Simplifying Reading: Applying the Simplicity Principle to Reading

Janet I. Vousden, a Michelle R. Ellefson, b Jonathan Solity, c Nick Chater d

aDepartment of Psychology, Coventry University
bPsychology and Neuroscience in Education, Faculty of Education, University of Cambridge
cEducational Psychology Group, Department of Psychology, University College London
dBehavioural Science Group, Warwick Business School, University of Warwick

Received 3 March 2009; received in revised form 26 May 2010; accepted 28 May 2010

Abstract

Debates concerning the types of representations that aid reading acquisition have often been influenced by the relationship between measures of early phonological awareness (the ability to process speech sounds) and later reading ability. Here, a complementary approach is explored, analyzing how the functional utility of different representational units, such as whole words, bodies (letters representing the vowel and final consonants of a syllable), and graphemes (letters representing a phoneme) may change as the number of words that can be read gradually increases. Utility is measured by applying a Simplicity Principle to the problem of mapping from print to sound; that is, assuming that the “best” representational units for reading are those which allow the mapping from print to sounds to be encoded as efficiently as possible. Results indicate that when only a small number of words are read whole-word representations are most useful, whereas when many words can be read graphemic representations have the highest utility.

Keywords: Psychology; Representational units of reading; Mathematical modeling

1. Introduction

The ability to translate a printed word into its spoken form is a fundamental skill beginning readers must master. There are a number of different ways that this may be achieved. For example, it might be possible to learn to associate the shape of a written word with its phonological (spoken) form. Another strategy might be to learn to associate smaller “chunks” of text, for example, on a letter-by-letter basis, with their phonological form and
to build words by decoding each chunk in turn. Thus, different representational units may be used to read. Identifying which representational units are used during reading acquisition in English has generated considerable debate and has been frequently influenced by examining the types of phonological units that pre- and beginning readers are able to manipulate (e.g., Hulme et al., 2002). However, it is also clear that the choice of representational unit will be influenced by the transparency between the print and sound of a language. Furthermore, the transparency between print and sound within a language may change over time—as the number of words that can be read increases—and the preferred type of representational unit may also vary accordingly.

The aim of the current article is to examine which representational units should be most useful for reading acquisition by focusing attention on the structure of the print to sound mapping in English. A key objective is to consider how the type of representational unit best suited to facilitate reading acquisition may change as the reading vocabulary increases. In pursuing these questions, we aim to develop a theoretically motivated account of why particular types and specific instances of representational units might be preferred over others, and to consider the implications of our findings for reading instruction. We examine these objectives using the Simplicity Principle, which operates on the basis of choosing simpler explanations of data over complex ones—and here favors representational units that allow the mapping from print to sound to be specified as simply as possible. This approach emphasizes maximizing the utility of various representational units by trading off the complexity of the solution against the ability to account for the data. A number of mechanisms may exist that can implement the solution identified, although the approach is not directly dependent upon specific underlying representations or processes at an implementational level.

1.1. Simplifying reading: Applying the Simplicity Principle to reading

The spelling-to-sound mapping of English can be distinguished from other languages by the fact that the same spelling may be pronounced in different ways; for example, the letters “ea” are pronounced differently depending on the context in which they occur (e.g., beach, real, head, great, etc.). This contrasts with many other European languages in which pronunciation is determined unambiguously by spelling (e.g., Serbo-Croat, Greek), and character-based languages such as Chinese, where the written language provides an even less transparent guide to pronunciation. Thus, some English spelling patterns are consistent (e.g., ck—⁄k⁄) and a simple, small unit (grapheme–phoneme) rule suffices, but others (e.g., ea) require a more sophisticated unit rule for accurate decoding. This inconsistency clearly increases the difficulty of learning to read compared with learning a language using a more consistent orthography (Landerl, 2000; Wimmer & Goswami, 1994). Most often, the vowel together with the following consonant(s) (i.e., a word body, such as -eap) provides a better cue to pronunciation (Treiman, Mullennix, Bijeljacbabic, & Richmondwelt, 1995). However, even word bodies can be inconsistent (Vousden, 2008). Alternatively, frequently encountered words can be learned “by sight,” without any decoding, and inconsistency is largely eliminated. A word can therefore be decoded by reference to units of different sizes: whole words (beach), bodies (b-each), or graphemes (b-ea-ch). In general terms, the smaller
the unit, the less there is to remember, but this comes at the expense of greater inconsist-

cency. The question faced by the cognitive system is then: What is the optimal way to repre-
sent such an orthographic to phonological mapping system? The answer to this question is

important for teaching beginning readers: If the optimal representation can be found, then it

should be of value when considering instructional materials.

The optimal representational units for reading may differ across languages and ortho-

graphic systems (Ziegler & Goswami, 2005). More regular orthographies, for example, can

more reliably be read using small linguistic units; highly irregular orthographies appear to

require encoding via larger “chunks.” But note, too, that even within a language and an

orthography, the optimal representation may also shift over time. We should expect that, just

as the cognitive system is able to adapt to different orthographies, so too is it able to adapt

to changes in the different number or types of words encountered within the same orthogra-

phy. Early readers are exposed to subsets of written words of the language as they learn; we

define these subsets here as a reading vocabulary. Note that this is different from a receptive

or productive vocabulary, which may include many words that are known but that cannot

yet be read. In general, we should not expect that the best representations for capturing a

small reading vocabulary will be the same as the best representations for capturing a much

larger reading vocabulary. This expectation raises the interesting possibility that, as readers

progress, their optimal representations should change—the increase in reading vocabulary

providing a potential driving force for developmental change, independent of any exogenous

changes in the cognitive machinery of learning, during development (e.g., changes in

aspects of short-term verbal memory). The possibility that, as the beginning reader’s linguist-
ic exposure gradually increases, the reader’s optimal representation of the print-sound map-
ing changes, provides an interesting possible explanation for representation change driving

reading development, and one that we shall explore further below. For a small reading

vocabulary, for example, it may be that simply memorizing the pronunciation of each word

requires fewer cognitive resources than learning many more seemingly arbitrary sublexical

decoding strategies to read a small number of words.

The structure of this article is as follows. We begin with the section Representational

Units and Learning to Read by reviewing the evidence that different representational units

are involved during reading acquisition. Next, in Spelling-to-Sound Consistency and Learn-
ing to Read we review cross-linguistic evidence that shows how inconsistency in the print to

sound mappings complicates the choice of representational units for English. In The Sim-

plicity Principle, we present a nontechnical description of the Simplicity Principle and

describe how it can be applied to reading to trade off maximizing reading outcome against

minimizing complexity. The findings from the simplicity-based analyses are presented next.

Analysis 1 explores which general type of representational unit should best facilitate reading

acquisition, while Analysis 2 examines which specific representational units should be most

useful. Analysis 3 considers the effect of an increasing vocabulary on the choice of repre-
sentational unit, and Analysis 4 explores the extent to which choosing among inconsistent

units can be aided by considering context-sensitive units. Thus, this is not a direct study of

human reading, but an evaluation of some hypotheses concerning the representations of

orthographic to phonological translation that might be learned during reading acquisition.
Finally, in the Discussion, we discuss the results of the analyses and the implications for instruction.

