in several words are more readily available in the reader's PSC knowledge, when encountering this spelling pattern. If some of the words in which this PSC occurs are also highly frequent, this should slightly increase the likelihood of using the pronunciation even further. By contrast, if there is a highly frequent word with an alternative pronunciation for the same spelling pattern, this should slightly reduce the likelihood of using the PSC supported by the several existing words.

These findings are also compatible with the results from the previous empirical work (Andrews & Scarratt, 1998; Treiman et al., 1990), which suggests that type frequency plays a larger role in nonword reading than token frequency does. As stated in the previous section, the findings reported in the current PhD project avoid some of the limitations of the previous studies, and as such provide compelling evidence of the importance of type frequency of PSCs in nonword reading.

7.1.5 Can PSC knowledge of skilled readers be assessed using a nonword rating method instead of a nonword naming method?

In Chapter 6, I evaluated the nonword rating task, in which participants are presented with a written form of a nonword and an aurally presented pronunciation assigned to it, after which the participants give an acceptability rating for how well the pronunciation fits the written form of the nonword. The ratings obtained from the rating task were compared to nonword naming responses for the same items, in several ways.

Firstly, the specificity and sensitivity of the rating method was assessed by inspecting the acceptability ratings given to nonwords that were paired with an implausible pronunciation (such as *dwal* pronounced as /jEsts/) and ratings to nonwords paired with the most common naming response amongst participants (such as *glatt* pronounced as /gl{t/). Low acceptability

ratings for implausible pronunciations and high ratings for common pronunciations would indicate high specificity and sensitivity of the rating method, respectively.

Secondly, rating responses in the rating task were compared to the naming responses in the naming task as a within-subjects comparisons. In these comparisons, the same group of participants first named the experimental nonwords and subsequently rated the same items, as these items were paired with pronunciations containing either a regular or an irregular vowel pronunciation (such as an item *bealm* paired with pronunciations /bilm/ or /bElm/).

Thirdly, the pattern of rating responses from participants in my experiments were compared to that of naming responses from previous studies (Treiman et al., 2007, Exp. 1; Treiman & Kessler, 2019, Exp. 1). These previous studies have shown a relatively clear pattern of naming responses to nonwords with *c* and *g* onsets, divided into critical items (onset *c* or *g* followed by vowels *e* or *i)*, as existing words with these spelling patterns often have a soft pronunciation (as in *cell* and *gene*) and control items (onset *c* or *g*, followed by vowels *a, o* or *u),* as existing words with these spelling patterns always have a hard pronunciation (as in *cat* and *game*). Most importantly, the proportion of soft pronunciations for the critical nonwords were approximately .80, (onset *c*) and .15 (onset *g*), while the proportion of soft pronunciations for the control items was at most .01, suggesting that skilled readers apply these context sensitive pronunciations to new items based on how these types of pronunciations occur in existing items (critical versus control items), although the incidence of soft pronunciations was not as high for the critical items as would be expected based on how often they occur in the existing words. Thus, this pattern of results was expected also in the acceptability ratings, where a subset of critical and control nonwords from Treiman et al. (2007) study was paired with both soft and hard pronunciations each.

The most important findings from these evaluations were that the rating method appears to have a high level of sensitivity and specificity (both at 92%, as an average score from four different groups of participants). Additionally, the within-subjects comparisons showed reliable convergence between the naming and rating responses skilled readers give to nonwords: participants rated regular vowel pronunciations as more acceptable than irregular pronunciations for nonwords they had previously assigned a regular pronunciation (dz = 1.64, Chapter 6, Exp1; dz = 1.67, Chapter 6, Exp. 2). Similarly, participants favoured irregular vowel pronunciations over regular pronunciations for items in the rating task for which they

had previously assigned irregular pronunciations in the naming task (*dz* = 1.22, Chapter 6, Exp1; *dz* = 2.15, Chapter 6, Exp. 2).

The comparison of the pattern of rating responses to that of naming responses from previous studies revealed, most importantly, that the same general pattern was seen in both types of responses. The rating responses for items with *c* and *g* onsets were analysed from four groups of participants, two of which had also named these items before rating them. Although some divergence from the pattern of naming responses from the previous studies were found, most of these could be explained, for instance, with the sample of nonwords used in my experiments compared to the previous studies. Furthermore, some of the discrepancies between the rating and naming responses could be interpreted as a strength of the rating method. This possibility was based on the findings that even though the incidence of soft pronunciations for control items with *g*-onset was at most .01 – both in the naming responses reported in previous studies and in the naming responses from the two groups of participants in my experiments – the *g*-control items paired with soft pronunciations in the rating task were not rejected reliably by at least two groups of participants[53]. As such, while the naming responses alone suggest that skilled readers' PSC knowledge allows them to differentiate between appropriate and inappropriate contexts for soft pronunciation of *g*, the rating responses suggest that this is not the case. The rating method may thus provide more detailed information about the PSC knowledge of skilled readers than can be obtained with the naming method.

The evaluation of the rating method reported in Chapter 6 is in line with the handful of previous work regarding an alternative for the traditional nonword naming method (e.g., Treiman et al., 2003; Treiman & Zukowski, 1988), namely, that the rating method produces the same general pattern of results as the naming method does. The findings from the current PhD work demonstrate the feasibility and potential strengths of the rating method in investigating human nonword reading, much like Gubian et al. (2022) demonstrated the usefulness of this method in evaluating computational models.

[53] Depending on how conservatively these analyses are protected against increased type I error rate due to multiple comparisons

7.2 Implications of findings

7.2.1 Findings regarding computational modelling

The computational investigations presented in this thesis have theoretical implications regarding generalisation in reading aloud. Most importantly, the relative success of the WSP model compared to the DRC and CDP++ models suggests, broadly speaking, that the mechanisms by which these models convert letter strings into speech sounds are comparably fitting explanations for the processes involved in human nonword reading. However, each model has particular areas of generalisation in reading aloud that they simulate relatively well and areas that they struggle with. These strengths and weaknesses of each model should inform further model development.

For instance, the CDP++ and the WSP models utilise more varied set of statistical properties of PSCs compared to the DRC model, where only grapheme-sized PSCs with the highest type frequency are applied in nonword reading. Consequently, the CDP++ and the WSP models produced much more accurate responses to items with irregular word bodies than the DRC model does, although, the two models also overestimated the incidence of the irregular pronunciations. As another example, due to the CDP++ and WSP's sensitivity to a richer set of statistical properties of English, across varied unit sizes, nonwords with *th*-onset were often pronounced as in *the* by these models, rather than as in *think,* which is the pronunciation most human participants used. As suggested in Chapter 3 (Section 3.5.3), this issue may be avoided by considering the part of speech information of the words that are included in the WSP's vocabulary that the model's PSC knowledge is based on, or vocabulary and the simple PSCs that the CDP++ model is trained on.

The model comparisons reported in the current dissertation also add to the discussion of what types of models may be the most informative in investigations of reading aloud. Broadly speaking, the two symbolic models included in these comparisons, the DRC and the WSP models, had the strongest performance across a variety of data sets. This is even though the success of the two models was based on quite different patterns of output: the DRC model outperformed other models on any set of nonwords where regular pronunciations were prevalent amongst human participants (e.g., the data set by Pritchard et al., 2012; the items in Chapter 4), whereas the WSP model tended to outperform other models when irregular pronunciations were common in the human responses (e.g., the data set by Treiman et al., 2003; the items in Chapter 5). Compared to connectionist models, symbolic models have the

benefit of a clearer account of the exact mechanisms by which these models produce reading aloud output. For instance, the source of the high incidence of irregular responses to Chapter 4 and 5 nonwords by the WSP model could be identified as the overwhelmingly strong word body PSCs, which resulted in the output of the model being mostly based on the word body parsing style. By contrast, some patterns of output from connectionist models (such as the CDP++ and the Psim1 model by Plaut et al. (1996, Sim. 1)) are difficult to explain in terms of the way these models operate. It should be noted, however, that the set of models included in the comparisons here is by no means comprehensive, and the performance of several other connectionist models would undoubtedly be informative regarding the issues investigated in this dissertation.

The evaluation of the computational models in the current PhD work also highlights the importance of considering what type of testing sets are used for model comparisons. For instance, the explicit inclusion of the antibody, word body and small segment parsing styles in the WSP model proved very useful in simulating human nonword reading responses to a set of items where pronunciations corresponding to each parsing style were required (Treiman et al., 2003). However, the same feature of the WSP model was less helpful for a data set where a vast majority of human naming responses consisted of mostly context insensitive pronunciations (Pritchard et al., 2012). While these sets of nonwords may test different aspects of a model's reading aloud performance, they appear to also reflect list context effects in nonword reading, which the models included in these comparisons are not designed to simulate. For instance, there is evidence to suggest that the composition of items in the data set by Pritchard et al. (2012) encourages context insensitive reading strategy in human participants (Perry, 2018). The current PhD work added to this evidence by demonstrating how this data set consists of mostly regular bodied items and was best simulated by models that rely on regular pronunciations (the DRC and the WSP model optimised for the Pritchard et al.'s set), at the expense of these models' performance on other data sets where irregular pronunciations were more prevalent. Therefore, comparison of computational models should always be accompanied with a consideration of the type of nonwords used as the testing set, particularly regarding what other behaviours than those related to skilled readers' PSC knowledge the responses to a particular set of nonwords may reflect.

Finally, findings regarding the variable output produced by the WSP model compared to the model by Zevin and Seidenberg (2006) demonstrated that model's simulating within-subjects

variability (the WSP model) and those simulating between-subjects variability (Zevin and Seidenberg, 2006) may produce output similar to that from a group of human participants. As such, more detailed comparisons, with data sets that allow separation of the two types of variability, are needed for further computational investigations of variability in nonword reading.

7.2.3 Empirical findings

The empirical findings regarding the role of token and type frequency of PSCs in nonword processing have theoretical implications. The role of token frequency in nonword reading is not clear in some computational models of reading, such as the CDP++ model[54], or it is not included at all, as is the case with the DRC model. The current findings thus indicate a need for modifications, if these models are to simulate human nonword reading comprehensively.

These findings also bear relevance to further development of the WSP model – while the evaluation of the model in Chapter 3 did not provide a clear answer to whether the statistical properties of the PSCs should be based on type or token frequency, the empirical findings suggest that both are needed, with a likely stronger influence of type frequency. It thus appears that the most appropriate choice are measures based on summed token frequency, which includes both type and token frequency information. This particular measure has also been shown to be more important in word reading than type frequency alone (e.g., Jared, 1990).

7.2.3 Findings regarding methodology

The evaluation of the nonword rating method bears relevance to future research in reading aloud. As a feasible alternative to the traditional nonword naming method, the rating method has clear strengths that allow it to both avoid some of the shortcomings of the naming method, as well as potentially provide information that cannot be obtained using the naming method.

[54] For instance, even though the CDP++ model's efficiency of learning PSCs should be sensitive to token frequency, the CDP++ model did not show an effect of token frequency in Chapter 4, but rather a numerical trend to the opposite direction.

7.3 Strengths and original contribution

7.3.1 Computational investigations

The WSP model, developed as a part of the current PhD project, offers a flexible approach to investigating the role of different statistical properties of the writing system in reading aloud, as well as a means for investigating the properties of different human nonword reading data sets. For instance, the WSP model allows for the isolation of the influence of different statistical properties of the writing system, and thus enables more focused investigations of specific properties of PSCs on nonword reading.

While the explicit inclusion of three different parsing styles, based on antibody-coda, onset-word body and onset-vowel-coda segments, in the WSP model is not new (e.g., Norris, 1994), the WSP model differs from the previous modelling work in important ways. For instance, as opposed to the Multiple-levels model by Norris (1994), where the competition between pronunciations associated with different parsing styles is based on a pre-determined hierarchy (such as the word body parsing style being favoured over the antibody parsing style if the two parsing styles result in different pronunciations), the competition in the WSP model is resolved based on the statistical properties of the PSCs in each parsing style. The WSP model also includes another factor that affects this competition, namely, weights applied to each parsing style, which aim to simulate the global tendency to favour certain parsing styles over others. This feature of the WSP model is another source of flexibility: for instance, large weight for the small segment parsing style produces reading behaviour similar to that of the DRC model. Therefore, different tendencies of parsing a letter string can be explored with the WSP model, including extremes such as the 'GPC-sized segments only' parsing style embodied in the DRC model.

The WSP model also produces variable output to letter strings, with the aim of simulating variability in nonword reading, which is based on probabilities derived from statistical properties of the PSCs of varying sizes, corresponding to the three parsing styles. Unlike previous modelling work regarding variability in the form of individual differences in nonword reading (Perry et al., 2019; Zevin and Seidenberg, 2006; Ziegler et al., 2008), the approach I took focuses on the within-subjects variability. In addition to the attempt to simulate variability in skilled readers' nonword naming responses, I tested this underrepresented feature in computational models. In doing so, I presented new ways to evaluate computational models that produce variable output, such as using human-human

correlations as a benchmark for human-model correlations of the proportions of different pronunciation options (see Chapter 3, Section 3.3.1.1).

Additionally, I presented several detailed comparisons of some of the current computational models (the newest versions of the DRC and CDP++ models) on three existing nonword naming data sets, as well as on two new data sets (Chapters 4 and 5).

7.3.2 Empirical investigations

The main empirical findings were that token and type frequency of PSCs both play a role in nonword processing, but type frequency of PSCs appears to have a larger influence. Importantly, these findings do not suffer from some of the shortcomings of previous research aiming to clarify the role or the relative importance of these properties in nonword reading (cf. Andrews and Scarratt, 1998). Therefore, the findings from the current PhD project are a valuable addition to answering the question regarding the role of type and token frequency in nonword reading.

7.3.3 Methodological investigations

The evaluation of the nonword rating method reported in Chapter 6 goes beyond the previous work (e.g., Johnson, 1970; Treiman et al., 2003; Treiman & Zukowski, 1988) in several ways. Most notably, the current work provided detailed comparisons between the nonword rating and the nonword naming responses for several types of nonwords, from several groups of participants. These evaluations also included within-subjects comparisons, thus ruling out the potential for between-group differences causing discrepancies in the rating and naming responses. Yet, rating responses without previous naming of the same items, or any items at all, were also investigated, providing information about nonword rating behaviour without potential influence from previous naming of the same stimuli. The specificity and sensitivity of the rating method were also quantified relative to nonword naming responses. While this approach for measuring specificity and sensitivity of the rating method was inspired by the work of Gubian et al. (2022), to my knowledge, this type of assessment of the rating method for the purposes of investigating nonword processing in skilled readers has not been conducted before.

7.4 Limitations

Several limitations of the WSP model were discussed in Chapter 3 (Section 3.5). Furthermore, an important limitation was identified in Chapters 4 and 5, namely, that the

model produced far higher incidence of irregular (word body analogy) pronunciations for items with unique word bodies than what was found in the human data. This behaviour of the model, especially when the model was optimised for its vocabulary, was due to the consistency of the word body sized PSCs having too strong an influence in the competition between different parsing styles, thus leading to most of the pronunciations of the model being based on word body parsing style. See Section 7.5 for further consideration of this limitation.

It is also worth mentioning that the vocabulary-optimised versions of the WSP model had a surprisingly large global tendency to use the antibody parsing style in reading aloud. This is not in line with empirical evidence (e.g., Kessler & Treiman, 2001; Treiman & Zukowski, 1988), and warrants further investigation on whether the model, the vocabulary or the optimisation procedure need to be adjusted.