1.2. Representational units and learning to read

The status of different unit sizes in terms of reading outcome has received much attention in recent decades. There is a general consensus that phonological awareness (the ability to deal explicitly with sound units at a subsyllabic level), together with knowledge of how spelling patterns represent speech sounds, is central to reading outcomes (Goswami & Bryant, 1990; Gough, Ehri, & Treiman, 1992; Wagner & Torgesen, 1987; Wyse & Goswami, 2008). Tasks that measure either phoneme awareness, for example, how well children can say words after deleting the last sound (CAT -> CA), or rime (VC) awareness, for example, whether children can judge whether two words rhyme (do HAT and CAT rhyme?), are strongly correlated with reading ability. However, there has been less agreement on whether early measures of phoneme awareness (Hulme et al., 2002; Hulme, Muter, & Snowling, 1998; Muter, Hulme, Snowling, & Taylor, 1997; Nation & Hulme, 1997) or rime awareness (Bradley & Bryant, 1978; Bryant, 1998; Bryant, Maclean, Bradley, & Crossland, 1990; Maclean, Bryant, & Bradley, 1987) best predict later reading ability. There is more agreement, however, on the development of phonological awareness; children become aware of large phonological units (words, syllables, rimes) before small units such as phonemes (Carroll, Snowling, Hulme, & Stevenson, 2003; Goswami & Bryant, 1990). Furthermore, theories of reading development have assumed development occurs in stages (Frith, 1985; Marsh, Friedman, Welch, & Desberg, 1981), from a large unit, logographic stage, through a small unit alphabetic (or grapheme–phoneme) stage, and finally a stage where more advanced context-dependent associations are available.

However, recent evidence suggests that prereaders are not restricted to a purely logographic strategy and are able to apply sublexical knowledge very early on in development. For example, prereaders with some knowledge of letter names learned phonetically motivated spelling-sound pairs such as AP—ape more easily than arbitrary (large-unit or logographic) pairs such as ID—oat (Bowman & Treiman, 2008), and letters with phonologically similar letter names and sounds, for example, /bi:/ and /bθ/ for the letter B, are learned before letters without that phonological similarity, for example, /waɪ/ and /jæ/ for the letter Y (Ellefson, Treiman, & Kessler, 2009). Furthermore, Treiman and colleagues have shown that context-dependent associations are also available early on in development. For example, in a task where young children are taught pronunciations of novel graphemes in pseudowords, they are able to make use of appropriate context-sensitive associations (onset–vowel, or head, and vowel–coda, or rime) in a subsequent transfer task (Bernstein & Treiman, 2004; Treiman, Kessler, Zevin, Bick, & Davis, 2006).

These results suggest that children are able to take advantage of the statistical structure of text, as and when it arises, regardless of the type of unit per se. However, although there is a wealth of evidence suggesting that language acquisition is facilitated by exploiting multiple cues in the input (e.g., Christiansen, Allen, & Seidenberg, 1998; Monaghan, Chater, & Christiansen, 2005), the exploitation of such findings in terms of instruction has not been
apparent. This observation is particularly salient given that environmental factors (e.g., exposure to books, growing up in a stimulating environment) are likely to contribute to literacy problems (Bishop, 2001). Our assumption, consistent with rational analysis theory (Anderson, 1990; Anderson & Schooler, 1991; Oaksford & Chater, 1998), is that these statistical properties (at multiple levels) should play an important role in learning to read because they guide our adaptation to the vocabulary to be acquired. We note that both the specific lexical knowledge, in the form of exactly which words children experience when learning to read, and the number of words they have learned (i.e., vocabulary size) will act together to form the lexical vocabulary to which the cognitive system adapts. This perspective may lead to shifts of optimal strategy as both factors change during learning. In other words the optimal strategy for a small specific vocabulary may not be the same as for a much larger general vocabulary.

### 1.3. Spelling-to-sound consistency and learning to read

The problem of inconsistency for English spelling is most obvious at the grapheme level. For example, a recent listing of grapheme–phoneme probabilities lists some vowel graphemes as having nine pronunciations (Gontijo, Gontijo, & Shillcock, 2003). Estimates of the orthographic depth of English (e.g., the average number of pronunciations each grapheme has) range from 2.1 (Berndt, Reggia, & Mitchum, 1987) to 2.4 (Gontijo et al., 2003) for polysyllabic text, to 1.7 (Vousden, 2008) for monosyllabic text. This result can be contrasted with languages such as Serbo-Croat that has an orthographic depth of 1. Thus, even monosyllabic text, items to which beginning readers of English are more likely to be exposed, is still more inconsistent than other languages. This inconsistency is problematic for beginning readers for obvious reasons: How does a beginning reader choose between alternative pronunciations of a given grapheme? It is the inconsistency of English that sets it apart from other orthographies—when compared with other major European languages, it is judged to be the least consistent (Borgwaldt, Hellwig, & De Groot, 2005; Seymour, Aro, & Erskine, 2003). This inconsistency is thought to be the cause of difficulty in learning to read, when compared with other more consistent orthographies (e.g., Henderson, 1982; Landerl, 2000; Wimmer & Goswami, 1994; Ziegler & Goswami, 2005).

Comparisons of reading performance between learners of consistent and inconsistent language have been well documented (for a review, see Ziegler & Goswami, 2005). Typically, the acquisition of spelling-to-sound knowledge is measured using nonword reading performance because decoding nonwords must be done by the application of spelling-to-sound rules. Early comparisons of monolingual studies reveal that while children who learn consistent orthographies perform at between 80% and 89% on nonword reading tasks, where percent correct is determined according to whether the regular or most dominant pronunciation of the graphemes in the nonword is produced (Cossu, Gugliotta, & Marshall, 1995; Porpodas, Pantelis, & Hantzou, 1990; Sprenger-Charolles, Siegel, & Bonnet, 1998), English children lag behind at 45% (Frith, Wimmer, & Landerl, 1998). Cross-linguistic studies, where tighter control over items and subjects can be exerted, show similar results, with English children performing at between 12% and 51% correct, compared with children that learn more
consistent non-English orthographies performing typically above 90% correct (Ellis & Hopper, 2001; Frith et al., 1998; Goswami, Gombert, & de Barrera, 1998; Seymour et al., 2003).

The Psycholinguistic Grain Size (PGS) account (Ziegler & Goswami, 2005) offers an explanation of these cross-linguistic differences. The core assumption behind PGS is that the word reading process adapts to the consistency of the orthography with which it is faced: Consistent orthographies can rely on grapheme–phoneme correspondences (GPCs), whereas more inconsistent orthographies require the formation of larger sublexical units to resolve those inconsistencies, in addition to GPCs (see Ziegler & Goswami, 2005 for a review of the evidence supporting these claims). In addition, learning an inconsistent language is constrained by the granularity problem—there are many more mappings to learn for larger psycholinguistic units (e.g., rimes). Therefore, according to PGS, it seems to make sense to highlight the consistencies within the language across grain sizes and optimize the orthographic-phonological learning task accordingly to maximize consistency, minimize the effects of granularity, and hence maximize overall performance. While there is plenty of evidence that cross-linguistic differences in consistency account for differences in early reading performance (e.g., Seymour et al., 2003), there is also some data that suggest even a basic attempt to maximize consistency within an orthography improves performance for inconsistent orthographies (Shapiro & Solity, 2008).