Experiments aiming to investigate the role of token and type frequency in nonword processing (Chapters 4 and 5, respectively), utilised materials where all (Chapter 4) or half (Chapter 5) of the nonwords had unique word bodies, i.e., word bodies that only occur in a single existing word in English. Using these types of nonwords had the benefit of keeping the consistency of the word body sized PSCs identical between the different groups of nonwords that were compared (i.e., the word body sized PSCs in the words that the nonwords were based on were comparable). However, it is possible that these singleton or unique items are special, and the process of assigning a pronunciation for them is different from other nonwords, which embody more typical spelling patterns. In other words, although the evidence I provide for the role of token and type frequency of PSCs in nonword naming is a clearly beneficial addition to the current literature, more definitive conclusion about the influence of these properties in nonword naming requires evidence from more representative samples of nonwords.

The investigations of the token frequency of PSCs in nonword processing relied on the difference of token frequencies of the words that the critical nonwords were based on. However, this difference may not have been sufficiently strong; that is, the strength of the manipulation in this study may explain why the effect of token frequency was not found consistently, or why the effect was small, when found. The requirements for the stimuli in this experiment greatly reduced the number of available items that could be used. Further data loss due to the participants' insufficient familiarity with the base words for the

experimental items was also problematic, and it may partly explain why the findings regarding the role of token frequency in nonword processing were somewhat inconsistent. This limitation is also relevant in regards to comparing the effects of type and token frequency in the current PhD work. While it appears that type frequency of PSCs has a stronger influence in nonword processing, I cannot rule out the possibility that a stronger manipulation of token frequency would have made this difference less clear.

Turning to the evaluation of the nonword rating method, apart from the limitations discussed in Chapter 6 (Section 6.9), the focus of these investigations was on context sensitive onset (the *c* and *g* items) and context sensitive vowel pronunciations (irregular word body items with varied token or type frequency). As such, the contrast between context sensitive and insensitive pronunciations in the audio recordings paired with the written form of the same nonword item, for instance, *gealth* paired with pronunciations /gilT/ and /gElT/, may not have been the sole focus of the participants. In other words, the acceptability ratings from the participants may have been based on how appropriate they deemed the onset, vowel and coda pronunciations for each item, not only the acceptability of the segments that were of interest in these experiments. Or, indeed, the participants may have based their ratings on only one of these segments. This is not a concern for the majority of the experimental items, as the segments that were not focused on in these investigations were typically not ambiguous and thus the acceptability of the pronunciation of these segments should be deemed equally high amongst most skilled readers. Nevertheless, this aspect should be taken into account in future investigations using the rating method, particularly regarding stimuli construction. One of the special characteristics of the nonword rating task is that we do not know what criteria participants use in their acceptability ratings, or if they ignore parts of the stimuli, a feature that is not an issue with the nonword naming task, where producing a pronunciation requires consideration of the whole letter string.

7.5 Future directions

7.5.1 WSP model's knowledge of PSCs

Several avenues for improving the WSP model were considered in Chapter 3 (Section 3.5). As briefly discussed in this section, bringing the PSC knowledge of the WSP model closer to that of skilled readers would likely help with increasing the similarity between the output from the model and human reading responses. Three ways of achieving this are worth considering. Firstly, the model's vocabulary could be expanded to upscale the types of words

the PSC knowledge is based on – and by extension, the type of letter strings the WSP model can read – to disyllabic and multisyllabic words.

Secondly, to ensure the vocabulary on which the PSCs available to the model are based on consist of words likely known by an average reader, the words included in the model's vocabulary could be restricted to only words with high enough prevalence (i.e., known by high enough percentage of skilled readers). I have made some attempts to modify the WSP model's PSC knowledge, based on a vocabulary that only included items known by 90% of skilled readers, based on the word prevalence data base by Brysbaert et al. (2019). However, further testing is required to uncover the potential benefits of this modified vocabulary and PSC knowledge of the WSP model. These modifications would be the most helpful if the WSP's vocabulary was larger, based on disyllabic words as well, as this would reduce the likelihood of losing PSCs that an average reader knows, even though they might not base their knowledge of these PSCs on an uncommon monosyllabic word or words.

Thirdly, it may be helpful to include proper nouns, such as names of people and places, in the WSP's vocabulary. These items are also part of the written material skilled readers encounter and can learn PSCs from. In my view, proper nouns should thus be considered as items representing PSCs, just like all the other letter strings. For example, some common names may include PSCs that are uncommon, and thus including them may change the statistics for this particular PSC, which might not be consistent enough to be included as a pronunciation option in the model unless proper nouns were included. These three modifications to the WSP's vocabulary would likely benefit both the model's deterministic mode and the variable mode.

7.5.2 Processes involved in print-to-sound conversion

As suggested in Chapter 3 (Section 3.5.3), including part of speech information in the WSP's vocabulary may be beneficial for addressing the issue of uncommon pronunciations, such as items with *th*-onset. Campbell and Besner (1981) offered a potential explanation for why skilled readers may assume that nonwords read aloud in isolation (rather than in a sentence context) are nouns, adjectives and verbs rather than function words (this, that, therefore, etc.). They suggest that because function words are highly frequent and a limited set of words in a language, a skilled reader wouldn't expect to encounter new function words, whereas coming across a new noun is more plausible. Therefore, constructing a computational model that

converts text into speech sounds may require more global knowledge of the language than mere statistics of the print-to-sound correspondences.

In line with this idea, other linguistic information is likely required to improve the WSP model, especially when attempting to expand the model's reading abilities beyond monosyllabic letter strings. Successful reading of disyllabic and multisyllabic words and nonwords also require knowledge about syllabic stress (e.g., Mousikou et al., 2017). This feature might be relatively straightforward to implement in the WSP model, as stress of syllables for multisyllabic words is readily available in different data bases, such as WebCelex. The first attempt at incorporating stress information into the WSP's PSC knowledge would involve extracting the most consistent pronunciations for different segments, as in the current versions of the WSP, but additionally including the stress, so that, for instance, an initial *o* as a stressed segment corresponds to /5/ as in *open,* but the same segment without stress corresponds to /@/ as in *omit.* This approach relies on the assumption that certain spelling patterns are associated with stress while others are not. Indeed, Rastle and Coltheart (2000) developed an algorithm for disyllabic nonword reading, which relies on this assumption. In this approach, naming a letter string is preceded by morphological decomposition and a set of affixes categorised as stress taking or not, which then defines on which syllable the stress is assigned.

On a related note, morphological information is likely needed for the WSP model to handle morphologically complex items, which are much more common in multisyllabic items compared to monosyllabic ones. Morphological knowledge is needed because morphological decomposition likely precedes word recognition (e.g., Rastle et al., 2004), which would also suggest that this process takes place before naming a morphologically complex nonword. My preliminary attempts at implementing this in the WSP model would also involve a knowledge base of prefixes and suffixes, similar to Rastle and Coltheart's (2000) approach.

Additional linguistic information implemented in the model could also provide a solution for some non-adjacent letter context effects reported in the literature, such as a Latinate suffix increasing the likelihood of soft pronunciation for nonwords with onset *c* or *g* (Treiman et al., 2007, Exp. 2). In addition to morphological decomposition, a classification of the resulting suffixes based on their source (e.g., Germanic or Latinate), would be needed to produce an advantage for soft onset pronunciation option with certain suffixes. However, this solution is akin to the manner in which the WPS handles context sensitive c and g onsets – i.e., as a

special case, supported by an additional feature of the model, rather than the general process of parsing a letter string. As such, a more global solution for the model to produce context sensitive pronunciations would be more informative.

Finally, expanding the WSP model beyond monosyllabic letter strings requires additional considerations regarding segmentation of disyllabic or multisyllabic letter strings. For instance, it is not immediately clear whether *m* in *lemon* or *tt* in *otter* should be parsed as an onset of the second syllable or as a coda of the first syllable. In another model designed to simulate disyllabic reading, the CDP++ model (Perry et al., 2010), this issue was solved by onset maximisation, such that consonants between two vowels are assigned as onsets whenever possible. I would start by applying this same principle, and inspect whether additional linguistic constraints might be necessary for a successful segmentation of multisyllabic letter strings.

The general approach taken in the WSP model may also be extended to other alphabetic writing systems. It should be possible to extract similar, varied sized PSCs from the vocabulary of the relevant language and optimise the weights for the vocabulary. However, different parsing styles and their competition may be redundant for models of shallow orthographies, such as Italian or Finnish, where single-letter correspondences are largely sufficient for accurate print-to-sound conversion of existing words. As such, the pronunciations most consistently associated with a particular orthographic segment would nearly always be the same regardless of how long this segment is in these writing systems. Extending the WSP model to non-alphabetic languages, such as Mandarin Chinese, would require more extensive modifications. This is because the characters or features of characters in such writing systems also include semantic information, while the focus in the current WSP model is on the relationship of orthography and phonology, disregarding semantic or morphological information.

7.5.3 Competition of different pronunciation options

It became apparent in Chapters 4 and 5, that the consistency of word body sized PSCs has too strong an influence on the WSP model's output, compared to human nonword naming responses. As the word body sized segments of the experimental stimuli were perfectly consistent in existing words (consistency value 1), this resulted in the WSP model's output being based on the word body parsing style almost exclusively. In these chapters, it was also found that frequency measures of the vowel segments were associated with the proportion of

word body-based naming responses to nonwords (Sections 4.3.4.2 and 5.3.1.3). This relationship was such that if a regular or context insensitive pronunciation of the vowel segment occurred in several words or highly frequent words, nonwords containing this vowel segment were less likely to be pronounced irregularly as a word body analogy.

Therefore, one potential modification of the WSP model would be to reduce the influence the word body consistency has on the final output depending on the frequency of the vowel segment. Although the investigations in Chapters 4 and 5 did not include antibody segments, I implemented this modification such that it reduced the influence of both body and antibody segments relative to the frequency of the vowel segments. This was based on the assumption that if high certainty about the pronunciation of the vowel segment reduces reliance on larger unit PSCs, this should apply in both directions: that is, for both the preceding and the following consonants. Further testing of the WSP model is needed to see whether this modification is beneficial for the overall performance of the model (including potential re-optimisation of the model with this modification in place).

In addition to the influence of consistency of word body sized PSCs, the consistency of onsets and codas appeared to influence the competition between the three parsing styles in unexpected ways. Sometimes consistent onsets and codas would lead to a small segment parsing style winning, even though each parsing style produced the same pronunciation for these segments (as discussed in Chapter 5). It appears that a way forward from this issue would be to base the competition between the three parsing styles on only the strength of the segments these parsing styles would produce a different pronunciation for. For instance, if the three parsing styles produce a different vowel pronunciation, then only the strength of the antibody, the word body and the vowel segments would compete with each other, excluding the strengths of onsets and codas from each parsing style.

7.5.4 Simulating variability in nonword reading

As mentioned in Chapter 3 (Section 3.5), the WSP model's ability to simulate variability in nonword reading is promising, but it requires more work. While some progress was made in simulating within-participants variability, the between-participants variability was neglected. This feature could be added by optimising the model with slightly different sets of items, similar to the idea employed in Zevin and Seidenberg (2006).

Another interesting approach for testing how well the modified WSP model could simulate between-subjects variability in nonword reading would be to compare the model's output to

detailed nonword naming responses (i.e., including responses from every participant for every item) from participants who received reading instruction focusing on GPCs and participants whose reading instruction focused on larger segments (as in Thompson et al., 2009). Here, the groups differed in their general tendency to parse letter strings into smaller and larger segments. Similar general tendencies can be produced by the WSP model, by either advantaging the larger or the smaller segment parsing styles.

Furthermore, as the WSP model is one of the few models producing variable output to a set of stimuli, the best practice for evaluating such a model is still lacking. For instance, if the empirical evidence from Ulicheva et al. (2021) is considered, the responses from even the same group of participants for the same nonwords may be quite different from one testing session to the next. As such, detailed naming responses (from every participant and for every item) from different groups of participants for the same item set would be needed to test whether variability produced by the WSP model's variable mode (multiple simulation runs method) is similar to the variability found between different participant groups.

However, investigations of variability in nonword reading require reliable differentiation between intended naming responses and mispronunciations or lapses of concentration. Procedures of collecting naming responses may in part help reduce the number of careless errors, such as instructing participants to re-pronounce an item if they think they mis-pronounced it the first time. Perhaps one way to distinguish between truly variable, intentional responses from careless errors is designing experiments with a large enough number of items that share a particular spelling pattern (e.g., *plall, rall, scrall* to tap into alternative pronunciations of the word body *all*). If a participant pronounces the same spelling pattern repeatedly in different ways, such that each alternative pronunciation is produced at least twice or three times, this would be more suggestive of deliberate pronunciations, based on what a given participant considers plausible, rather than careless errors. Considering group-level variability, pronunciations produced by several participants should generally be considered intentional.

It is also worth considering whether accurate simulation of central tendencies should be achieved before attempting to simulate variability in nonword reading. The cognitive processes involved in reading are undoubtedly more complex than the mechanisms included in current computational models. The models' inadequate success in simulating central tendencies may be, at least in part, due to factors that influence human reading behaviour but

that are not included in the models. In my view, it is not necessarily the case that certain aspects of human reading need to be successfully captured before others can be simulated. In fact, considering some additional aspects of nonword reading may also help discovering ways to improve the models' ability to simulate central tendencies. In addition to variability in nonword reading, attempts to simulate reading of disyllabic nonwords have also been made before monosyllabic reading is fully mastered by a particular model or framework (e.g., the CDP++ model by Perry et al., 2010). Findings from disyllabic reading behaviour may inform us about the general constraints in reading that cannot be discovered based on monosyllabic items only (such as stress assignment and influence of non-adjacent letter contexts). Similarly, endeavours of simulating variability in nonword reading may reveal aspects about the reading process as a whole that facilitate more comprehensive characterisation of reading, including more accurate simulation of central tendencies.

7.5.5 Other considerations

Throughout the thesis, I have contrasted the large segment and the small segment parsing styles, which are also a central part of the WSP model. The support for the idea of these parsing styles being utilised by skilled readers comes from empirical studies showing that, for instance, the nonword item *moup* is pronounced according to the GPC rules (/m6p/) by some participants and as a word body analogy (/mup/) by other participants (Andrews and Scarratt, 1998). However, it is not clear whether the latter, irregular or word body analogy responses are indeed a result of parsing the letter string and applying PSCs based on larger segments, or whether the vowel pronunciation in the word body analogy pronunciation is a context sensitive pronunciation, based on a context-sensitive rule or PSC: *ou* followed by *p* is pronounced as /u/. To my knowledge, the difference between larger segment or context sensitive rules/correspondences has not been investigated directly in empirical work. One way of trying to tease apart the two would be by using nonword items with word bodies where in addition to the vowel, the coda is also pronounced in a non-standard way, such as nonwords sharing a word body with *folk* and *yolk*. If these nonwords are pronounced with the non-standard vowel and coda, /5/ and /k/, this would suggest that word body sized PSCs are used. If these nonwords are pronounced with a non-standard vowel, /5/, but with a standard coda, /lk/, this would suggest utilization of context sensitive PSCs instead.