Evidence that highlighting the most consistent aspects of the orthography improves performance comes from recent modeling of the cross-linguistic differential in reading performance (Hutzler, Ziegler, Perry, Wimmer, & Zorzi, 2004), and behavioral data that compares teaching methods for children learning to read English (Landerl, 2000; Shapiro & Solity, 2008; Solity & Shapiro, 2008). In a recent attempt to model the cross-linguistic differences in reading performance between German and English children, Hutzler et al. (2004) found that a two-layer associative network (Zorzi, Houghton, & Butterworth, 1998) was only able to simulate the cross-linguistic reading data once the model had been pre-trained with a set of highly consistent grapheme–phoneme correspondences. Comparison of the model data before and after pretraining, for both English and German words, showed a marked improvement in performance, suggesting that learning even an inconsistent orthography such as English can be improved by highlighting the most consistent correspondences at the grapheme level.

Behavioral data appear to support this theoretical suggestion. In two studies designed to tease apart the effects of instruction and orthographic consistency on reading performance, Landerl (2000) compared the reading performance of German and English children who had received different instruction. Phonics instruction involves conveying an understanding of how spelling patterns (graphemes) relate to speech sounds (phonemes), and how this knowledge can be applied to “sound out” words. In Landerl’s study, English children either received a structured phonics approach (grapheme–phoneme correspondences are introduced in a predetermined and systematic manner, e.g., Lloyd, 1992), or a mix of phonics and whole-word instruction. German children received phonics instruction, the dominant approach in Germany. English children receiving only phonics instruction outperformed English children who received the mixed approach, and they were almost as accurate as the German children when reading nonwords. More recent work backs up these findings. Chil-
dren from the United Kingdom who were taught phonic skills using a restricted set of frequent, highly consistent grapheme–phoneme correspondences not only improved their reading performance significantly faster than children who were taught a wider range of grapheme–phoneme correspondences (including less consistent ones), but significantly fewer of them were classified as having reading difficulties (Shapiro & Solity, 2008).

In English, most of the orthographic inconsistency stems from the multiple pronunciations for vowel graphemes. Extensive analyses have shown that the problems of grapheme inconsistency can be alleviated by considering a larger unit: the body (Kessler & Treiman, 2001; Peereeman & Content, 1998; Treiman et al., 1995)—that is, the consonants following the orthographic vowel generally predict its pronunciation better than the preceding consonants (Treiman & Zukowski, 1988). Generally, body units are more consistent than graphemes. In an analysis of English monosyllabic text, Vousden (2008) found that while 39% of graphemes are inconsistent (i.e., 39% of all graphemes can be pronounced in more than one way—for example, ea can be pronounced as in head, beach, great, etc., but the rest are only ever pronounced one way—t is always pronounced as in tap), only 16% of onset units (e.g., wh- can be pronounced as in who, white) and 18% of body units (e.g., -arm can be pronounced as in warm, farm) are inconsistent. In terms of predicting pronunciation, onset and body units together predict pronunciation better than graphemes (Vousden, 2008) due to their greater consistency. However, the apparent benefit afforded by bodies is compromised by the granularity problem (Ziegler & Goswami, 2005): The number of orthographic units to learn increases as the size of the orthographic unit increases. Thus, there are many more body units than graphemes. Vousden (2008) has shown that if the most frequent mappings at each grain size are chosen, then graphemes are better at predicting pronunciation when a smaller number of mappings are known; however, if a large number of mappings are known, then onsets and bodies are better at predicting pronunciation.

In sum, it would appear that reducing inconsistency might improve reading performance for beginning readers. As highlighted above, one way to accomplish this for English may be to concentrate on larger “chunks” (Ziegler & Goswami, 2005). However, learning larger “chunks” comes at the expense of an increased load for the cognitive system. The question then raised concerns how an optimal reading system trades off the load of representing more mappings against increasing predictability of pronunciation, and how the most consistent units across multiple levels can be combined within the same system. Our aim is to explore to what extent different sized representational units are useful in English, an inconsistent language, using the simplicity principle.

### 1.4. The Simplicity Principle

At a fairly general level, learning to read could be interpreted as a search for patterns in the orthographic to phonological translation of the language, which is broadly consistent with both rational analysis and PGS. According to Ehri (1992, 1998), these patterns act as the basis of the links between the orthographic and phonological representations of words, which then form the basis of later automatic sight reading—where skilled readers can automatically access pronunciations without resorting to the slower decoding processes characteristic of
beginning readers. As such, the orthography to be searched (the data) may be consistent with any number of possible patterns, and the problem is how to choose between alternative patterns.

The Simplicity Principle is based on the assumption (similarly to Ockham’s razor) that simpler explanations of data should be preferred to complex explanations (Chater, 1997, 1999; Chater & Vitányi, 2003). The Simplicity Principle can be mathematically expressed using the theory of Kolmogorov complexity (Kolmogorov, 1965), which defines the complexity of an individual object as the length of the shortest program in a universal programming language (any conventional programming language) that re-creates the object (Li & Vitányi, 1997) and therefore provides an objective measure of complexity. Importantly, the length of the program is independent of the particular universal programming language chosen, up to a constant. Of course, the preferred explanation must be able to re-create the data. According to simplicity, the preferred explanation is one that re-creates the data but can be described most succinctly. Thus, the cognitive system should strive to maximally compress the data such that the data can be reconstructed with maximum accuracy from its compressed form. In this sense, the Simplicity Principle provides an objective basis for choosing among many compatible patterns. The Simplicity Principle appears to be consistent with a range of empirical data across a range of cognitive domains, including perception (Chater, 1996, 2005), categorization (Feldman, 2000; Pothos & Chater, 2002), similarity judgments (Hahn, Chater, & Richardson, 2003), and various data within the linguistic domain (Brent & Cartwright, 1996; Brighton & Kirby, 2001; Chater, 2004; Dowman, 2000; Ellison, 1992; Goldsmith, 2002, 2007; Goldsmith & Riggle, 2010; Onnis, Roberts, & Chater, 2002; Perfors, Tenenbaum, & Regier, 2006; Roberts, Onnis, & Chater, 2005).

A practical methodology in statistics and information theory that embodies the principles of simplicity and Kolmogorov complexity is the Minimum Description Length (MDL) principle (Rissanen, 1989; see also Wallace & Freeman, 1987). MDL is based on the notion that the more regularity there is in the data, the more it can be compressed, and hence described succinctly. It requires simply that the length of code required to specify an object (the description length) be computed, and not the actual code itself. The description length can be calculated using information theory (Shannon, 1948) if a probability can be associated with an event, or pattern. Using standard information theory in this way, more probable events are associated with shorter code lengths. Data that contain no regularities or patterns cannot be compressed and are best described by (reproducing) the data itself. However, where patterns exist and some events are highly probable, the description length of an object will be reduced. Therefore, MDL provides a measure of simplicity by specifying the length of the binary code necessary to describe that object. The binary code must represent two components. First, it must provide a measure of the hypothesis (spelling-to-sound mappings) that describes the route from print to sound. Second, it must provide a measure of how well the data can be re-created, given the hypothesis under consideration. Put more formally:

\[ L = L(H) + L(D|H) \]

where \( L \) is the total description length, \( L(H) \) is the description length of the current hypothesis, and \( L(D|H) \) is the description length of the data under the current hypothesis.
In the current context, it is useful to think of (a) a set of spelling-to-sound mappings as a hypothesis (H), and (b) how well those spelling-to-sound mappings describe the target pronunciations as the data given the hypothesis (D, given H). From a simplicity perspective, the optimal reading system should trade-off the number of spelling-to-sound mappings represented (complexity of the hypothesis, L(H)) against the ability to pronounce existing and novel words as concisely as possible (goodness-of-fit to data, L(D|H)). The description length of each hypothesis will vary, as will the description length of the data given each different hypothesis. The total description length of a hypothesis (L) is therefore a measure of its simplicity. Full algorithmic details concerning the calculation of the description length using information theory (Shannon, 1948) for both hypotheses and the data, given each hypothesis, are given in Appendix; a brief summary of the mechanistic calculation of word pronunciation is given below.