The issue with this type of experiment is that there might not be many items in the English language that contain the required characteristics. Additionally, skilled readers may utilise

both types of PSCs in their reading. Nevertheless, context sensitive PSCs rather than PSCs based on larger segments could be implemented in the WSP model: while the competition of the parsing styles could still determine which vowel pronunciation is applied, the pronunciation of the onset and the coda could be based on pronunciations of the grapheme-sized segments. This modification would also address some of the issues relating to the pronunciation options produced by the variable mode of the WSP model. Comparing the performance of this 'context sensitive' version of the WSP model to the versions of the model described and tested in the present dissertation may prove useful in trying to clarify whether skilled readers primarily utilise context sensitive PSCs or PSCs based on larger orthographic segments.

However, effects of letter context on pronunciation of letter clusters are not restricted to adjacent letter contexts, such as the influence of onset or coda on the vowel. Non-adjacent contexts may also influence pronunciations. Treiman et al. (2007, Exp. 2) demonstrated that a Latinate suffix in disyllabic nonwords increased the likelihood of soft pronunciation assigned to the onset *c* or *g* of these nonwords (*cebic* or *gebic*), compared to native suffixes (e.g., *cebful* or *gebful*), which is in concordance with these types of onset pronunciations being more common in existing words with Latinate suffixes. While context sensitive pronunciations in monosyllabic nonwords can be explained reasonably well with unit size of PSCs, this explanation does not accommodate non-adjacent contextual effects. Thus, when moving beyond monosyllabic nonword naming, the approach reliant on varied unit size needs to be reconsidered. Whether a possibility of combining the two approaches exists, is yet to be seen. For instance, implementing a context sensitive, rather than varied unit size based approach for modelling nonword reading requires a mechanism for connecting the adjacent or non-adjacent parts of the letter string that define the pronunciation of certain parts of a nonword. I do not have a proposed implementation for this, but it would appear that a combination of varied unit size and context sensitive approach may also be an option – this is particularly so if further empirical work demonstrates that the influence of non-adjacent letter contexts are limited to only a few cases. In this situation, the morphological information required for morphological decomposition of polysyllabic letter strings may also include classifications of the sources of suffixes, while most other context sensitive influences would still be achieved by varied unit size.

Turning to the empirical investigations, two important statistical properties of the writing system, the regularity and the consistency of PSCs, were not investigated empirically in the

current PhD project. This was mainly because the question of different types of frequency measures was more pressing, as an answer to this question would also bear relevance to which type of frequency the measure of consistency should be based on. Although a large proportion of findings in the nonword naming literature could be interpreted as the influence of consistency (especially if this applies to varied unit sizes, in which case regular pronunciations would also be explained by consistency), there are some findings that are not explained fully with consistency (e.g., Andrews and Scarratt, 1998, discussed in Chapter 1, Section 1.1.2). Although there is evidence to suggest that regularity and consistency are separate properties from word reading studies (e.g., Andrews, 1982; Jared, 2002), whether and how these properties relate to each other in nonword reading is still unclear. These investigations, left for future research, could also inform further development of the WSP model. This is so particularly given that the influence of the consistency of PSCs on the model's output appeared to be problematic in some versions of the model.

Finally, further evaluation of the nonword rating method is needed to both 1) to further clarify under which circumstances and for what purposes should the nonword rating method be used instead of the nonword naming method and 2) to fine-tune the form in which the nonword rating task would provide maximal amount of information. However, previous literature provides some constraints for what type of modifications may be the most promising to explore. For instance, it appears that scales with seven, nine or 10 response categories should be preferred, as indicated by the highest participant preference, discriminatory power, reliability and validity of these types of scales over those with fewer or higher number of categories (Preston & Colman, 2000). Additionally, previous research using a multiple-choice method in nonword processing (Treiman et al., 2003, Exp. 2) suggests that a neutral response option is likely not beneficial in this type of task. With these constraints in mind, instead of the six-point, verbally labelled response categories used in the investigations in Chapter 6, a different number and form of response categories, such as a numeric, 10-point scale, may produce more detailed information about the pronunciation preferences and print-to-sound knowledge skilled readers have.

7.6 Concluding remarks

Reading aloud new words requires an ability to generalise linguistic knowledge acquired via experience in reading. For decades, empirical investigations and the development of verbal theories and computational models have been used with the aim of uncovering the cognitive

processes involved in the human ability to generalise from previous reading experience. One of the biggest challenges and a benchmark test for computational models of reading is whether they can simulate this generalisation, that is, whether they read new words or pronounceable letter strings (nonwords) in the same way as humans do.

In this PhD project, I aimed to shed light on the cognitive process of generalisation in reading, via empirical and computational investigations of nonword reading. I also evaluated a new method of investigating nonword processing.

The most important conclusions from the computational investigations in this thesis were that the WSP model, which is sensitive to different statistical properties of the writing system, across varying sized print-to-sound correspondences, simulates central tendencies in human nonword reading responses as well as other models of reading (the DRC and the CDP++ model). However, none of the models compared here predicted the human modal responses adequately. Most importantly, each model failed to find the right balance between standard and context-sensitive vowel pronunciations for nonwords, compared to the pattern of human nonword responses found in several data sets.

The WSP's variable output was an attempt to simulate variability in nonword reading. While moderate success was achieved, several areas of improvement were also identified. Most importantly, these investigations generated some ideas for how to assess models that produce variable naming responses. It was also concluded that differentiating within-participants and between-participants variability or comparing models that produce either type of variability, requires detailed nonword naming data sets from several testing sessions.

The empirical investigations demonstrated that skilled readers are sensitive to both type and token frequencies of print-to-sound correspondences, which suggests that modifications are needed for computational models in which these properties are not included.

Detailed comparisons of the nonword rating responses and the nonword naming responses demonstrated the feasibility of the nonword rating method, with its strengths and weaknesses discussed, relative to the nonword naming method.

These conclusions bear relevance to future computational modelling of reading aloud as well as further empirical investigations in this area. I hope the insights provided by this PhD work serve future endeavours to improve our understanding of generalisation in reading.

References

Andrews, S. (1982). Phonological recoding: Is the regularity effect consistent?. *Memory & Cognition*, *10*(6), 565-575. https://doi.org/10.3758/bf03202439

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. https://doi.org/10.1037/0096-1523.24.4.1052

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, *52*(1), 388-407. https://doi.org/10.3758/s13428-019-01237-x

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (CD-ROM). University of Pennsylvania, Philadelphia: Linguistic Data Consortium.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of experimental psychology: General*, *133*(2), 283-316. https://doi.org/10.1037/0096-3445.133.2.283

Baron, J., & Strawson, C. (1976). Use of orthographic and word-specific knowledge in reading words aloud. *Journal of Experimental Psychology: Human perception and performance*, *2*(3), 386-393. https://doi.org/10.1037/0096-1523.2.3.386

Bowers, J. S., Davis, C. J., & Hanley, D. A. (2005). Automatic semantic activation of embedded words: Is there a 'hat' in 'that'? Journal of Memory & Language, 52, 131–143. https://doi.org/10.1016/j.jml.2004.09.003

Bowey, J. A. (1990). Orthographic onsets and rimes as functional units of reading. *Memory & Cognition*, *18*(4), 419-427. https://doi.org/10.3758/bf03197130

Bowey, J. A., & Hansen, J. (1994). The development of orthographic rimes as units of word recognition. *Journal of Experimental Child Psychology*, *58*(3), 465-488. https://doi.org/10.1006/jecp.1994.1045

Brown, G. D., & 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

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior research methods*, *51*(2), 467-479. https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in psychology*, *7*, 1116. https://doi.org/10.3389/fpsyg.2016.01116

Campbell, R., & Besner, D. (1981). This and THAP–-constraints on the pronunciation of new, written words. *The Quarterly Journal of Experimental Psychology*, *33*(4), 375-396. https://doi.org/10.1080/14640748108400799

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. Journal of social sciences, 3 (3), 106-116. https://doi.org/10.3844/jssp.2007.106.116

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464-473. https://doi.org/10.1177/1948550619875149

Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, *1*(2), 120-131. https://doi.org/10.1017/xps.2014.5

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, 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

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), 204-256. https://doi.org/10.1037/0033-295x.108.1.204

Coltheart, M., & Ulicheva, A. (2018). Why is nonword reading so variable in adult skilled readers?. *PeerJ*, *6*, e4879. https://doi.org/10.7717/peerj.4879

Compton, D. L., Steacy, L. M., Petscher, Y., Rueckl, J. G., Landi, N., & Pugh, K. R. (2019). Linking Behavioral and Computational Approaches to Better Understand Variant Vowel Pronunciations in Developing Readers. *New directions for child and adolescent development*, 2019(165), 55-71. https://doi.org/10.1002/cad.20294

Deavers, R., Solity, J., & Kerfoot, S. (2000). The effect of instruction on early nonword reading strategies. *Journal of Research in Reading*, *23*(3), 267-286. https://doi.org/10.1111/1467-9817.00122

De Simone, E., Beyersmann, E., Mulatti, C., Mirault, J., & Schmalz, X. (2021). Order among chaos: Cross-linguistic differences and developmental trajectories in pseudoword reading aloud using pronunciation Entropy. *PloS one*, *16*(5), e0251629. https://doi.org/10.1371/journal.pone.0251629

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175-191. https://doi.org/10.3758/bf03193146

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of verbal learning and verbal behavior*, *12*(6), 627-635. https://doi.org/10.1016/s0022-5371(73)80042-8

Forster, K. I., & Taft, M. (1994). Bodies, antibodies, and neighborhood-density effects in masked form priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 844-863. https://doi.org/10.1037/0278-7393.20.4.844

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

Gubian, M., Blything, R., Davis, C. J., & Bowers, J. S. (2022) Does that sound right? A novel method of evaluating models of reading aloud: Rating nonword pronunciations. *Behavior research methods*. https://doi.org/10.3758/s13428-022-01794-8

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological review*, *106*(3), 491-528. https://doi.org/10.1037/0033-295x.106.3.491

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

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of experimental psychology*, *41*(6), 401-410. https://doi.org/10.1037/h0056020

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

Johnson, D. (1970). Factors related to the pronunciation of vowel clusters. Technical Report No. 149. Madison: Wisconsin Research and Development Center for Cognitive Learning

Johnson, D. D., & Venezky, R. L. (1975). Models for predicting how adults pronounce vowel digraph spellings in unfamiliar words (Tech. Rep. 346). *Madison: Wisconsin Research and Development Center for Cognitive Learning*.

Kay, J., & Marcel, A. (1981). One process, not two, in reading aloud: Lexical analogies do the work of non-lexical rules. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 397-413. https://doi.org/10.1080/14640748108400800

Kessler, B. (2009). Statistical learning of conditional orthographic correspondences. *Writing Systems Research*, *1*(1), 19-34. https://doi.org/10.1093/wsr/wsp004

Kessler, B., & Treiman, R. (2001). Relationships between sounds and letters in English monosyllables. *Journal of memory and Language*, *44*(4), 592-617. https://doi.org/10.1006/jmla.2000.2745

Kim, J., Gabriel, U., & Gygax, P. (2019). Testing the effectiveness of the Internet-based instrument PsyToolkit: A comparison between web-based (PsyToolkit) and lab-based (E-Prime 3.0) measurements of response choice and response time in a complex psycholinguistic task. *PloS one*, *14*(9), e0221802. https://doi.org/10.1371/journal.pone.0221802

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

Nation, K., & Cocksey, J. (2009). Beginning readers activate semantics from sub-word orthography. Cognition, 110(2), 273–278. https://doi.org/10.1016/j.cognition.2008.11.004

Newman, A. J., Ullman, M. T., Pancheva, R., Waligura, D. L., & Neville, H. J. (2007). An ERP study of regular and irregular English past tense inflection. *NeuroImage*, *34*(1), 435-445. https://doi.org/10.1016/j.neuroimage.2006.09.007

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, *15*(5), 625-632. https://doi.org/10.1007/s10459-010-9222-y

Norris, D. (1994). A quantitative multiple-levels model of reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 1212–1232. https://doi.org/10.1037/0096-1523.20.6.1212

Patterson, K. E., & Morton, J. (1985). From orthography to phonology: An attempt at an old interpretation. In K. Patterson, J. C. Marshall, & M. Coltheart (Eds.), *Surface dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading* (pp. 335-359). Hove, England: Erlbaum. https://doi.org/10.4324/9781315108346-20

Perry, C. (2018). Reading orthographically strange nonwords: Modelling backup strategies in reading. *Scientific Studies of Reading*, *22*(3), 264-272. https://doi.org/10.1080/10888438.2018.1433673

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), 106-151. https://doi.org/10.1016/j.cogpsych.2010.04.001

Perry, C., Zorzi, M., & Ziegler, J. C. (2019). Understanding dyslexia through personalized large-scale computational models. *Psychological science*, *30*(3), 386-395. https://doi.org/10.1177/0956797618823540

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. https://doi.org/10.1037/0033-295x.103.1.56

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, *104*(1), 1-15. https://doi.org/10.1016/s0001-6918(99)00050-5

Pritchard, S. C., Coltheart, M., Marinus, E., & Castles, A. (2016). Modelling the implicit learning of phonological decoding from training on whole-word spellings and pronunciations. *Scientific studies of reading*, *20*(1), 49-63. https://doi.org/10.1080/10888438.2015.1085384

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*(5), 1268-1288. https://doi.org/10.1037/a0026703

R Core Team (2020). 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. (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

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., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic bulletin & review*, *11*(6), 1090-1098. https://doi.org/10.3758/bf03196742

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339-1362. https://doi.org/10.1080/02724980244000099

Robidoux, S., & Pritchard, S. C. (2014). Hierarchical clustering analysis of reading aloud data: a new technique for evaluating the performance of computational models. *Frontiers in Psychology*, *5*, 267. https://doi.org/10.3389/fpsyg.2014.00267

Rosson, M. B. (1983). From SOFA to LOUCH: Lexical contributions to pseudoword pronunciation. *Memory & Cognition*, *11*(2), 152-160. https://doi.org/10.3758/bf03213470

Ryder, R. J., & Pearson, P. D. (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

Schmalz, X., Marinus, E., Robidoux, S., Palethorpe, S., Castles, A., & Coltheart, M. (2014). Quantifying the reliance on different sublexical correspondences in German and English. *Journal of Cognitive Psychology*, *26*(8), 831-852. https://doi.org/10.1080/20445911.2014.968161

Schmalz, X., Robidoux, S., Castles, A., Coltheart, M., & Marinus, E. (2017). German and English bodies: No evidence for cross-linguistic differences in preferred orthographic grain size. *Collabra: Psychology*, *3*(1). https://doi.org/10.1525/collabra.72

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

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Siegelman, N., Kearns, D. M., & Rueckl, J. G. (2020). Using information-theoretic measures to characterize the structure of the writing system: the case of orthographic-phonological regularities in English. *Behavior Research Methods*, *52*(3), 1292-1312. https://doi.org/10.3758/s13428-019-01317-y

Snell, J., Grainger, J., & Declerck, M. (2018). A word on words in words: How do embedded words affect reading?. *Journal of Cognition*, *1*(1). https://doi.org/10.5334/joc.45

Stanback, M. L. (1992). Syllable and rime patterns for teaching reading: Analysis of a frequency-based vocabulary of 17,602 words. *Annals of Dyslexia*, *42*(1), 196-221. https://doi.org/10.1007/bf02654946

Steacy, L. M., Compton, D. L., Petscher, Y., Elliott, J. D., Smith, K., Rueckl, J. G., ... & Pugh, K. R. (2019). Development and prediction of context-dependent vowel pronunciation in elementary readers. *Scientific Studies of Reading*, *23*(1), 49-63. https://doi.org/10.1080/10888438.2018.1466303

Taraban, R., & McClelland, J. L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and language*, *26*(6), 608-631. https://doi.org/10.1016/0749-596x(87)90105-7

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

Treiman, R., Goswami, U., & Bruck, M. (1990). Not all nonwords are alike: Implications for reading development and theory. *Memory & Cognition*, *18*(6), 559-567. https://doi.org/10.3758/bf03197098

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., Kessler, B., & Evans, R. (2007). Anticipatory conditioning of spelling-to-sound translation. *Journal of Memory and Language*, *56*(2), 229-245. https://doi.org/10.1016/j.jml.2006.06.001

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: General*, *124*(2), 107-136. https://doi.org/10.1037/0096-3445.124.2.107

Treiman, R., & Zukowski, A. (1988). Units in reading and spelling. *Journal of Memory and Language*, *27*(4), 466-477. https://doi.org/10.1016/0749-596x(88)90068-x

Trudgill, P. (1984). *Language in the British isles*. Cambridge University Press.