The underlying mechanism that produces pronunciations from mappings in the simplicity analyses that follow assumes a probabilistic rule-based mechanism that could be instantiated in a number of ways; for example, implicitly as a two-layer associative network, consistent with models that are able to represent the statistical regularities that exist between orthography and phonology (Perry, Ziegler, & Zorzi, 2007; Zorzi et al., 1998). However, because we are interested in comparing different sizes of units (the effects of different sized units occur within the same architecture in connectionist models) and different mappings within unit sizes, it has been implemented here as a simple, explicit, probability-based production system. Multiple rules for orthographic units are represented and applied probabilistically; therefore, the production system is more similar to two-layer associative networks (Perry et al., 2007; Zorzi et al., 1998) than the symbolic sublexical rule system of the DRC (Coltheart, Curtis, Atkins, & Haller, 1993).

A simple, probabilistic rule-based mechanism for generating pronunciations is appealing in the context of this study because it is straightforward to see how using such a mechanism could apply to instruction (i.e., applying spelling-sound knowledge through blending to form words). The rules are applied probabilistically, based on the directionless type frequency with which individual orthographic and phonological representations are associated. Such frequencies (termed sonograph frequencies; see Appendix for further discussion of frequency; see Data S1 and Data S2 for the actual grapheme–phoneme frequencies used throughout the following analyses) play a significant role in predicting young children’s reading accuracy (Spencer, 2010). The probabilistic rule system implemented here is able to produce the target (correct) pronunciation for all words, although sometimes the target pronunciation may have a lower probability than an alternative pronunciation. The relative probability with which the target pronunciation is output by the rule system provides the basis for the calculation of the description length of the data under the current hypothesis, as described above. Thus, the description length of the data when the target pronunciation is the most probable output will be shorter than when the target pronunciation has a lower probability than an alternative (incorrect) pronunciation. When the target pronunciation is not the most probable output, additional description is required to specify which output is the target pronunciation. Thus, all words can be pronounced correctly, but the description length for all target pronunciations will vary according to how many words can be assigned
the shortest code length (their pronunciations are most probable) and how many require additional description, over and above that specified by the hypothesis. The less likely the current hypothesis finds the target pronunciation to be, the longer the code will be to describe it under the current hypothesis.

We do not assume that skilled reading is based on the same explicit mechanism, only that a simple mechanism that assumes a direct association between orthography and phonology forms the basis of the underlying associations between the orthographic and phonological representations of words on which later automatization occurs. We return to the issue of mechanisms and models in the Discussion section.

In the next section, we describe a series of analyses and discuss how well they match relevant human data. The first two analyses show how the utility of different types of representational units varies considerably within large reading vocabularies, but is qualitatively similar across children’s and adults’ reading vocabularies. To anticipate the results, the data are most concisely described by grapheme-sized units. The analyses also show how, within a type of representational unit, some specific units (graphemes) are more useful than others. The third analysis shows how the utility of different types of representational units varies with the size of the reading vocabulary. This reveals a shift in preference from large units to small units as the vocabulary increases. The last analysis reveals a pattern in which large representational units have high utility for large vocabularies, when carefully selected within the context of preferred small units.

2. Analysis 1: Which type of representational unit should best facilitate reading acquisition?

It has previously been shown that rime (VC) units are more consistent and provide a more accurate guide to pronunciation than both head (CV) and individual grapheme units (Treiman et al., 1995), although head units can in some circumstances influence pronunciation as well as rimes (Bernstein & Treiman, 2004; Treiman, Kessler, & Bick, 2003; Treiman et al., 2006). Overall, pronunciation is better predicted by the application of onset and rime sized correspondences than by the application of grapheme–phoneme correspondences (Vousden, 2008), but many more onset–rime units are required to achieve accurate pronunciation (Solity & Vousden, 2009; Vousden, 2008), thereby increasing the complexity of the solution to be adopted by the cognitive system, an issue that has received little consideration. For example, Solity and Vousden (2009) showed that the amount of monosyllabic adult text that could be accurately decoded by applying 64 grapheme–phoneme correspondences would require the application of 63 onset and 471 rime correspondences. Following the literature that found a developmental progression from large to small unit phonological awareness, and a number of studies that suggested training children about onset and rime could be beneficial for reading acquisition (Goswami, 1986; Wise, Olson, & Treiman, 1990), there emerged a large literature comparing different methods of instruction based on different unit sizes, from whole words (e.g., I. S. Brown & Felton, 1990) and rimes (e.g., Greaney, Tunmer, & Chapman, 1997), to graphemes (e.g., Stuart, 1999).
Our aim for this analysis was to ask which type of representational unit should best facilitate reading acquisition. Four different types of spelling-to-sound mapping (whole-word, head–coda, onset–rime, and grapheme) were considered separately as hypotheses about the data (the spelling-to-sound translation of words from each database), and the simplicity of each hypothesis was calculated. This facilitated an exploration of the impact of different cognitive variables (embodied by each hypothesis) on potential reading. The separate analyses for each unit size allow comparison with the existing reading instruction literature, and they also serve as a baseline for later analyses where hypotheses based on different unit sizes are optimized and combined.

2.1. Method

2.1.1. Databases

To examine the text that an adult reader is exposed to, we used the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995), which uses frequency counts from the COBUILD/Birmingham corpus (Sinclair, 1987). This 17.9 million-word token corpus mainly consists of written text from 284 mostly British English sources. For children’s text, a word frequency list based on children’s early reading vocabulary was used (Stuart, Dixon, Masterson, & Gray, 2003). This database was based on the transcription of 685 children’s books used by 5- to 7-year-olds (i.e., the first 3 years of formal instruction) in English schools.

Both databases were restricted to monosyllabic words only and were reduced by removing proper names, abbreviations, interjections, and nonwords. This procedure resulted in a total of 3,066 word types for the children’s database and 7,286 word types for the adults’ database.

2.1.2. Procedure

The simplicity of the whole-word, head–coda, onset–body, and grapheme hypothesis was calculated by following the calculations described in detail in the Appendix, resulting in a total description length measure (L) for each hypothesis. This measure was the sum of two lengths—the description length of the hypothesis, L(H), that measured the complexity of the hypothesis, and the description length of the data, L(D|H), that measured the goodness-of-fit to the data. Each hypothesis had a separate measure of simplicity (L = L(H) + L(D|H)) for each database.

In order to explore how well mappings derived from children’s text could generalize to describe the spelling-to-sound translation for adult text, another measure of simplicity was calculated. This measure of simplicity consisted of the description length of hypotheses generated from the children’s database—L(HChild), and the description length of the data from the adults’ database, given those hypotheses—L(DAdult|HChild). This resulted in a total description length for each hypothesis, which measured generalization (L = L(HChild) + L(DAdult|HChild)).

2.2. Results and discussion

The number of mappings required to describe the spelling-to-sound translation for each database is listed in Table 1. First, hypotheses that incorporate large segment sizes—the
head–coda and onset–body segmentations—require many more mappings to describe the same amount of data. Both head–coda and onset–body hypotheses require between six and seven times more mappings than the grapheme-sized hypothesis. Second, it appears that the sets of large-segment mappings required to describe children’s reading vocabulary are less adequate to also describe adults’ reading vocabulary than the grapheme-sized hypothesis. This result can be seen by the greater increase in number of mappings required to describe the CELEX versus the Stuart et al. (2003) database for head–coda and onset–body mappings compared to a more modest increase for grapheme-based mappings. Thus, overall, head–coda and onset–body mappings appear more complex and less generalizable than grapheme-sized mappings.

A comparison of the different sized mappings in terms of total description length ($L$) can be seen in Fig. 1A and B for the children’s and adults’ databases, respectively. The pattern is the same for both databases: Grapheme-sized mappings have the shortest total description length, and head–coda and onset–body mappings have total description lengths that are shorter than the data itself (as shown by the total description length for whole words) but longer than that of grapheme-based mappings. The description length of the hypotheses ($L(H)$) is a function of both the complexity of the mappings (i.e., larger segment mappings will have longer description lengths) and the total number of mappings that make up the hypothesis. Thus, because head–coda and onset–body mappings are more complex and numerous (Table 1) than grapheme-sized mappings, their description lengths are longer.

However, inspection of Fig. 1A and B reveals a different picture for the description lengths of the data, given the hypothesis ($L(D|H)$). Again, the pattern is similar for both databases—pronunciation is more concisely described by onset–body mappings than either head–coda or grapheme-based mappings. This result reflects the greater certainty with which the target pronunciation occurs when the mapping size is based on onset–body units as shown in Fig. 1C and D for children’s and adults’ text, respectively. Fig. 1C and D shows the proportion of target pronunciations that are output as most probable (i.e., accurate pronunciations) under each hypothesis; this is a measure of how accurately a hypothesis can account for data and is the basis of the description length of that data. The measure illustrates the extent to which a hypothesis is able to minimize its description of the data—the shortest description lengths of data ($L(D|H)$) will be obtained for hypotheses where a large proportion of the data is accurately reproduced—rather than a comparison of human performance data.

Table 1
Total number of mappings for each hypothesis, for both databases

<table>
<thead>
<tr>
<th>Mapping size</th>
<th>Stuart et al. (2003)</th>
<th>CELEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole words</td>
<td>3,066</td>
<td>7,286</td>
</tr>
<tr>
<td>Heads and codas</td>
<td>1,285</td>
<td>1,944</td>
</tr>
<tr>
<td>Onsets and rimes</td>
<td>1,141</td>
<td>2,070</td>
</tr>
<tr>
<td>Graphemes</td>
<td>237</td>
<td>311</td>
</tr>
</tbody>
</table>
Fig. 2A shows the description lengths for the adult database given the head–coda, onset–body, and grapheme hypotheses derived from the children’s database. The description length of each hypothesis ($L(H_{Child})$) is the same as that plotted in Fig. 1A; however, the description length of the adult database ($L(D_{Adult}|H_{Child})$) has increased—most notably for the head–coda and onset–body hypotheses. Fig. 2B shows the size of this increase—2.3 times and 5.2 times more code, respectively, was needed to describe the adult database using the head–coda and onset–body hypotheses based on the children’s database than when using hypotheses based on the adult database ($L(D_{Adult}|H_{Adult})$). In contrast, the increase in description length for the adult database was a modest 1.18 times for the grapheme hypothesis. The differential increase in description length occurs because the graphemes derived from the children’s database generalize better to the adult database than the head–coda and onset–body mappings.

Fig. 2C compares the proportion of words in the adult database where the target pronunciation was the most probable for hypotheses derived from the children’s and adults’ database. The results reflect those in Fig. 2B. For the onset–body hypothesis, fewer target pronunciations were output as the most probable when the hypothesis was derived from the children’s database than when it was derived from the adults’ database. However, this pattern is not the case for the grapheme hypothesis; similar numbers of target pronunciations
were output as the most probable regardless of whether the hypothesis was derived from the children’s or adults’ database. This suggests that grapheme-sized mappings derived from children’s text generalize well to adults’ text.

These results are consistent with earlier findings that show rime units to be better predictors of pronunciation than both head units and graphemes (Treiman et al., 1995; Vousden, 2008): Pronunciation is described most concisely by the onset–body hypothesis for both databases (Fig. 1A and B). However, the brevity with which onset–body mappings describe the data does not outweigh the length of code needed to describe the mappings themselves: The total description length is greater than that required by grapheme-sized mappings.

The finding that sublexical sized mappings offer a simpler account of the data than whole-word sized mappings can be compared to findings from empirical studies in which young children are trained to read English words using either whole-word, onset–body, head–coda, or grapheme-based reading strategies.
In a short, 6-week training study, Haskell, Foorman, and Swank (1992) compared first graders who were trained to read using whole words, onsets and bodies, or graphemes. Those in the whole-word group performed significantly worse than those in the other two groups, although there were no differences between the sublexical groups. In another short study, Levy and Lysynchuk (1997) trained beginning readers and poor grade two readers to read words using one of the three strategies above, or head–coda sized mappings. Word reading, retention, and generalization were poorest in the whole-word group, but again, there were no differences between the sublexical groups. Similarly, Levy, Bourassa, and Horn (1999) trained poor second grade readers to read words using either whole-word, onset–body, or grapheme strategies. Overall, they found that the sublexical strategies were superior to whole-word methods, and that retention was worst for the onset–body strategy. Tests of generalization revealed performance was best for graphemes and worst for whole-words.

These results give a representative picture of whether sublexical reading strategies provide a more favorable approach to reading than whole-word strategies, as borne out by more recent meta-analyses and reviews (e.g., Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001). Ehri, Nunes, Stahl, and Willows (2001) compared the effect of systematic phonics instruction with unsystematic or no phonics instruction as part of the U.S. National Reading Panel meta-analysis on reading (National Institute of Child Health and Human Development, 2000). Collating the results from 60 treatment-control comparisons, Ehri et al. (2001) compared small unit (graphemes, 39 comparisons), large unit (included onset–body units, 11 comparisons), and miscellaneous phonics programs (10 comparisons) with control conditions that did not include any systematic phonics (e.g., whole-language and whole-word approaches). All types of phonics instruction produced mean effect sizes that were significantly greater than zero, meaning that they were more effective than reading programs that did not contain systematic phonics instruction. The largest effect size was observed for the small unit programs ($d = .45$), followed by large unit programs ($d = .34$), and the smallest effect size observed was for miscellaneous programs ($d = .27$), although these effect sizes were not significantly different. These findings are consistent with the simplicity analysis presented above, in which sublexical approaches to decoding are preferred over whole-word representations, and grapheme sized units are preferred over onset–body sized units. A similar finding was observed in a more recent meta-analysis by Torgerson, Brooks, and Hall (2006), in which a small ($d = .27$) but significant effect size was found for systematic phonics over no systematic phonics instruction (e.g., whole-language and whole-word approaches). Evidence of small unit effects can also be observed in adult skill reading. Pelli and Tillman (2007) showed that letter-by-letter decoding accounted for 62% of adult reading rate, with whole-word recognition (16%) and sentence context (22%) accounting for the rest.