Trudgill, P., & Gordon, E. (2006). Predicting the past: Dialect archaeology and Australian English rhoticity. English World-Wide, 27(3), 235-246. https://doi.org/10.1075/eww.27.3.02tru

Turner, G. W. (1994). English in Australia. The Cambridge history of the English language, 5, 277-327. https://doi.org/10.1017/chol9780521264785.007

Ulicheva, A., Coltheart, M., Grosseck, O., & Rastle, K. (2021). Are people consistent when reading nonwords aloud on different occasions?. *Psychonomic Bulletin & Review*, 28(5), 1679-1687. https://doi.org/10.3758/s13423-021-01925-w

Ullman, M. T. (1999). Acceptability ratings of regular and irregular past-tense forms: Evidence for a dual-system model of language from word frequency and phonological neighbourhood effects. *Language and Cognitive Processes*, *14*(1), 47-67. https://doi.org/10.1080/016909699386374

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly journal of experimental psychology*, *67*(6), 1176-1190. https://doi.org/10.1080/17470218.2013.850521

Waters, G. S., & Seidenberg, M. S. (1985). Spelling-sound effects in reading: Time-course and decision criteria. *Memory & Cognition*, *13*(6), 557-572. https://doi.org/10.3758/bf03198326

Welbourne, S. R., Woollams, A. M., Crisp, J., & Lambon Ralph, M. A. (2011). The role of plasticity-related functional reorganization in the explanation of central dyslexia. *Cognitive Neuropsychology*, 28(2), 65–108. https://doi.org/10.1080/02643294.2011.621937

Zevin, J. D., & Balota, D. A. (2000). Priming and attentional control of lexical and sublexical pathways during naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 121–135. https://doi.org/10.1037/0278-7393.26.1.121

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

Ziegler, J. C., Castel, C., Pech-Georgel, C., George, F., Alario, F. X., & Perry, C. (2008). Developmental dyslexia and the dual route model of reading: Simulating individual differences and subtypes. *Cognition*, 107(1), 151–178. https://doi.org/10.1016/j.cognition.2007.09.004

Ziegler, J., 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

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

Appendices

Appendix 1

Phoneme symbols in DISC character set used throughout the thesis

Table 1A

Phoneme symbols in DISC and IPA character sets

Example	DISC	IPA
p**a**t	{	a
p**e**t	E	ɛ
p**i**t	I	ɪ
p**o**t	Q	ɒ
p**u**tt	V	ʌ
p**u**t	U	ʊ
anoth**er**	@	ə
b**ar**n	#	ɑː
b**ea**n	i	iː
b**or**n	$	ɔː
b**oo**n	u	uː
b**ur**n	3	əː
b**ay**	1	eɪ
b**uy**	2	ʌɪ
b**oy**	4	ɔɪ
n**o**	5	əʊ
br**ow**	6	aʊ
p**eer**	7	ɪə
p**air**	8	ɛː
p**oor**	9	ʊə

Example	DISC	IPA
bat	b	b
cad	k	k
cheap	J	ʧ
dad	d	d
fat	f	f
game	g	g
had	h	h
jeep	_	dʒ
mea**s**ure	Z	ʒ
lad	l	l
mad	m	m
nat	n	n
ba**ng**	N	ŋ
pat	p	p
rat	r	r
sap	s	s
sheep	S	ʃ
tack	t	t
thin	T	θ
then	D	ð
vat	v	v
why	w	w
yank	j	j
zap	z	z

Appendix 2

Performance of the WSP model in the variable mode after optimisation using larger range of weights

Weights and performance of the WSP-type version of the model (variable mode), optimised for Andrews and Scarratt, Treiman and Pritchard sets, using a larger range of weights for the parsing styles (from 1 to 50 in increments of 5).

Table 2A

Performance of WSP-type model in variable mode (raw probabilities method) optimised for three nonword data sets using a range of weights from 1 to 50

data set	item group	human-model correlation	p-value	match proportion	weights
		optimised for Andrews & Scarratt set			
Andrews & Scarratt set	regular	0.86	< .001	1.00	CV-C: 1
	irregular	0.75	< .001	1.00	C-VC: 46
	reg-irreg diff.	0.84	< .001	1.00	C-V-C: 26
Treiman set		0.54	0.17	-	
Pritchard set	1st response	0.37	< .001	0.92	
	2nd response	0.17	0.04	0.40	
	3rd response	0.09	0.53	0.26	
		optimised for Treiman set			
Andrews & Scarratt set	regular	0.52	0.04	1.00	CV-C: 41
	irregular	0.08	0.76	1.00	C-VC: 46
	reg-irreg diff.	0.31	0.25	1.00	C-V-C: 1
Treiman set		0.82	0.01	-	
Pritchard set	1st response	0.30	< .001	0.92	
	2nd response	0.14	0.10	0.40	
	3rd response	0.20	0.17	0.26	
		optimised for Pritchard set			
Andrews & Scarratt set	regular	0.55	0.03	1.00	CV-C: 6
	irregular	0.64	0.01	1.00	C-VC: 11
	reg-irreg diff.	0.61	0.01	1.00	C-V-C: 21
Treiman set		0.60	0.11	-	
Pritchard set	1st response	0.42	< .001	0.92	
	2nd response	0.19	0.03	0.40	
	3rd response	0.12	0.41	0.26	

Note. cv-c = antibody-coda parsing style (e.g., *wa-sk*); c-vc = onset-word body parsing style (e.g., *w-ask*); c-v-c = small segment parsing style (e.g., *w-a-sk*).

Appendix 3

Stimuli Properties in investigations of token frequency of PSCs in nonword processing (Chapter 4)

Table 3A

Non-Unique Base words for Experimental Items

Non-Unique Base words			
BW	**BW2**	**BW2 Freq**	**BW type**
borne	airborne	3.39	Regular-high
	seaborne	1.97	
	waterborne	1.93	
	forborne	no Freq	
	overborne	no Freq	
bulb	flashbulb	1.39	Regular-high
	lightbulb	2.68	
curve	recurve	1.17	Regular-high
dealt	misdealt	no Freq	Irregular-high
gauge	rain-gauge	no Freq	Irregular-high
	wind-gauge	no Freq	
heart	sweetheart	4.28	Irregular-high
leash	unleash	3.36	Regular-low
month	twelvemonth	1.54	Irregular-high
soap	soft-soap	no Freq	Regular-high
waist	shirtwaist	1.17	Regular-high
watt	kilowatt	2.57	Irregular-low
world	dreamworld	1.74	Irregular-high
	underworld	3.21	

Note. BW = a baseword of an experimental nonword, BW2 = the lexical item or items that share a word body with the baseword, BW2 Freq = Zipf frequency of the BW2, BW type = the type of the baseword. 'no Freq' in the BW2 Freq column indicates that there was no frequency value available for the relevant item (i.e. this item does did exist in the SUBTLEX-UK database).

Table 3B

List of Experimental Stimuli in the Naming Task and Proportion of Regular and Irregular Pronunciations Assigned to Each Item

Item	Item type	Base word	Frequency (Zipf)	Number of responses	Regular responses	Irregular responses
BEALM	Irregular-low	realm	3.53	58	0.45	0.31
BROULT	Irregular-low	moult	2.7	39	0.31	0
CHUAVE	Irregular-low	suave	2.91	51	0.12	0.45
CRAUCHE	Irregular-low	gauche	2.56	0		
DUAVE	Irregular-low	suave	2.91	53	0.17	0.6
DWURGH	Irregular-low	burgh	2.84	1	0	0
FLIRSCH	Irregular-low	kirsch	2.5	1	1	0
FRORL	Irregular-low	whorl	1.3	36	0.75	0
GHUEDE	Irregular-low	suede	3.07	38	0.5	0.32
GLATT	Irregular-low	watt	3.45	57	0.91	0
GLOURGE	Irregular-low	scourge	2.99	10	0.7	0.3
HOULT	Irregular-low	moult	2.7	41	0.1	0.1
JAUCHE	Irregular-low	gauche	2.56	0		
LUSQUE	Irregular-low	brusque	2	15	0.67	0.2
MIRSCH	Irregular-low	kirsch	2.5	1	0	1
NEANSE	Irregular-low	cleanse	3.13	64	0.7	0.09
PHEALM	Irregular-low	realm	3.53	57	0.28	0.56
PSORL	Irregular-low	whorl	1.3	34	0.85	0
PSUGUE	Irregular-low	fugue	2.54	24	0	0.71
SHUGUE	Irregular-low	fugue	2.54	26	0.04	0.73
SMUSQUE	Irregular-low	brusque	2	15	0.2	0.4
SNULLE	Irregular-low	tulle	2.2	29	0.52	0.28
STEANSE	Irregular-low	cleanse	3.13	65	0.72	0.22
TUEDE	Irregular-low	suede	3.07	38	0.42	0.26
TWOURGE	Irregular-low	scourge	2.99	10	0.7	0.3
VULLE	Irregular-low	tulle	2.2	29	0.48	0.31
WURGH	Irregular-low	burgh	2.84	1	0	0
ZATT	Irregular-low	watt	3.45	57	0.98	0
BLEIZE	Irregular-high	seize	3.77	64	0.34	0.45
BREIRD	Irregular-high	weird	4.75	65	0.05	0.65
CHOUNG	Irregular-high	young	5.51	61	0.21	0.2
CRORST	Irregular-high	worst	4.92	66	0.8	0.02
DONTH	Irregular-high	month	5.08	67	1	0
DWONGE	Irregular-high	sponge	4.12	66	0.91	0.05
FLOUSSE	Irregular-high	mousse	3.76	67	0.25	0.61
FOUSSE	Irregular-high	mousse	3.76	64	0.13	0.78
FREART	Irregular-high	heart	5.3	66	0.2	0.24
GHEALT	Irregular-high	dealt	4.3	51	0.31	0.49

Table 3B continued

Item	Item type	Base word	Frequency (Zipf)	Number of responses	Regular responses	Irregular responses
GLEART	Irregular-high	heart	5.3	65	0.12	0.26
GLILST	Irregular-high	whilst	4.63	52	0.96	0.02
HAUGE	Irregular-high	gauge	3.69	25	0.4	0.2
JEALT	Irregular-high	dealt	4.3	53	0.32	0.38
LORST	Irregular-high	worst	4.92	65	1	0
MEIRD	Irregular-high	weird	4.75	67	0	0.75
NORLD	Irregular-high	world	5.88	68	0.82	0.01
PHOUTE	Irregular-high	route	4.6	58	0.16	0.72
PSORLD	Irregular-high	world	5.88	68	0.87	0.06
SHOUNG	Irregular-high	young	5.51	60	0.25	0.25
SMACHT	Irregular-high	yacht	3.77	60	0.85	0.05
SNAUGE	Irregular-high	gauge	3.69	25	0.32	0.04
STONTH	Irregular-high	month	5.08	67	0.97	0
TILST	Irregular-high	whilst	4.63	53	1	0
TWEIZE	Irregular-high	seize	3.77	64	0.02	0.81
VOUTE	Irregular-high	route	4.6	60	0.28	0.65
WONGE	Irregular-high	sponge	4.12	67	0.9	0.03
ZACHT	Irregular-high	yacht	3.77	60	0.85	0.1
BAIPSE	Regular-low	traipse	2.16	69	0.87	-
BREINT	Regular-low	feint	1.95	67	0.36	-
CHORGUE	Regular-low	morgue	3.12	66	0.94	-
CRAICE	Regular-low	plaice	3.02	68	0.94	-
DONCH	Regular-low	conch	2.81	69	1	-
DWOPSE	Regular-low	copse	2.59	68	0.71	-
FLAIPSE	Regular-low	traipse	2.16	69	0.86	-
FOMPT	Regular-low	prompt	3.13	67	0.97	-
FRAUZE	Regular-low	gauze	2.53	69	0.55	-
GHOMPT	Regular-low	prompt	3.13	67	1	-
GLIEK	Regular-low	shriek	2.78	69	0.7	-
GLOPSE	Regular-low	copse	2.59	68	0.84	-
HEINT	Regular-low	feint	1.95	69	0.16	-
JALC	Regular-low	talc	2.4	69	0.87	-
LORGUE	Regular-low	morgue	3.12	69	0.96	-
MEASH	Regular-low	leash	3.12	67	1	-
NOOTHE	Regular-low	soothe	2.76	69	0.97	-
PHULPT	Regular-low	sculpt	2.61	68	0.9	-
PSONCH	Regular-low	conch	2.81	66	0.97	-
SHAICE	Regular-low	plaice	3.02	69	0.75	-
SMEASH	Regular-low	leash	3.12	67	0.96	-
SNALC	Regular-low	talc	2.4	67	0.87	-
STILGE	Regular-low	bilge	2.53	69	0.96	-
TULPT	Regular-low	sculpt	2.61	69	0.93	-

Table 3B continued

Item	Item type	Base word	Frequency (Zipf)	Number of responses	Regular responses	Irregular responses
TWOOTHE	Regular-low	soothe	2.76	69	0.96	-
VAUZE	Regular-low	gauze	2.53	69	0.58	-
WILGE	Regular-low	bilge	2.53	69	0.96	-
ZIEK	Regular-low	shriek	2.78	68	0.82	-
BORPSE	Regular-high	corpse	3.41	68	0.99	-
CHOATHE	Regular-high	loathe	3.16	67	0.91	-
CROILT	Regular-high	spoilt	3.65	66	0.76	-
DWEK	Regular-high	trek	3.72	68	0.99	-
DWORPSE	Regular-high	corpse	3.41	68	0.93	-
DYNCH	Regular-high	lynch	3.44	68	0.87	-
FLURVE	Regular-high	curve	3.91	68	0.97	-
FOATHE	Regular-high	loathe	3.16	69	0.96	-
FROAP	Regular-high	soap	4.1	68	0.81	-
GHONZE	Regular-high	bronze	4.41	66	0.89	-
GLORNE	Regular-high	borne	3.44	69	1	-
HONZE	Regular-high	bronze	4.41	69	0.96	-
JOAP	Regular-high	soap	4.1	68	0.91	-
LOSQUE	Regular-high	mosque	3.81	67	0.88	-
MEACE	Regular-high	peace	4.74	69	0.99	-
NAIST	Regular-high	waist	3.71	69	0.78	-
PHOOB	Regular-high	boob	3.21	68	1	-
PSAIST	Regular-high	waist	3.71	66	0.73	-
SHULB	Regular-high	bulb	3.68	69	0.93	-
SMEACE	Regular-high	peace	4.74	69	0.99	-
SNYNCH	Regular-high	lynch	3.44	69	0.93	-
STULB	Regular-high	bulb	3.68	68	0.91	-
SWEK	Regular-high	trek	3.72	68	0.96	-
TURVE	Regular-high	curve	3.91	68	1	-
TWOSQUE	Regular-high	mosque	3.81	68	0.75	-
VOOB	Regular-high	boob	3.21	69	1	-
WOILT	Regular-high	spoilt	3.65	68	0.76	-
ZORNE	Regular-high	borne	3.44	69	0.99	-