Other studies have found generalization from body units to be problematic (Muter, Snowling, & Taylor, 1994; Savage, 1997; Savage & Stuart, 1998), or inferior to generalization from grapheme units (Bruck & Treiman, 1992; Levy et al., 1999). This is the pattern of results predicted by the Simplicity Principle (Fig. 2).

Analysis 1 focused on how well various hypotheses accounted for spelling-to-sound translations of large reading vocabularies. The results suggest that one of the main determinants of simplicity appears to be the number of mappings required by each
hypothesis—those hypotheses that require a large number of mappings to describe the data also have longer descriptions (Table 1), although they are able to describe the data more concisely. This raises the question of whether each hypothesis can be simplified by reducing the number of mappings each employs. Closer inspection of the mappings employed by each hypothesis shows that many mappings occurred only once (19.2% of grapheme–phoneme mappings and 45.5% of onset–rime mappings) and therefore possibly contribute little to the overall statistical structure of the hypotheses. Therefore, in Analysis 2, our aim was to explore whether some mappings were more useful than others, and whether the total description length of the onset–body and grapheme hypotheses could be reduced by omitting some of the least useful mappings each employs.

3. Analysis 2: Are all mappings equally useful?

The goal of Analysis 2 was to compare the simplest forms of the onset–body and grapheme hypotheses, and to compare the utility of individual mappings. Here, each hypothesis must be expressed in its simplest form; that is, in a form that requires the shortest total description length. Given that both hypotheses contain a sizeable proportion of mappings occurring at very low frequency, the description length of the hypotheses themselves could be shortened by omitting mappings that have less impact on the description of the data, given the hypothesis. In order to find their simplest forms, a procedure must be followed to determine which mappings would result in an overall shortening of total description length when removed from the hypothesis, and of the remaining mappings, to identify their relative utility.

3.1. Method

The onset–body and grapheme hypotheses generated from the children’s database in Analysis 1 were used as initial hypotheses. The number of mappings for each hypothesis was therefore the same as given in Table 1.

For each hypothesis, an individual mapping was removed from the complete set, and the total description length (based on all mappings bar one) was recalculated. The mapping was replaced, and the procedure was repeated for each mapping in the complete set. This resulted in a set of total description lengths reflecting the removal of each mapping. Thus, the impact of removing any particular mapping could be ascertained by comparing the total description length for the hypothesis without that mapping to the total description length of the hypothesis based on the complete set of mappings. If the total description length was shorter after the removal of a mapping, then it indicated that the hypothesis might be simplified by omitting it; if it was higher, then that mapping formed part of a simpler hypothesis and should be retained. Starting with the mapping whose removal produced the shortest total description length, each mapping was removed (without replacement) from the complete set of mappings one at a time, with recalculation of the total description length occurring after each removal. If the resulting total description length was shorter, then the procedure was
repeated with the next mapping in the list. If the resulting total description length was larger, then the mapping was replaced and the procedure was repeated with alternative mappings to find the largest reduction in total description length. The procedure stopped when the total description length could no longer be reduced.

3.2. Results and discussion

The shortest total description lengths for each simplified hypothesis are presented in Fig. 3A. First, the pattern of results is qualitatively the same as for Analysis 1: The shortest description length was obtained for the grapheme hypothesis. Both hypotheses were simplified considerably by removing a substantial number of mappings (compared to Fig. 1A). The shortest total description lengths were obtained by reducing the number of onset–body mappings from 1,141 to 349, and the number of grapheme mappings from 237 to 118. Even after the hypotheses were simplified as much as possible, the total description length of the grapheme hypothesis was still considerably shorter than that of the onset–body hypothesis; it also contained approximately a third of the mappings. Even though the simplified grapheme hypothesis contained fewer mappings, the proportion of words for which the target pronunciation was the most probable was greater than that of the simplified onset–body hypothesis (73.0% vs. 60%, respectively). Furthermore, the proportion of words for which the target pronunciation was the most probable was less than from each complete hypothesis, but the reduction was much greater for the onset–body hypothesis, as depicted in Fig. 3B. The difference between the two hypotheses is in part a reflection of the relatively large proportion of onset–body mappings that occur only once (45.5% vs. 19.2% for graphemes). These mappings contribute little to the regularity of the hypothesis (while the individual mappings may be regular, because they occur only once they do not describe a pattern within the data as a whole) and so the brevity with which they describe the data does not outweigh the length of code needed to describe them. However, because there are so

![Fig. 3. Description lengths (A) of optimized hypotheses and performance on target pronunciations (B) of complete and optimized hypotheses.](image)
many of them, a correspondingly large proportion of words depend on them for correct pronunciation. The number of inconsistent orthographic units (i.e., those that map to more than one phonemic representation) represented by the simplified hypotheses was reduced from 15.2% to 0.6% for the onset–body hypothesis, and from 37.2% to 25.8% for the grapheme hypothesis.

The results yielded an ordered list of mappings for each hypothesis (see Appendix S1 for the ordered list of grapheme-sized mappings), with those ranked high on the list representing the most useful mappings. These mappings clearly account for more of the regularity in the data than those nearer the end of the list. The highest ranked mappings in the grapheme hypothesis are mainly single letter graphemes, with increasingly longer graphemes occupying lower ranks (median rank for single-letter graphemes = 20, two-letter graphemes = 63.5, three-letter graphemes = 71, four-letter graphemes = 95).

These findings are potentially relevant for reading instruction because they make explicit the potential benefits of learning different representational units, both at a general type-of-unit level and more specifically at an individual unit level. However, relatively few studies have directly compared the impact of teaching grapheme versus onset–body sized correspondences. According to the results above, the preferred representation is clearly for grapheme-sized correspondences. Knowledge of grapheme-sized correspondences should therefore be advantageous to beginning readers. A small but insignificant advantage was found for small units in a recent meta-analysis (Torgerson et al., 2006); however, only three studies were included in the analysis and the authors concluded that there was insufficient evidence on which to draw any firm conclusions. Several intervention studies not included in the Torgerson et al. (2006) meta-analysis have directly compared the impact of teaching grapheme versus onset–body sized correspondences. Christensen and Bowey (2005) compared children taught over a 14-week period where instruction was based on graphemes, bodies, or a control condition. Children were assessed on reading accuracy and speed for words taught during the program and transfer words. The grapheme group were consistently better on all assessments, significantly so for the transfer words. In addition, the grapheme group’s reading age on posttest was 9 months ahead of those in the body group. Similar results were found by Savage, Abrami, Hipps, and Deault (2009), who compared children taught over a 12-week period using either grapheme or body-based interventions, or a control condition. They found that both interventions produced significant improvements over the control condition for a range of literacy measures, but the grapheme intervention produced larger effect sizes for the key skills such as word blending and reading accuracy, both at immediate and delayed posttest, whereas the body intervention produced a more general effect.

Several studies have explored the effects of orthographic complexity on grapheme acquisition. These data can be compared to the list of grapheme correspondences listed in Appendix S1; the more useful correspondences (those higher up the list) should be easier to learn. In a study comparing poor and normal readers, Manis (1981) (as cited in Morrison, 1984) found that both groups read simple word-initial consonants in nonwords with greater accuracy than medial short vowels, which in turn were read more accurately than medial long vowels. The correlation between the accuracy that normal readers read each of the 15
graphemes and the rank order of those graphemes in Appendix S1 was high, at .7. Laxon, Gallagher, and Masterson (2002) found that 6- to 7-year-olds read simple, short vowels with greater accuracy than complex (digraph) vowels in both familiar words and nonwords. Similarly, the mean rank order of simple, short vowels from Appendix S1 was less than the mean rank order of the complex vowels (18 vs. 49) from that study. Thus, the order in which different types of graphemes are acquired appears to be well aligned with their utility in decoding text, with more useful graphemes acquired before less useful ones.

Analyses 1 and 2 focused on how well various hypotheses accounted for spelling-to-sound translations of large reading vocabularies. In Analysis 3, we explored whether similar results would be obtained if vocabularies of increasing size, drawn from the children’s database, were considered.

4. Analysis 3: Does optimal unit size change with reading development?

The aim of this analysis was to compare the optimal hypotheses obtained from Analysis 2 (onset–body and grapheme) with the whole-word hypothesis, for different sized reading vocabularies. The above results suggest that grapheme-sized mappings result in the simplest hypothesis about the data when the data encompasses a sizeable vocabulary. However, a reading vocabulary is acquired incrementally, and the same (grapheme) hypothesis may not result in the simplest hypothesis for smaller data sets.

4.1. Method

The children’s database was split into incrementally larger vocabulary sets. The different sized vocabulary sets from the children’s database were formed by saving the first 10, 50, 100, 500, 1,000, and 3,000 most frequent words to a different vocabulary set, and they represented an approximate progression of the number of words a child can read. Thus, each subsequent set included all words in the next smallest set plus a number of less frequent words from the database.

The simplicity of the whole-word, onset–body, and grapheme hypothesis for each vocabulary size was calculated by following the calculations described in the Appendix, but restricting the mapping set for the onset–body and grapheme hypotheses to those from the optimal hypotheses resulting from Analysis 2, rather than using the complete set (as in Analysis 1).

4.2. Results and discussion

The total description length for each hypothesis is plotted in Fig. 4 as a function of vocabulary size. For small vocabularies (up to 50 words), the simplest hypothesis about the data is represented by whole-word mappings. However, once the vocabulary size exceeds around 50 words, grapheme-sized mappings provide the simplest hypothesis about the data.
Fig. 4 indicates that the regularity evident in all but the smallest vocabularies is best captured by grapheme-sized mappings.

It is clearly suggested from the literature (e.g., Carroll et al., 2003) that children are aware of larger units before they are aware of smaller units, and furthermore, that awareness of phonemes develops (at least in part) alongside instruction. Thus, children come to the task of learning to read with an established large unit phonological lexicon, which may explain why learning a small vocabulary by sight is quick and easy—the appropriate phonological representations to which orthographic representations must be associated are already present. However, our analyses explain why this initially useful strategy, evident in some models of reading development (e.g., Ehri, 1992), is less efficient as the reading vocabulary develops—as has been demonstrated by a number of meta-analyses comparing “whole-word” approaches to instruction with systematic small unit approaches—because grapheme-sized representations offer a much more efficient representation of the data for larger vocabularies. We return to this issue in the General Discussion. In contrast, not only do children have to develop their knowledge of graphemes and phonemes, but for small vocabularies, grapheme-sized mappings are less likely to be repeated (most grapheme-sized mappings in the smaller vocabularies occur only once); therefore, the number of grapheme-sized mappings needed to describe the vocabulary will be greater than the number of whole-word mappings. Hence, the simplest hypothesis for small vocabularies is the whole-word hypothesis.

In a recent study, Powell, Plaut, and Funnell (2006) investigated children’s ability to read words and nonwords (a measure of the ability to apply grapheme–phoneme knowledge) at the beginning of formal instruction in the United Kingdom, and again 6 months later. They found that the children read significantly more words (where errors were largely lexical errors) than nonwords at Time 1, but read a similar number of each at Time 2 (although lexical errors still dominated). Thus, their initial ability, when few words were known, was marked more by a whole-word approach, whereas later ability appeared to be characterized by a more equal input from both whole-word and grapheme–phoneme knowledge. Likewise, in an early training study, Vellutino and Scanlon (1986) found that children trained by
whole-word methods read more pretrained words than children trained with small-unit phonetic methods on early trials, yet the pattern reversed for later trials. The pattern predicted by the simplicity analysis above matches these patterns of acquisition. It explains why when children are taught a small set of sight words in a particular session, the whole-word approach will work best but will fail to integrate with the rest of their knowledge, therefore leading to poorer overall performance in the longer term (Vellutino & Scanlon, 1986).

However, English contains many words that are irregular (the pronunciations for some of the letters they contain are not the most frequent pronunciation), and it is possible that hypothesis preference is determined by whether words are regular, rather than vocabulary size. In order to explore this further, the children’s database was split according to whether a word was regular. Different sized vocabularies were then formed as before, containing either all regular, or all irregular, words. The total description lengths for each vocabulary size were recalculated by following the calculations described in the Appendix. For a meaningful comparison between the regular and irregular vocabularies, description lengths were calculated using the complete sets of (nonoptimized) mappings, as in Analysis 1 and not the optimized set of mappings derived in Analysis 2 (although a qualitatively similar result was obtained using the optimized grapheme hypothesis from Analysis 2, even for small vocabularies). If regularity, rather than vocabulary size, is the critical factor, then the grapheme hypothesis for the smallest regular vocabularies should have shorter description lengths than the whole-word hypothesis.

Fig. 5B and C shows the total description lengths for each hypothesis as a function of vocabulary size, for regular and irregular words, respectively; Fig. 5A shows the corresponding description lengths for all words (regular and irregular), for comparison. The advantage for whole-word mappings still exists for small vocabularies, but it is moderated by regularity. The advantage for grapheme-sized mappings is apparent for small regular vocabularies (Fig. 5B) but only for relatively larger irregular vocabularies (Fig. 5A and C). As well as regularity moderating the point at which the grapheme hypothesis becomes most preferable, the number and complexity of graphemes in an orthography is likely to have an effect. Thus, it is possible for other more consistent languages where the orthography contains fewer and less complex graphemes that the advantage for whole-word learning might disappear even for small vocabularies. For example, for German, small-unit phonics teaching typically dominates from the beginning of reading instruction (e.g., Landerl, 2000), with considerable success (Wimmer, 1993).

Fig. 5B and C indicates that the total description length for irregular words is much larger than for regular words, suggesting regular words should be learned more easily than irregular words. This appears to be the case both across and within languages. Thus, the reading (decoding) ability of British children lags consistently behind that of children who learn more regular languages (Seymour et al., 2003; Ziegler & Goswami, 2005), whereas children who learn less regular languages, for example, Chinese, spend considerably longer learning an equivalently sized vocabulary (Rayner et al., 2001). Hanley, Masterson, Spencer, and Evans (2004) followed up the Welsh and English readers of a previous study to find that at age 10, the English group still read irregular words worse than regular and nonwords, which were read to a similar proficiency as the Welsh readers. These data are consistent
with the idea that the simplicity of an orthography affects the ease with which it can be learned.

Analyses 1–3 compared hypotheses composed of one type of mapping with those composed of another. The results consistently show that the grapheme hypothesis yields the shortest total description length. However, the optimal grapheme hypothesis, as
identified in Analysis 3, also contained some graphemes with multiple pronunciations. Our aim for Analysis 4 was to consider whether reducing the inconsistency in the grapheme hypothesis by providing some larger unit mappings would further reduce the description length.