Table 3C

List of Filler Items in the Naming Task

Naming Task Fillers				
BALSH	DORT	GERD	NYTH	SUNCE
BELSH	DRERN	GHIMN	OL	SWUS
BILTH	DRICHE	GILSH	PHISP	TELTH
BLALM	DULTH	GIPTH	PHONK	THAFE
BLUGE	DWAL	GIRSH	PHOZ	THIM
BLYPE	DWALP	GISE	PHRUP	THOUN
BRASK	DWARB	GORD	PLAIL	THWIE
BRORK	DWI	GRORD	PLANGE	TULTH
CALSH	DWYM	GRUIT	PLOFT	TWALPH
CALTH	FALTH	GULTH	POY	TWARK
CELTH	FATH	GWADD	PUDD	TWEIL
CEPTH	FENE	GWI	RELTH	TWING
CERSH	FENTH	GWIEL	RERNS	TWOVE
CHIEL	FEPTH	GWOB	RHUPS	TWULT
CHUILT	FIPTH	HELTE	RILSH	VAPSE
CIFF	FIRSH	JOCH	ROP	VATE
CILSH	FLANE	KEALD	ROUCHE	VEBB
CILTH	FLOAF	KNEAM	ROWSE	VIPTH
CIRSH	FLOLL	KNUSH	SCAWP	VORNS
CLUFT	FLUST	KULSE	SICH	WEFF
CORSH	FORSH	LECS	SKEWT	WHA
CREUM	FRA	LERSH	SKOAL	WHOLT
CRICHE	FRABE	LIRSH	SKUBE	WRAUK
CUPTH	FRARC	LUB	SKUNT	WROID
DALSH	FUPTH	LUT	SLONT	YELF
DALTH	FUSK	MUNE	SLYS	ZERE
DANGE	GALSH	NALK	SMILL	ZERPS
DAWSE	GALTH	NEPTH	SPAGS	ZI
DERSH	GELSH	NIS	SPEVE	ZOINS
DILSH	GELTH	NORB	SPLEZ	ZORT

Table 3D

List of Experimental Stimuli in the Rating Task

Rating Task Experimental Items							
Spelling	**Type**	**Pronunciation**		**Spelling**	**Type**	**Pronunciation**	
		Irregular	*Regular*			*Irregular*	*Regular*
BEALM	Irreg-low	bElm	bilm	BLEIZE	Irreg-high	bliz	bl1z
BROULT	Irreg-low	br5lt	br6lt	BREIRD	Irreg-high	br7d	br8d
CHUAVE	Irreg-low	Jw#v	J1v	CHOUNG	Irreg-high	JVN	J6N
CRAUCHE	Irreg-low	kr5S	kr$S	CRORST	Irreg-high	kr3st	kr$st
DUAVE	Irreg-low	dw#v	dw1v	DONTH	Irreg-high	dVnT	dQnT
DWURGH	Irreg-low	dwVr@	dw3g	DWOUTE	Irreg-high	dwut	dw6t
FLIRSCH	Irreg-low	fl7S	fl3S	FLOUSSE	Irreg-high	flus	fl6s
FRORL	Irreg-low	fr3l	fr$l	FOUSSE	Irreg-high	fus	f6s
GHUEDE	Irreg-low	g1d	gud	FREART	Irreg-high	fr#t	fr7t
GLATT	Irreg-low	glQt	gl{t	GHEALT	Irreg-high	gElt	gilt
GLOURGE	Irreg-low	gl3_	gl$_	GLEART	Irreg-high	gl#t	gl7t
HOULT	Irreg-low	h5lt	h6lt	GLILST	Irreg-high	gl2lst	glIlst
JAUCHE	Irreg-low	_5S	_$S	HAUGE	Irreg-high	h1_	h$_
LUSQUE	Irreg-low	lusk	lVsk	JEALT	Irreg-high	_Elt	_ilt
MIRSCH	Irreg-low	m7S	m3S	LORST	Irreg-high	l3st	l$st
NEANSE	Irreg-low	nEns	nins	MEIRD	Irreg-high	m7d	m8d
PHEALM	Irreg-low	fElm	film	NORLD	Irreg-high	n3ld	n$ld
PSORL	Irreg-low	s3l	s$l	PHONGE	Irreg-high	fVn_	fQn_
PSUGUE	Irreg-low	sug	sVg	PSORLD	Irreg-high	s3ld	s$ld
SHUGUE	Irreg-low	Sug	SVg	SHOUNG	Irreg-high	SVN	S6N
SMUSQUE	Irreg-low	smusk	smVsk	SMACHT	Irreg-high	smQt	sm{Jt
SNULLE	Irreg-low	snul	snVl	SNAUGE	Irreg-high	sn1_	sn$_
STEANSE	Irreg-low	stEns	stins	STONTH	Irreg-high	stVnT	stQnT
TUEDE	Irreg-low	tw1d	tjud	TILST	Irreg-high	t2lst	tIlst
TWOURGE	Irreg-low	tw3_	tw$_	TWEIZE	Irreg-high	twiz	tw1z
VULLE	Irreg-low	vjul	vVl	VOUTE	Irreg-high	vut	v6t
WURGH	Irreg-low	wVr@	w3g	WONGE	Irreg-high	wVn_	wQn_
ZATT	Irreg-low	zQt	z{t	ZACHT	Irreg-high	zQt	z{Jt

Table 3E

List of Filler Items in the Rating Task

Rating Task Fillers								
Spelling	**Pronunciation**			**Spelling**	**Pronunciation**	**Spelling**	**Pronunciation**
	Option1	*Option2*	*Option3*				
BLALM	bl#m	bl{lm		BALSH	b{lS	LUT	lVt
BRASK	br#sk	br{sk		BELSH	bElS	MUNE	mjun
CHIEL	J2l	Jil		BILTH	bIlT	NEPTH	nEpT
CHUILT	Jilt	Jult		BLUGE	bluZ	NORB	n$b
DANGE	d1n_	d{n_		BLYPE	bl2p	PHISP	fIsp
DRICHE	dr2J	driS		BRORK	br$k	PHRUP	frVp
DWALP	dwQlp	dw{lp		CREUM	krum	PLAIL	pl1l
DWARB	dw$b	dw#b		DAWSE	d$s	PLOFT	plQft
DWI	dw2	dwI		DERSH	d3S	POY	p4
FATH	f#T	f{T		DILSH	dIlS	RELTH	rElT
FLOLL	fl5l	flQl		DRERN	dr3n	RHUPS	rVps
FRA	fr#	fr{		DWYM	dwIm	ROP	rQp
GWIEL	gwIl	gw2l	gwil	FALTH	f{lT	SCAWP	sk$p
NALK	n$k	n{lk		FENTH	fEnT	SICH	slJ
NIS	nIs	nIz		FIPTH	fIpT	SKEWT	skjut
NYTH	n2T	nIT		FORSH	f$S	SKOAL	sk5l
OL	5l	Ql		FRARC	fr#k	SKUNT	skVnt
PHONK	f5nk	fQnk		FUPTH	fVpT	SLONT	slQnt
PHOZ	f5z	fQz		FUSK	fVsk	SPLEZ	splEz
PLANGE	pl1n_	pl{n_		GHIMN	gIm	SUNCE	sVns
ROUCHE	r5J	r6S		GISE	g2s	THIM	Tim
ROWSE	r5z	r6s		GRUIT	grut	THOUN	T6n
SKUBE	skVb	skub		GWADD	gw{d	TULTH	tVlT
SLYS	sl2s	slIs		HELTE	hElt	TWEIL	tw1l
SPLICHE	splIS			JOCH	_QJ	TWING	twIN
SWUS	swus	swVs		KEALD	kild	TWOVE	tw5v
TWALPH	twQlf	tw{lf		KNEAM	nim	TWULT	twVlt
TWARK	tw$k	tw#k		KNUSH	nVS	VATE	v1t
WHA	w1	w#		KULSE	kVls	WEFF	wEf
WHOLT	w5lt	wQlt		LECS	lEks	WRAUK	r$k
ZERE	z8	z3	z7	LIRSH	l3S	YELF	jElf
ZI	zi	z2	zI	LUB	lVb	ZORT	z$t

Additionally, 10 C-initial, 10 G-initial, 10 Error and 10 Odd items were included in the rating task, these items are listed in Appendix 11.

Table 3F

List of Words and Their Definition Options in the Vocabulary Task

Words and Four Definition Options in the Vocabulary task				
Word	**Option 1**	**Option 2**	**Option 3**	**Option 4**
WHORL	*spiral*	square	triangle	ellipse
BRUSQUE	careful	*abrupt*	slow	cheerful
TULLE	a woollen blanket	velvet drapes	*a thin cloth*	thick fabric
KIRSCH	wine made from dates	whiskey made from rice	vodka made from wheat	*brandy made from cherries*
FUGUE	*a loss of awareness of one's identity*	a state of exhaustion	a bad-tempered person	an increased amount of optimism
GAUCHE	exaggeratedly enthusiastic	*awkward*	rebellious	dishonest
MOULT	cause a large amount of damage	perplex (someone)	*shed old hair or skin*	renounce or reject (something)
BURGH	a remote village	a metropolitan area	an independent state	*a chartered town*
SUAVE	*charming*	timid	sentimental	hostile
SCOURGE	a showy and purely ornamental thing	*thing that causes great trouble or suffering*	the scope or bounds of something	the point at which something is at its best
SUEDE	cotton	silk	*leather*	wool
CLEANSE	empty	resize	twist	*purify*
WATT	*unit of power*	unit of time	unit of length	unit of depth
REALM	an international organisation	*a kingdom*	an unrecognized state	a military dictatorship
GAUGE	summarise	steer	*measure*	advertise
MOUSSE	a type of drink	a type of nutritional supplement	a type of soup	*a type of dessert*
SEIZE	*grab*	throw	break	push
YACHT	a racing car	*a sailing boat*	a rowing boat	a cycle with a single wheel
SPONGE	a device for sharpening razors	material providing heat insulation for a water tank and pipes	*a piece of a soft and absorbent substance used for cleaning*	a medium-paced French dance
DEALT	(past tense) estimate	(past tense) steal	(past tense) approach	*(past tense) distribute*
ROUTE	*course*	hight	duration	consistency
WHILST	in a straight line	*at the same time*	at a fairly brisk speed	in a smooth flowing manner
WEIRD	superficial	loyal	*very strange*	popular
WORST	the finest	the most common	the most expensive	*of the poorest quality*
MONTH	*4 weeks*	2 weeks	7 days	14 weeks

Table 3F continued

Word	Option 1	Option 2	Option 3	Option 4
HEART	a glandular organ involved in many metabolic processes	*a muscular organ that pumps blood through the circulatory system*	an abdominal organ involved in the production and removal of blood cells	organ of balance and hearing embedded in the temporal bone
YOUNG	middle-aged	elderly	*immature*	intermittent
WORLD	a star	an asteroid	a moon	*a planet*

Note. The correct answer is in italics.

Appendix 4

Conversion of phonemic transcription from Plaut et al. (1996) to DISC

The key for the phonemic transcription and the corresponding example word given in Plaut et al. (1996, Appendix B and C, p. 114-115) was followed. Additionally, the choice of DISC transcription for the phonemes a, o, ur and Or in Plaut et al. was supported by inference from the example transcriptions of existing words in Appendix C of their paper (e.g. 'BROAD - /brod/', 'WANT - /want/', 'WERE - /wur/' and 'SWARM - /swOrm/').

Table 4A

Conversion of phonemic transcription from Plaut et al. (1996) to DISC

Plaut et al.	example	DISC
a	pot	Q
@	cat	{
e	bed	E
i	hit	I
o	dog	$
u	good	U
A	make	1
E	keep	i
I	bike	2
O	hope	5
U	boot	u
W	now	6
Y	boy	4
^	cup	V
N	ring	N
S	she	S
C	chin	J
Z	beige	Z
T	thin	T
D	this	D
j	jeep	_
y	yes	j
ur[a]	were	3
Or[a]	swarm	$

Note. The rest of the consonant phonemes are written as in DISC, as Plaut et al. state that the remaining phonemes are transcribed in the 'conventional way (e.g. /b/ in BAT)'.

[a] These print-to-sound correspondences were inferred from the Appendix C (Plaut et al., 1996, p. 115).

Appendix 5

Analyses of naming and rating responses to Irregular items with stricter criteria for individual means (Chapter 4)

The samples in these analyses only contain participants with a minimum of 10 valid responses in each condition. See Table 5A for descriptive statistics for the naming and rating data.

Table 5A

Proportion and mean ratings of irregularly pronounced Irregular items in naming and rating tasks

Data	Group	Irregular-low			Irregular-high		
		n	*Mean*	*SD*	*n*	*Mean*	*SD*
Naming	*Naming-Rating*	59	0.26	0.14	59	0.28	0.1
Rating	*Naming-Rating*	60	4.88	0.51	60	4.80	0.54
	Rating-Only	56	4.35	0.51	56	4.50	0.55

Comparison of proportion of irregular pronunciations assigned to Irregular-low and Irregular-high items as a paired-samples t-test revealed no reliable differences ($t(58) = 0.93$, $p = 0.18$, Cohen's $dz = 0.12$). The rating data for paired samples t-tests comparing the mean ratings to irregularly pronounced Irregular items is reported in Table 5B.

Table 5B

Comparison of the mean acceptability ratings to irregularly pronounced low and high frequency nonwords

Group	Mean diff.	t-value	df	p-value	dz	Min. dz
Naming-Rating	-0.08	-1.82	59	0.96	-0.24	0.32
Rating-Only	0.15	2.32	55	0.01	0.31	0.34

Note. Min. dz shows the minimum detectable effect sizes from sensitivity analyses computed for 1-tailed, paired samples t-tests with an alpha level of .05, power of .8 and sample size of either 60 or 56.

Appendix 6

Human and model responses to Irregular nonwords with varied token frequency

Table 6A

Human and model responses to Irregular nonwords

		Human participants						Computational models					
Item	*Type*	*first*	*n*	*second*	*n*	*third*	*n*	*DRC*	*CDP++*	*Psim1*	*WSP-type*	*WSP-token*	*WSP-token-T*
BEALM	Low	bilm	23	bElm	16	bIlm	11	bilm	bElm	bilm	bElm	bElm	bElm
BLEIZE	High	bliz	29	bl1z	21	bl2z	10	bl1z	bl1z	bl1z	bliz	bliz	bliz
BREIRD	High	br7d	40	br8d	11	brid	3	br1rd	br8d	br1rd	br7d	br7d	br7d
BROULT	Low	brUlt	10	brQlt	8	br6t	7	br6lt	br$lt	br5lt	br5lt	br$lt	br6lt
CHOUNG	High	JQN	16	JVN	12	J6N	10	J6N	J6N_	k6N	JVN	JVN	J6N
CHUAVE	Low	Jw#v	13	J6v	5	Jw1v	4	JV1v	Jw#v	JQv	J#v	J#v	J#v
CRAUCHE	Low	-	-	-	-	-	-	kr$S	kr$J	kr5S	kr5S	kr5S	kr5S
CRORST	High	kr$st	48	krQst	8	k$st	2	kr$st	kr$st	kr$st	kr3st	kr3st	kr$st
DONTH	High	dQnT	65	dQnt	1	dQNT	1	dQnT	dVnT	dVnT	dVnT	dVnT	dVnT
DUAVE	Low	dw#v	32	dw1v	9	d$v	2	dV1v	dwuv	dQv	d#v	d#v	d#v
DWONGE	High	dwQn_	39	dwQN	17	dwVN	3	dwQn_	dwQn	dw$n_	dwVn_	dwVn_	dwQn_
DWURGH	Low	dw#	1	-	-	-	-	dw3g	dw3	dw3	dwV	dwV	dw3
FLIRSCH	Low	fl3S	1	-	-	-	-	fl3S	fl3zJ	fl3sJ	fl7S	fl7S	fl7S
FLOUSSE	High	flus	39	fl6s	15	flQs	2	fl6si	fl6s	fl6s	flus	flus	flus
FOUSSE	High	fus	45	f6s	7	fus1	5	f6si	f6s	f6s	fus	fus	fus

Table 6A continued

		Human participants						Computational models					
Item	*Type*	*first*	*n*	*second*	*n*	*third*	*n*	*DRC*	*CDP++*	*Psim1*	*WSP-type*	*WSP-token*	*WSP-token-T*
FREART	High	fr#t	15	fr3t	15	fr7t	12	fr7t	fr7t	frirt	fr#t	fr#t	fr7t
FRORL	Low	fr$l	24	frQl	6	$	1	fr$l	fr$l	fr$l	fr3l	fr3l	fr$l
GHEALT	High	gElt	18	gilt	14	gIlt	5	gilt	gElt	gElt	gElt	gElt	gilt
GHUEDE	Low	gud	16	gw1d	10	gwid	3	gjud	gjud	gjud	g1d	g1d	g1d
GLATT	Low	gl{t	51	glVt	3	gl#t	2	gl{t	gl{t	gl{t	glQt	glQt	gl{t
GLEART	High	gl3t	37	gl#t	16	gl7t	8	gl7t	gl7t	gl$t	gl#t	gl#t	gl7t
GLILST	High	glIlst	33	glIst	7	glIst	5	glIlst	glIlst	glIlst	gl2lst	gl2lst	glIlst
GLOURGE	Low	gl$_	6	gl3_	2	gl$g	1	gl$_	gl$_	gl$_	gl3_	gl3_	gl3_
HAUGE	High	h$_	6	h1g	4	h6g	4	h$_	h$_	h$	h1_	h1_	h1_
HOULT	Low	hQlt	21	hUlt	6	h5lt	4	h6lt	h6lt	h5lt	h5lt	h5lt	h5lt
JAUCHE	Low	-	-	-	-	-	-	_$S	_$J	5	_5S	_5S	_5S
JEALT	High	_Elt	20	_ilt	16	_Ilt	13	_ilt	_Elt	_ilt	_Elt	_Elt	_ilt
LORST	High	l$st	63	l$s	2	-	-	l$st	l$st	l$st	l$st	l$st	l$st
LUSQUE	Low	lVsk	9	lusk	3	l{sk1	1	lVsk	lVsk	lus	lusk	lusk	lVsk
MEIRD	High	m7d	50	m3d	5	m8d	5	m1rd	m7d	mird	m7d	m7d	m7d
MIRSCH	Low	m7S	1	-	-	-	-	m3S	m3sJ	m3sJ	m7S	m7S	m7S
NEANSE	Low	nins	41	nEns	6	ni{ns	3	nins	nEns	ninz	nEnz	nEnz	nEnz
NORLD	High	n$ld	50	nQld	7	n$d	4	n$ld	n$ld	n$ld	n$ld	n$ld	n$ld
PHEALM	Low	fElm	30	film	16	fIlm	6	film	fElm	ilm	fElm	fElm	fim
PHOUTE	High	fut	40	f6t	9	f5t	3	f6t	f6t	f5t	fut	fut	f6t
PSORL	Low	ps$l	13	s$l	12	s$	2	s$l	p$l	sp$l	s$l	s$l	s$l
PSORLD	High	s$ld	34	ps$ld	14	s$d	3	s$ld	p$ld	sp$ld	s$ld	s3ld	s$ld
PSUGUE	Low	sug	6	psug	2	sjug	2	sug	pVg	spV_	sVg	sVg	sVg

Table 6A continued

		Human participants						Computational models					
Item	*Type*	*first*	*n*	*second*	*n*	*third*	*n*	*DRC*	*CDP++*	*Psim1*	*WSP-type*	*WSP-token*	*WSP-token-T*
SHOUNG	High	SVN	15	SQN	13	S6N	12	S6N	SVN	SV	SVN	SVN	SVN
SHUGUE	Low	Sug	13	Su_	4	Sugu	2	Sjug	SVg	Sug	Sug	Sug	SVg
SMACHT	High	sm{kt	21	sm{Jt	9	sm{J	6	sm{Jt	sm{tt	sm{J	smQt	smQt	smQt
SMUSQUE	Low	smusk	3	smUsk	2	musk	1	smVsk	smVsk	smVs	smusk	smusk	smVsk
SNAUGE	High	sn6_	6	sn$_	4	sn$g	4	sn$_	sn$_	sn1_	sn1_	sn1_	sn1_
SNULLE	Low	snVl	15	snul	8	snUl	4	snVl	snul	snVl	snul	snul	snVl
STEANSE	Low	stins	44	stEns	13	i	1	stins	stEns	stin	stEnz	stEnz	stEnz
STONTH	High	stQnT	61	st5nT	2	dQnT	1	stQnT	stVnT	stVnT	stVnT	stVnT	stVnT
TILST	High	tIlst	50	tIls	2	dIlst	1	tIlst	tIlst	tIlst	t2lst	t2lst	t2lst
TUEDE	Low	tud	13	twid	10	tw1d	9	tjud	tjud	tud	t1d	t1d	t1d
TWEIZE	High	twiz	52	tw2z	8	tw2	2	tw1z	tw1z	tw2z	twiz	twiz	twiz
TWOURGE	Low	tw$_	4	tw$g	3	tw3Z	2	tw$_	tw$_	t3	tw3_	tw3_	tw$_
VOUTE	High	vut	39	v6t	17	v5t	2	v6t	v6t	v6t	vut	vut	v6t
VULLE	Low	vVl	14	vul	9	vUl	4	vVl	vjul	vjUl	vul	vul	vVl
WONGE	High	wQn_	33	wQN	22	w5n_	3	wQn_	wQn	wVn_	wVn_	wVn_	wQn_
WURGH	Low	w3g@	1	-	-	-	-	w3g	w3	w3	wV	wV	wV
ZACHT	High	z{kt	31	z{Jt	5	z{J	4	z{Jt	z{tt	QJ	zQt	zQt	zQt
ZATT	Low	z{t	55	{	1	sVt	1	z{t	z{t	z{t	zQt	zQt	z{t

Table 6B

Model output not produced by any participant

Item	DRC	CDP++	Psim1	WSP-type	WSP-token	WSP-token-T
BREIRD	br1rd		br1rd			
BROULT		br$lt	br5lt	br5lt	br$lt	
CHOUNG		J6N_	k6N			
CHUAVE	JV1v		JQv			
CRORST				kr3st	kr3st	
DONTH		dVnT	dVnT	dVnT	dVnT	dVnT
DUAVE	dV1v	dwuv	dQv	d#v	d#v	d#v
DWONGE		dwQn	dw$n_	dwVn_	dwVn_	
DWURGH	dw3g	dw3	dw3	dwV	dwV	dw3
FLIRSCH		fl3zJ	fl3sJ	fl7S	fl7S	fl7S
FLOUSSE	fl6si					
FOUSSE	f6si					
FREART			frirt			
FRORL				fr3l	fr3l	
GHUEDE	gjud	gjud	gjud			
GLEART			gl$t			
GLATT				glQt	glQt	
GLILST				gl2lst	gl2lst	
HAUGE			h$	h1_	h1_	h1_
LORST						
LUSQUE			lus			
MEIRD	m1rd		mird			
MIRSCH	m3S	m3sJ	m3sJ			
NEANSE				nEnz	nEnz	nEnz
PHEALM			ilm			fim
PSORL		p$l	sp$l			
PSORLD		p$ld	sp$ld			
PSUGUE		pVg	spV_	sVg	sVg	sVg
SHOUNG			SV			
SHUGUE	Sjug	SVg				SVg
SMACHT		sm{tt				
SMUSQUE			smVs			
SNAUGE			sn1_	sn1_	sn1_	sn1_
STONTH		stVnT	stVnT	stVnT	stVnT	stVnT
TILST				t2lst	t2lst	t2lst
TWOURGE			t3			
VULLE		vjul	vjUl			
WURGH	w3g	w3	w3	wV	wV	wV
ZACHT		z{tt	QJ	zQt	zQt	

Appendix 7

Comparison of vowel segment properties (Chapter 4)

The frequency measures of vowel segments of low and high (token) frequency nonwords were compared with a Welch's t-test for both Irregular and Regular item groups. The tests were conducted as 1-tailed tests, in the direction indicated by the descriptive statistics of the vowel segments. Table 7A summarises the results of these comparisons.

Table 7A

Welch's t-tests comparing properties of vowel segments between low and high items of Irregular and Regular item groups

Comparison direction	Property	t-value	df	p-value
	Regular items			
low > high	type freq	1.24	49.58	0.11
low > high	sum token freq	1.11	50.27	0.14
high > low	max token freq	0.17	51.19	0.43
	Irregular items			
low > high	type freq	0.7	45.1	0.24
low > high	sum token freq	0.57	46.1	0.29
high > low	max token freq	0.93	39	0.18

Appendix 8

Stimuli Properties in investigations of type frequency of PSCs in nonword processing (Chapter 5)

Table 8A

Properties of the Experimental Stimuli in the Naming Task and proportion of irregular and regular pronunciations assigned to each item

Item	Item type	Base words	Max frequency (Zipf)	Type frequency	Number of responses	Regular responses	Irregular responses
BRALF	Irregular-Many	half, calf, behalf	5.55	3	55	0.84	0.05
FLALD	Irregular-Many	bald, scald	3.72	2	55	0.78	0.07
FLALT	Irregular-Many	halt, malt, salt…	4.69	5	55	0.69	0.04
FRALM	Irregular-Many	balm, calm, palm…	4.72	5	55	0.69	0.25
GEALTH	Irregular-Many	health, stealth, wealth	5.14	3	54	0.15	0.44
GHASK	Irregular-Many	cask, mask, task…	4.66	6	54	0.67	0.33
GHIGN	Irregular-Many	sign, align, benign…	4.98	10	49	0.39	0.45
GLALK	Irregular-Many	balk, chalk, talk…	5.52	5	55	0.65	0.11
GLIGH	Irregular-Many	high, sigh, thigh…	5.53	4	54	0.39	0.46
GNOLK	Irregular-Many	folk, yolk	4.24	2	54	0.87	0
KEALTH	Irregular-Many	health, stealth, wealth	5.14	3	55	0.31	0.4
KYME	Irregular-Many	rhyme, thyme	4	2	54	0	0.98
MEARN	Irregular-Many	earn, learn, yearn	5.01	3	55	0.25	0.71
NALM	Irregular-Many	balm, calm, palm…	4.72	5	55	0.62	0.31
NALT	Irregular-Many	halt, malt, salt…	4.69	5	54	0.56	0.06
PHOUP	Irregular-Many	group, soup, recoup	5.19	6	55	0.11	0.87
PLALF	Irregular-Many	half, calf, behalf	5.55	3	54	0.91	0
PLIGN	Irregular-Many	sign, align, benign…	4.98	10	55	0.16	0.78
PSASK	Irregular-Many	cask, mask, task…	4.66	6	52	0.85	0.1
RHALD	Irregular-Many	bald, scald	3.72	2	53	0.74	0.06

Table 8A continued

Item	Item type	Base words	Max frequency (Zipf)	Type frequency	Number of responses	Regular responses	Irregular responses
RHOLK	Irregular-Many	folk, yolk	4.24	2	51	0.88	0.04
SMIQUE	Irregular-Many	pique, antique, unique…	4.66	11	52	0.13	0.65
SMYME	Irregular-Many	rhyme, thyme	4	2	53	0	0.89
SNAST	Irregular-Many	blast, cast, fast…	5.97	14	55	0.91	0.09
TWEARN	Irregular-Many	earn, learn, yearn	5.01	3	55	0.09	0.85
VOUP	Irregular-Many	group, soup, recoup	5.19	6	55	0.05	0.91
YAST	Irregular-Many	blast, cast, fast…	5.97	14	55	0.85	0.15
YIQUE	Irregular-Many	pique, antique, unique…	4.66	11	55	0.15	0.73
ZALK	Irregular-Many	balk, chalk, talk…	5.52	5	55	0.73	0.15
ZIGH	Irregular-Many	high, sigh, thigh…	5.53	4	55	0.22	0.65
BEALM	Irregular-Single	realm	3.53	1	54	0.48	0.33
BLEIZE	Irregular-Single	seize	3.77	1	54	0.24	0.57
BREIRD	Irregular-Single	weird	4.75	1	55	0.13	0.67
CHOUNG	Irregular-Single	young	5.51	1	54	0.33	0.07
CRORST	Irregular-Single	worst	4.92	1	55	0.84	0
DONTH	Irregular-Single	month	5.08	1	54	0.98	0.02
DWOUTE	Irregular-Single	route	4.6	1	54	0.3	0.56
FLOUSSE	Irregular-Single	mousse	3.76	1	55	0.25	0.64
FOUSSE	Irregular-Single	mousse	3.76	1	54	0.15	0.78
FREART	Irregular-Single	heart	5.3	1	52	0.25	0.38
GHEALT	Irregular-Single	dealt	4.3	1	54	0.17	0.48
GLEART	Irregular-Single	heart	5.3	1	52	0.29	0.21
GLILST	Irregular-Single	whilst	4.63	1	53	0.94	0.04
HAUGE	Irregular-Single	gauge	3.69	1	52	0.48	0.21
JEALT	Irregular-Single	dealt	4.3	1	55	0.2	0.49

Table 8A continued

Item	Item type	Base words	Max frequency (Zipf)	Type frequency	Number of responses	Regular responses	Irregular responses
LORST	Irregular-Single	worst	4.92	1	55	0.96	0.04
MEIRD	Irregular-Single	weird	4.75	1	53	0.02	0.92
NORLD	Irregular-Single	world	5.88	1	55	0.96	0.02
PHEALM	Irregular-Single	realm	3.53	1	54	0.2	0.57
PHONGE	Irregular-Single	sponge	4.12	1	54	0.83	0.11
PSORLD	Irregular-Single	world	5.88	1	53	0.91	0.04
SHOUNG	Irregular-Single	young	5.51	1	54	0.28	0.06
SMACHT	Irregular-Single	yacht	3.77	1	53	0.85	0.11
SNAUGE	Irregular-Single	gauge	3.69	1	54	0.35	0.09
STONTH	Irregular-Single	month	5.08	1	55	1	0
TILST	Irregular-Single	whilst	4.63	1	55	1	0
TWEIZE	Irregular-Single	seize	3.77	1	55	0.05	0.78
VOUTE	Irregular-Single	route	4.6	1	52	0.19	0.75
WONGE	Irregular-Single	sponge	4.12	1	55	0.95	0.02
ZACHT	Irregular-Single	yacht	3.77	1	53	0.77	0.17

Table 8B

Fillers in the Naming task

Naming Task Fillers				
BALSH	DWARM	GULTH	PLOFT	THOUN
BELSH	DWI	GWADD	POY	THWIE
BILTH	DWYM	GWI	PRU	THWYM
BLUGE	FALTH	GWIEL	PUDD	TREWN
BLYPE	FATH	GWOB	RADGE	TRISK
BRORK	FEECE	GWUTT	RELTH	TULTH
CALSH	FENE	HELTE	RERNS	TWALPH
CALTH	FENTH	JACE	RERV	TWEIL
CELTH	FEPTH	JOCH	RESS	TWING
CEPTH	FIPTH	KAUVE	RHESK	TWITE
CERSH	FIRSH	KEALD	RHUPS	TWOVE
CHIEL	FLANE	KNEAM	RILSH	TWULT
CHUILT	FLOAF	KNULB	ROP	VAPSE
CIFF	FLOLL	KNUSH	ROUCHE	VARP
CILSH	FLUST	KULSE	ROWSE	VATE
CILTH	FLUTH	LAWK	SCALC	VEBB
CIRSH	FORSH	LECS	SCAWP	VIPTH
CLARP	FRA	LERSH	SCRIF	VIVE
CLITE	FRABE	LIRSH	SICH	VOOCH
CLUFT	FRARC	LUB	SKEWT	VORNS
CORSH	FRAUL	LUN	SKOAL	VUD
CREUM	FRAVE	LUT	SKUBE	WEFF
CRICHE	FUPTH	MUNE	SKUNT	WERF
CRIFE	FUSK	NARSE	SLONT	WERGE
CUPTH	GALSH	NARVE	SLULK	WHA
DALSH	GALTH	NEAF	SLYS	WHOLT
DALTH	GELSH	NEPTH	SMILL	WHURF
DANGE	GELTH	NIS	SNEBE	WRAUK
DAWSE	GERD	NORB	SNOBE	WROID
DEET	GHIMN	NYTH	SPAC	YARL
DENGE	GHURF	OL	SPAGS	YELF
DERSH	GILSH	PEAF	SPEVE	YOAT
DILSH	GIPTH	PHIEK	SPLEZ	YUCH
DORT	GIRM	PHISP	SUNCE	ZARVE
DRERN	GIRSH	PHONK	SWURB	ZERE
DRICHE	GISE	PHOZ	SWUS	ZERPS
DULF	GLELP	PHROR	TELTH	ZI
DULTH	GLERT	PHRUP	THAFE	ZOINS
DWAL	GORD	PLAIL	THIM	ZORT
DWALP	GRORD	PLANGE	THOOT	ZOSE
DWARB	GRUIT			

Table 8C

Experimental items in the Rating task

Rating Task Experimental Items							
Spelling	**Type**	**Pronunciation**		**Spelling**	**Type**	**Pronunciation**	
		Irregular	*Regular*			*Irregular*	*Regular*
RHALD	IM	r$ld	r{ld	BEALM	IS	bElm	bilm
BRALF	IM	br#f	br{lf	HAUGE	IS	h1_	h$_
ZALK	IM	z$k	z{lk	FOUSSE	IS	fus	f6s
NALM	IM	n#m	n{lm	BLEIZE	IS	bliz	bl1z
FLALT	IM	fl$lt	fl{lt	ZACHT	IS	zQt	z{Jt
GHASK	IM	g#sk	g{sk	WONGE	IS	wVn_	wQn_
YAST	IM	j#st	j{st	JEALT	IS	_Elt	_ilt
KEALTH	IM	kElT	kilT	VOUTE	IS	vut	v6t
MEARN	IM	m3n	m7n	TILST	IS	t2lst	tIlst
GLIGH	IM	gl2	glI	MEIRD	IS	m7d	m8d
GHIGN	IM	g2n	gIn	LORST	IS	l3st	l$st
YIQUE	IM	jik	jIk	DONTH	IS	dVnT	dQnT
GNOLK	IM	n5k	nQlk	GLEART	IS	gl#t	gl7t
VOUP	IM	vup	v6p	SHOUNG	IS	SVN	S6N
KYME	IM	k2m	kIm	NORLD	IS	n3ld	n$ld
FLALD	IM	fl$ld	fl{ld	PHEALM	IS	fElm	film
PLALF	IM	pl#f	pl{lf	SNAUGE	IS	sn1_	sn$_
GLALK	IM	gl$k	gl{lk	FLOUSSE	IS	flus	fl6s
FRALM	IM	fr#m	fr{lm	TWEIZE	IS	twiz	tw1z
NALT	IM	n$lt	n{lt	SMACHT	IS	smQt	sm{Jt
PSASK	IM	s#sk	s{sk	PHONGE	IS	fVn_	fQn_
SNAST	IM	sn#st	sn{st	GHEALT	IS	gElt	gilt
GEALTH	IM	gElT	gilT	DWOUTE	IS	dwut	dw6t
TWEARN	IM	tw3n	tw7n	GLILST	IS	gl2lst	glIlst
ZIGH	IM	z2	zI	BREIRD	IS	br7d	br8d
PLIGN	IM	pl2n	plIn	CRORST	IS	kr3st	kr$st
SMIQUE	IM	smik	smIk	STONTH	IS	stVnT	stQnT
RHOLK	IM	r5k	rQlk	FREART	IS	fr#t	fr7t
PHOUP	IM	fup	f6p	CHOUNG	IS	JVN	J6N
SMYME	IM	sm2m	smIm	PSORLD	IS	s3ld	s$ld

Note. IM = Irregular-Many items; IS = Irregular-Single items

Table 8D

Filler items in the Rating task

Rating Task Fillers								
Spelling	**Pronunciation**			**Spelling**	**Pronunciation**	**Spelling**	**Pronunciation**	
	Option1	*Option2*	*Option3*					
BALSH	b{lS			BRORK	br$k	PHRUP	frVp	
BELSH	bElS			CREUM	krum	PLAIL	pl1l	
BILTH	bIlT			DAWSE	d$s	PLOFT	plQft	
BLUGE	bluZ			DERSH	d3S	POY	p4	
BLYPE	bl2p			DILSH	dIlS	RELTH	rElT	
CHIEL	J2l	Jil		DRERN	dr3n	RHUPS	rVps	
CHUILT	Jilt	Jult		DWYM	dwIm	ROP	rQp	
DANGE	d1n_	d{n_		FALTH	f{lT	SCAWP	sk$p	
DRICHE	dr2J	driS		FENTH	fEnT	SICH	sIJ	
DWALP	dwQlp	dw{lp		FIPTH	flpT	SKEWT	skjut	
DWARB	dw$b	dw#b		FORSH	f$S	SKOAL	sk5l	
DWI	dw2	dwI		FRARC	fr#k	SKUNT	skVnt	
FATH	f#T	f{T		FUPTH	fVpT	SLONT	slQnt	
FLOLL	fl5l	flQl		FUSK	fVsk	SPLEZ	splEz	
FRA	fr#	fr{		GHIMN	gIm	SPLICHE	splIS	
GWIEL	gwIl	gw2l	gwil	GISE	g2s	SUNCE	sVns	
NIS	nIs	nIz		GRUIT	grut	THIM	Tim	
NYTH	n2T	nIT		GWADD	gw{d	THOUN	T6n	
OL	5l	Ql		HELTE	hElt	TULTH	tVlT	
PHONK	f5nk	fQnk		JOCH	_QJ	TWEIL	tw1l	
PHOZ	f5z	fQz		KEALD	kild	TWING	twIN	
PLANGE	pl1n_	pl{n_		KNEAM	nim	TWOVE	tw5v	
ROUCHE	r5J	r6S		KNUSH	nVS	TWULT	twVlt	
ROWSE	r5z	r6s		KULSE	kVls	VATE	v1t	
SKUBE	skVb	skub		LECS	lEks	WEFF	wEf	
SLYS	sl2s	slIs		LIRSH	l3S	WRAUK	r$k	
SWUS	swus	swVs		LUB	lVb	YELF	jElf	
TWALPH	twQlf	tw{lf		LUT	lVt	ZORT	z$t	
WHA	w1	w#		MUNE	mjun			
WHOLT	w5lt	wQlt		NEPTH	nEpT			
ZERE	z8	z3	z7	NORB	n$b			
ZI	zi	z2	zI	PHISP	flsp			

Additionally, 10 C-initial, 10 G-initial, 10 Error and 10 Odd items were included in the Rating task, these items are listed in Appendix 11.

Appendix 9

Analyses with full set of nonword items (Chapter 5)

Nonword naming and rating responses were also compared with the original, full set of the nonword stimuli. Table 9A depicts the key properties of the full set of items. Table 9B shows the mean proportion of irregular pronunciations (standard deviations in brackets) assigned to the nonwords, mean ratings for irregularly pronounced nonwords and 1-tailed t-test results comparing the mean irregular naming proportions and ratings between Irregular-Single and Irregular-Many items.

Table 9A

Key properties of the full set of Irregular-Many and Irregular-Single items

Statistic	Irregular-Many base words	Irregular-Many Max Zipf	Irregular-Single Zipf	Irregular-Single known base words	Irregular-Single prop. of irreg. pron.
Mean	5.40	4.91	4.51	0.88	0.29
SD	3.54	0.61	0.72	0.16	0.28

Note. Irregular-Single known base words = proportion of participants that defined and pronounced at least the vowel of the base word correctly in the study reported in Chapter 4. Irregular-Single prop. of irreg. pron. = proportion of irregular pronunciations assigned to each nonword (calculated as pooled responses to two nonwords that both had the same word body) in the study reported in Chapter 4.

Table 9B

Mean proportions of irregular responses, mean ratings for irregularly pronounced nonwords and t-test results comparing IS and IM items

Data	Irregular-Single items	Irregular-Many items	t-value	df	p-value	Cohen's dz
Naming	0.31 (0.1)	0.39 (0.11)	5.64	54	<.001	0.76
Rating	4.62 (0.44)	5.31 (0.39)	14.35	54	<.001	1.94

Appendix 10

Human and model responses to experimental nonwords (Chapter 5)

Table 10A

Human and model responses to Irregular-Single (IS) and Irregular-Many (IM) nonwords

				Human participants						Computational models					
item	*type*	*TypeFreq*	*TokenFreq*	*first*	*n*	*second*	*n*	*third*	*n*	*DRC*	*CDP++*	*Psim1*	*WSP-type*	*WSP-token*	*WSP-type-T*
BEALM	IS	1	3.53	bilm	24	bElm	18	bIlm	8	bilm	bElm	bilm	bElm	bEIm	bElm
BLEIZE	IS	1	3.77	bliz	29	bl1z	11	bl2z	9	bl1z	bl1z	bl1z	bliz	bIiz	bliz
BRALF	IM	3	5.55	br{lf	46	brQlf	4	br#f	3	br{lf	br#lf	br$f	br#f	br#f	br#f
BREIRD	IS	1	4.75	br7d	34	br8d	7	brid	5	br1rd	br8d	br1rd	br7d	br7d	br7d
CHOUNG	IS	1	5.51	JQN	17	J6N	12	JuN	12	J6N	J6N_	k6N	JVN	JVN	J6N
CRORST	IS	1	4.92	kr$st	43	krQst	6	kr$s	2	kr$st	kr$st	kr$st	kr3st	kr3st	kr$st
DONTH	IS	1	5.08	dQnT	52	dQ	1	dVnT	1	dQnT	dVnT	dVnT	dVnT	dVnT	dVnT
DWOUTE	IS	1	4.6	dwut	29	dw6t	16	dw5t	2	dw6t	dwut	dw6t	dwut	dwut	dw6t
FLALD	IM	2	3.72	fl{ld	40	fl$d	4	flQld	3	fl{ld	fl{ld	fl{d	fl{ld	fI{Id	fl$ld
FLALT	IM	5	4.69	fl{lt	36	flQlt	5	fl6t	3	fl{lt	fl$lt	fl$t	fl{lt	fI{It	fl{lt
FLOUSSE	IS	1	3.76	flus	33	fl6s	14	flus1	3	fl6si	fl6s	fl6s	flus	fIus	flus
FOUSSE	IS	1	3.76	fus	39	f6s	8	fu	2	f6si	f6s	f6s	fus	fus	fus
FRALM	IM	5	4.72	fr{lm	35	fr#m	14	f{lm	1	fr{lm	fr#m	fr{m	fr#m	fr#m	fr#m
FREART	IS	1	5.3	fr#t	19	fr3t	13	fr7t	13	fr7t	fr7t	frirt	fr#t	fr#t	fr7t
GEALTH	IM	3	5.14	gElT	22	gIlT	21	gilT	7	_ilT	gElT	gElT	_ElT	_EIT	_ilT
GHASK	IM	6	4.66	g{sk	35	g#sk	18	g{Sk	1	g{sk	g#sk	g{sk	g#sk	g#sk	g{sk
GHEALT	IS	1	4.3	gElt	24	gIlt	17	gilt	7	gilt	gElt	gElt	gElt	gEIt	gilt
GHIGN	IM	10	4.98	g2n	17	gIn	12	gIN	4	gIn	g2n	In	g2n	g2n	gIn

Table 10A continued

				Human participants						Computational models					
item	*type*	*TypeFreq*	*TokenFreq*	*first*	*n*	*second*	*n*	*third*	*n*	*DRC*	*CDP++*	*Psim1*	*WSP-type*	*WSP-token*	*WSP-type-T*
GLALK	IM	5	5.52	gl{lk	33	gl$k	5	glQlk	4	gl{lk	gl{k	gl{k	gl$k	gl$k	gl$k
GLEART	IS	1	5.3	gl3t	22	gl7t	15	gl#t	9	gl7t	gl7t	gl$t	gl#t	gl#t	gl7t
GLIGH	IM	4	5.53	gl2	24	gllk	7	gli	5	gl2	gl2	gl2	gl2	gl2	gl2
GLILST	IS	1	4.63	gllist	39	gllst	4	gllst	4	gllist	gllist	gllist	gl2lst	gl2lst	gllist
GNOLK	IM	2	4.24	nQlk	32	gnQlk	13	n$k	4	nQlk	n5lk	n5	n5k	n5k	nQlk
HAUGE	IS	1	3.69	h$_	18	h1g	6	h6_	6	h$_	h$_	h$	h1_	h1_	h1_
JEALT	IS	1	4.3	_Elt	27	_llt	17	_ilt	10	_ilt	_Elt	_ilt	_Elt	_Elt	_ilt
KEALTH	IM	3	5.14	kElT	22	kilT	16	klIT	16	kilT	kilT	kilT	kElT	kElT	kElT
KYME	IM	2	4	k2m	53	kim	1	-	-	k2m	k2m	k2m	k2m	k2m	k2m
LORST	IS	1	4.92	l$st	53	l3s	1	l3st	1	l$st	l$st	l$st	l$st	l$st	l$st
MEARN	IM	3	5.01	m3n	38	m7n	14	min	2	m7n	m3n	mirn	m3n	m3n	m3n
MEIRD	IS	1	4.75	m7d	49	m3d	2	m2@d	1	m1rd	m7d	mird	m7d	m7d	m7d
NALM	IM	5	4.72	n{lm	34	n#m	16	nQlm	2	n{lm	n#m	nQm	n#m	n#m	n#m
NALT	IM	5	4.69	n{lt	30	nQlt	14	n$t	3	n{lt	n$lt	n{lt	n{lt	n$lt	n{lt
NORLD	IS	1	5.88	n$ld	45	n$d	8	n3ld	1	n$ld	n$ld	n$ld	n$ld	n$ld	n$ld
PHEALM	IS	1	3.53	fElm	30	fllm	11	film	10	film	fElm	ilm	fElm	fElm	fim
PHONGE	IS	1	4.12	fQn_	37	fQN	7	fVn_	4	fQn_	f5n_	f$n_	fVn_	fVn_	fQn_
PHOUP	IM	6	5.19	fup	46	f6p	4	f5p	1	f6p	f6p	up	fup	fup	f6p
PLALF	IM	3	5.55	pl{lf	39	f{lf	3	pl{f	3	pl{lf	pl#f	pl{f	pl#f	pl#f	pl#f
PLIGN	IM	10	4.98	pl2n	40	plIN	5	plIn	4	plIn	pl2n	pl2n	pl2n	pl2n	plIn
PSASK	IM	6	4.66	s{sk	30	ps{sk	5	ps{k	4	s{sk	p#sk	sp{sk	s#sk	s#sk	s{sk
PSORLD	IS	1	5.88	s$ld	20	ps$ld	15	s$l	4	s$ld	p$ld	sp$ld	s$ld	s3ld	s$ld
RHALD	IM	2	3.72	r{ld	39	rQld	9	r$d	2	r{ld	r{ld	hr{d	r$ld	r$ld	r{ld

Table 10A continued

				Human participants						Computational models					
item	*type*	*TypeFreq*	*TokenFreq*	*first*	*n*	*second*	*n*	*third*	*n*	*DRC*	*CDP++*	*Psim1*	*WSP-type*	*WSP-token*	*WSP-type-T*
RHOLK	IM	2	4.24	rQlk	45	r4k	2	r5k	2	rQlk	r5lk	r5k	r5k	r5k	rQlk
SHOUNG	IS	1	5.51	SQN	18	S6N	13	SuN	13	S6N	SVN	SV	SVN	SVN	SVN
SMACHT	IS	1	3.77	sm{kt	22	sm{J	12	sm{Jt	4	sm{Jt	sm{tt	sm{J	smQt	smQt	smQt
SMIQUE	IM	11	4.66	smik	32	sm2k	8	smIk	7	smIk	sm2k	sm2	smik	smik	smIk
SMYME	IM	2	4	sm2m	44	sm2mi	2	smim	2	sm2m	sm2m	sm2m	sm2m	sm2m	sm2m
SNAST	IM	14	5.97	sn{st	48	sn#st	4	n{st	2	sn{st	sn#st	sn{st	sn{st	sn{st	sn{st
SNAUGE	IS	1	3.69	sn6_	19	sn$_	9	sn$g	8	sn$_	sn$_	sn1_	sn1_	sn1_	sn1_
STONTH	IS	1	5.08	stQnT	55	-	-	-	-	stQnT	stVnT	stVnT	stVnT	stVnT	stVnT
TILST	IS	1	4.63	tIlst	53	tIst	1	zIls	1	tIlst	tIlst	tIlst	t2lst	t2lst	t2lst
TWEARN	IM	3	5.01	tw3n	47	tw7n	5	tw@n	1	tw7n	tw3n	tw3n	tw3n	tw3n	tw7n
TWEIZE	IS	1	3.77	twiz	42	tw2z	5	tw1z	3	tw1z	tw1z	tw2z	twiz	twiz	twiz
VOUP	IM	6	5.19	vup	49	v6p	3	fup	1	v6p	v6p	vup	vup	vup	v6p
VOUTE	IS	1	4.6	vut	38	v6t	10	fut	1	v6t	v6t	v6t	vut	vut	v6t
WONGE	IS	1	4.12	wQn_	32	wQN	18	w5n_	1	wQn_	wQn	wVn_	wVn_	wVn_	wQn_
YAST	IM	14	5.97	j{st	47	j#st	8	-	-	j{st	j#st	j{st	j{st	j{st	j{st
YIQUE	IM	11	4.66	jik	40	jIk	8	j2k	2	jIk	jik	j2	jik	jik	jIk
ZACHT	IS	1	3.77	z{kt	26	z{J	7	zQt	7	z{Jt	z{tt	QJ	zQt	zQt	z{t
ZALK	IM	5	5.52	z{lk	40	z$k	7	zQlk	6	z{lk	z$kk	z$lk	z$k	z$k	z$k
ZIGH	IM	4	5.53	z2	32	zIg	12	zi	4	z2	z2	2	z2	z2	zI

Table 10B

Model output not produced by any participant

Item	DRC	CDP++	Psim1	WSP-type	WSP-token	WSP-type-T
BREIRD	br1rd		br1rd			
CHOUNG		J6N_	k6N			
CRORST				kr3st	kr3st	
FREART			frirt			
FLALD						fl$ld
FLALT		fl$lt				
FLOUSSE	fl6si					
FOUSSE	f6si					
GEALTH	_ilT					_ilT
GHIGN			In			
GLEART			gl$t			
GLILST				gl2lst	gl2lst	
GNOLK		n5lk	n5	n5k	n5k	
MEARN			mirn			
MEIRD	m1rd		mird			
NALM			nQm			
NALT		n$lt			n$lt	
PHONGE			f$n_			
PHOUP			up			
PSASK			sp{sk			
PSORLD		p$ld	sp$ld		s3ld	
RHALD			hr{d			
RHOLK		r5lk				
SHOUNG			SV			
SMACHT		sm{tt				
SMIQUE			sm2			
STONTH		stVnT	stVnT	stVnT	stVnT	stVnT
TILST				t2lst	t2lst	t2lst
YIQUE			j2			
ZACHT		z{tt	QJ			z{t
ZALK		z$kk	z$lk			
ZIGH			2			zI

Appendix 11

Key Stimuli in Experiments 1 and 2 (Chapter 6)

Table 11A

List of Error and Odd Items in the Rating Task

Spelling	Pronunciation	Type
COSE	fr{kt	Error
DWAL	jEsts	Error
FRACT	prib	Error
GWI	mElS	Error
MELSH	rVpT	Error
PREBE	t$S	Error
RUPTH	zElms	Error
TORSH	vIpT	Error
YESTS	gw2	Error
ZELMS	k5z	Error
CHAIPSE	J$ps	Odd
DWEK	dwuk	Odd
FLOICE	fl#s	Odd
FROAP	fr1p	Odd
GLOOST	glEst	Odd
GOMPT	g6mpt	Odd
LONCH	l1nJ	Odd
NOATHE	niT	Odd
STOPSE	stIps	Odd
ZURVE	ziv	Odd

Table 11B

List of C and G-initial Items in the Rating Task

Item	Type	Hard pronunciation	Soft pronunciation
CELTH	C-critical	kElT	sElT
CEPTH	C-critical	kEpT	sEpT
CERSH	C-critical	k3S	s3S
CILSH	C-critical	kIlS	sIlS
CILTH	C-critical	kIlT	sIlT
CIRSH	C-critical	k3S	s3S
CALSH	C-control	k{lS	s{lS
CALTH	C-control	k{lT	s{lT
CORSH	C-control	k$S	s$S
CUPTH	C-control	kVpT	sVpT
GELSH	G-critical	gElS	_ElS
GELTH	G-critical	gElT	_ElT
GERD	G-critical	g3d	_3d
GILSH	G-critical	gIlS	_IlS
GIPTH	G-critical	gIpT	_IpT
GIRSH	G-critical	g3S	_3S
GALSH	G-control	g{lS	_{lS
GALTH	G-control	g{lT	_{lT
GORD	G-control	g$d	_$d
GULTH	G-control	gVlT	_VlT

Table 11C

List of nonwords named by the Unrelated-Rating group

Naming task items							
ANG	CLOUNT	FLOKE	HENCH	NITHE	SCRALK	SPRILT	TUNCH
ANK	CRAX	FOOF	HESE	NUM	SCRINTH	SQUEER	TUNG
BACHE	CRICHE	FOOK	HEST	PAULT	SHELSE	SQUOINT	TURGE
BAITH	CRITCH	FOW	HIBE	PAYST	SHESE	STAFT	TWEARN
BEAP	CROAST	FRALL	HINTH	PENTH	SHIG	STASK	TWOLL
BEM	CUKE	FRAUSE	HOUGH	PITE	SHILL	STAUSE	URE
BERGE	DANGE	FREC	HURST	PLALSE	SHINK	STENE	VACK
BIME	DECHE	FRULK	INE	PLARF	SHIVE	STIEGE	VIGN
BISE	DEIGH	FUSH	JEICH	PLEN	SHONG	STINE	VOSE
BIVE	DIFF	GEALTH	JILL	PLIGN	SHOULE	STUICE	VOUP
BLANCE	DOAL	GENVE	JOCK	POUGHT	SHRAS	STULL	WAIR
BLILL	DOAN	GEPTH	JOFF	POURT	SHRERE	SUILE	WAWL
BLOOT	DOM	GERR	KOH	POVE	SIVE	SUNT	WEATHE
BLOUCH	DOSH	GERSH	KUP	POWN	SKECHE	SURP	WEICH
BLUISE	DOUCH	GHETE	KYME	PRAUGH	SLEECH	TAM	WHAGUE
BOIST	DREAT	GHIGN	LAIT	PREL	SLELL	TARF	WHAISE
BOT	DRICHE	GIDGE	LALTZ	PRUMP	SLUE	TAY	WHETCH
BOUCHE	DRICK	GIEXT	LART	PSOOSH	SLYS	TEEP	WHYRE
BRAME	DRUIT	GINVE	LIDE	PUISE	SMINK	THAIPSE	WIB
BRETE	DRUMF	GLAPE	LILTH	QUIME	SMYS	THEAT	WOAT
BROAR	DULSE	GLEEZE	LIMPSE	QUOP	SNAITCH	THECHE	WOCK
BRULPT	DWAWSE	GLIGH	MANK	QUOUGH	SNEAR	THOG	WOLK
CEAT	EATH	GLUNT	MECK	RALL	SNEIR	TIX	WOOT
CELSH	EST	GNOMB	MEEP	RART	SOAST	TIZZ	WOTCH
CENVE	FAMP	GNOOSH	MELL	RHAWSE	SPAIL	TOWSE	WRAWSE
CHACH	FANT	GNUSE	MERSE	RHOISE	SPELP	TREAN	WRICHE
CHALM	FIDE	GOAK	MUNG	RIR	SPINT	TRELT	YAST
CIKE	FIFF	GRACH	NACH	ROFT	SPLAKE	TROME	YAUGHT
CINVE	FLALD	GROE	NALK	ROO	SPLALM	TROUNT	YIGHT
CIPTH	FLALSE	GRUZZ	NALT	ROWSE	SPLIE	TRURE	YIQUE
CLALF	FLEATHE	GUITT	NASTE	SCEACH	SPONCH	TUISE	ZAUSE
CLIB	FLEG	GWALF	NAWN	SCOP	SPOW	TUIT	
CLIX	FLO	HEAN	NIRST	SCRAFT	SPOWN	TUMPH	

Appendix 12

Results of the Experiment 1 Rating Data Analyses (Chapter 6)

Table 12A

Results of Repeated measures ANOVAs investigating the effects of Onset, Condition and Pronunciation on mean ratings for the C and G-initial items in each participant group

Effect	df	F	p-value	η_p^2
Rating-Only group (n = 68)				
Onset	1, 67	37.59	< .001	0.36
Cond	1, 67	102.64	< .001	0.61
Pron	1, 67	229.08	< .001	0.77
Onset x Cond	1, 67	3.10	0.08	0.04
Onset x Pron	1, 67	31.83	< .001	0.32
Cond x Pron	1, 67	242.87	< .001	0.78
Onset x Cond x Pron	1, 67	76.42	< .001	0.53
Naming-Rating group (n = 69)				
Onset	1, 68	5.52	0.02	0.08
Cond	1, 68	88.49	< .001	0.57
Pron	1, 68	196.34	< .001	0.74
Onset x Cond	1, 68	3.01	0.09	0.04
Onset x Pron	1, 68	46.95	< .001	0.41
Cond x Pron	1, 68	165.22	< .001	0.71
Onset x Cond x Pron	1, 68	87.14	< .001	0.56

Table 12B

Results of Repeated measures ANOVA investigating the effects of Condition and Pronunciation on mean ratings for the C and G-initial items at each level of Onset

Onset	Effect	df	F	p-value	η_p^2
Rating-Only group (n = 68)					
C	Cond	1, 67	73.6	< .001	0.52
C	Pron	1, 67	97.6	< .001	0.59
C	Cond x Pron	1, 67	208	< .001	0.76
G	Cond	1, 67	48.4	< .001	0.42
G	Pron	1, 67	244	< .001	0.78
G	Cond x Pron	1, 67	37.4	< .001	0.36
Naming-Rating group (n = 69)					
C	Cond	1, 68	42.7	< .001	0.39
C	Pron	1, 68	49.7	< .001	0.42
C	Cond x Pron	1, 68	152	< .001	0.69
G	Cond	1, 68	54.1	< .001	0.44
G	Pron	1, 68	265	< .001	0.80
G	Cond x Pron	1, 68	40.1	< .001	0.37

Appendix 13

Results of the Experiment 2 Rating Data Analyses (Chapter 6)

Table 13A

Results of Repeated measures ANOVAs investigating the effects of Onset, Condition and Pronunciation on mean ratings for the C and G-initial items in each participant group

Effect	df	F	p-value	η_p^2
Unrelated-Rating group (n = 62)				
Onset	1,61	26.13	< .001	0.30
Cond	1,61	128.72	< .001	0.68
Pron	1,61	228.36	< .001	0.79
Onset x Cond	1,61	2.51	0.12	0.04
Onset x Pron	1,61	12.81	< .001	0.17
Cond x Pron	1,61	162.73	< .001	0.73
Onset x Cond x Pron	1,61	72.11	< .001	0.54
Naming-Rating-type group (n = 55)				
Onset	1,54	2.91	0.09	0.05
Cond	1,54	64.11	< .001	0.54
Pron	1,54	203.26	< .001	0.79
Onset x Cond	1,54	2.38	0.13	0.04
Onset x Pron	1,54	26.06	< .001	0.33
Cond x Pron	1,54	89.22	< .001	0.62
Onset x Cond x Pron	1,54	43.56	< .001	0.45

Table 13B

Results of Repeated measures ANOVA investigating the effects of Condition and Pronunciation on mean ratings for the C and G-initial items at each level of Onset in each group

Onset	Effect	df	F	p-value	η_p^2
	Unrelated-Rating group (n = 62)				
C	Cond	1,61	76.70	< .001	0.56
C	Pron	1,61	85.60	< .001	0.58
C	Cond x Pron	1,61	142.00	< .001	0.70
G	Cond	1,61	77.40	< .001	0.56
G	Pron	1,61	159.00	< .001	0.72
G	Cond x Pron	1,61	37.40	< .001	0.38
	Naming-Rating-type group (n = 55)				
C	Cond	1,54	46.60	< .001	0.46
C	Pron	1,54	69.80	< .001	0.56
C	Cond x Pron	1,54	92.00	< .001	0.63
G	Cond	1,54	35.20	< .001	0.40
G	Pron	1,54	256.00	< .001	0.83
G	Cond x Pron	1,54	17.70	< .001	0.25