5. Analysis 4: Can the addition of large units reduce small unit inconsistency?

The results from Analysis 2 indicated that the benefit in describing the data when representing some alternative pronunciations in the optimal grapheme hypothesis outweighed the cost of describing them. However, it is possible that the benefit gained from those alternative pronunciations could be increased further by providing a representation of when they could be most usefully applied. For example, the grapheme “a” may be pronounced in many different ways (e.g., ball, fast, want), but the correct pronunciation can often be determined by considering the surrounding letter(s); “a” is nearly always pronounced /aː/ when it is followed by the letter “s” (e.g., fast, task), and often pronounced /d/ when it is preceded by the letter “w” (e.g., was, want). Likewise, the grapheme “c” is nearly always pronounced /s/ when it is followed by grapheme “e” (e.g., cell, cent), or appears in the coda position (e.g., face, rice, fierce). In Analysis 4, our goal was to optimize the grapheme hypothesis further by constructing an alternative hypothesis that contained additional mappings of other unit sizes, relevant to the inconsistent grapheme mappings. The main motivation for this analysis was to explore whether providing some contextual information for inconsistent graphemes would reduce the overall description length. In addition, we wished to consider whether taking account of the syllabic position of graphemes when describing mappings would have an impact on total description length, prompted by the finding that the distribution of consonants across syllable position is not uniform (Kessler & Treiman, 1997).

5.1. Method

Analysis 4 was restricted to the children’s database. The simplest grapheme hypothesis, as identified by Analysis 2, containing 118 grapheme–phoneme mappings, was used as the starting hypothesis to which larger unit mappings were added. This straightforward approach was taken to make the contribution of large units transparent, rather than to explore a less transparent analysis of the order in which larger and smaller units added alongside each other might have most impact.

5.1.1. Position-specific mappings

First, inconsistent consonant graphemes (e.g., c, th) that could be pronounced in more than one way in both onset and coda positions were identified from the simplest grapheme hypothesis. The mappings containing them were listed with two frequencies, reflecting how often each mapping occurred separately in the onset and coda of a syllable. For example, the mappings containing the grapheme th were listed as “th /θ/ 34, 32” and “th /ð/
15, 12’’ instead of ‘‘th /θ/ 66’’ and ‘‘th /ð/ 27.’’ The total description length of the simplest grapheme hypothesis was then calculated separately by including each inconsistent consonant grapheme with position-specific mapping frequencies. The impact of position-specific mappings was assessed by comparing the total description length of the simplest grapheme hypothesis with and without position-specific mapping frequencies for each inconsistent consonant grapheme. Those that resulted in a shorter description length were incorporated into the simplest grapheme hypothesis before large-unit mappings were tested, below.

5.1.2. Large unit mappings

A potential pool of large-unit mappings was created by identifying any large-unit mapping (onset, rime, head, and coda mappings from hypotheses described in Analysis 1) that contained an alternative pronunciation for an inconsistent grapheme. For example, the rime mapping ead—/ed/ was identified as containing an alternative pronunciation for the inconsistent grapheme ea (usually pronounced to rhyme with b ea ch), and war—/wɔː/ was identified as containing an alternative pronunciation for the inconsistent grapheme ar (usually pronounced to rhyme with h ard). The total description length was calculated for the simplest grapheme hypothesis plus one of the large unit mappings, for each large unit mapping in turn. The impact of adding any particular mapping was determined by comparing the total description length for the simplest grapheme hypothesis with and without the mapping in question. If the total description length was shorter after the addition of a mapping, then it indicated that the hypothesis might be simplified by including that mapping; otherwise it should be omitted. Starting with the mapping whose addition produced the shortest total description length, each mapping was added to the simplest grapheme hypothesis one at a time, with recalculation of the total description length occurring after each addition. If the resulting total description length was shorter, then the procedure was repeated with the next mapping in the list. If the resulting total description length was larger, or only a little shorter (<0.1%), then the mapping was replaced and the procedure was replaced with alternative mappings to find the largest reduction in total description length. The procedure stopped when the total description length could no longer be reduced.

5.2. Results and discussion

5.2.1. Position-specific mappings

Five inconsistent consonant graphemes were identified that mapped to more than one phoneme in both onset and coda positions: c, ch, g, s, and th. Each of these mappings, in turn, were listed with a separate frequency according to how often they occurred in the onset and coda positions, and the total description length was recalculated. Only the inclusion of a position-specific mapping for the grapheme c yielded a shorter total description length and was therefore included in the simplest grapheme hypothesis. This procedure reflected the fact that the grapheme c is more likely to be pronounced /s/ when it appears in the coda than when it appears in the onset of a syllable.
5.2.2. Large-unit mappings

There were many large-unit mappings that contained an alternative pronunciation for an inconsistent grapheme—most occurred with very low frequency. In contrast, the majority of the highest frequency large-unit mappings were regular (as defined by the most frequent grapheme–phoneme mapping). The pool of potential mappings was limited to the 20 most frequent mappings of three types (rime, head, and coda; there were no suitable onset mappings). These mappings are listed in Table 2, in descending order of frequency, with those that either reduce or did not increase the total description length when included on their own, shown in bold. Those mappings that reduced the total description length and were added to the simplest grapheme hypothesis are listed in Table 3, with those that reduced the total description length most listed first, along with the total description length after their addition, and the proportion of words where the target pronunciation was the most probable. As each large unit mapping (as listed in Table 3) was added in turn to the simplest grapheme hypothesis, the total description length fell, and the proportion of words where the target pronunciation was the most probable increased.

Most of the large-unit mappings in Table 3 provide consonantal context for the vowel grapheme (wa, all, wor, ought, ook, as). There were several instances where the description length was reduced after providing context for consonant graphemes, mainly reflecting a

Table 2
Large-unit mappings identified as potential mappings to reduce total description length of simplest grapheme hypothesis

<table>
<thead>
<tr>
<th>Heads</th>
<th>Rimes</th>
<th>Codas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphemes</td>
<td>Phonemes</td>
<td>Example</td>
</tr>
<tr>
<td>wor</td>
<td>wɔː</td>
<td>work</td>
</tr>
<tr>
<td>wa</td>
<td>wə</td>
<td>want</td>
</tr>
<tr>
<td>pu</td>
<td>pʊ</td>
<td>put</td>
</tr>
<tr>
<td>pa</td>
<td>pə</td>
<td>pass</td>
</tr>
<tr>
<td>war</td>
<td>wər</td>
<td>warm</td>
</tr>
<tr>
<td>bow</td>
<td>bɔʊ</td>
<td>bowl</td>
</tr>
<tr>
<td>ca</td>
<td>ɡa</td>
<td>cast</td>
</tr>
<tr>
<td>wa</td>
<td>wə</td>
<td>wall</td>
</tr>
<tr>
<td>brow</td>
<td>braʊ</td>
<td>brown</td>
</tr>
<tr>
<td>crow</td>
<td>kraʊ</td>
<td>crown</td>
</tr>
<tr>
<td>hoo</td>
<td>hʊ</td>
<td>hood</td>
</tr>
<tr>
<td>lea</td>
<td>ɡe</td>
<td>leapt</td>
</tr>
<tr>
<td>ma</td>
<td>maː</td>
<td>mast</td>
</tr>
<tr>
<td>ta</td>
<td>tə</td>
<td>tall</td>
</tr>
<tr>
<td>woo</td>
<td>wʊ</td>
<td>wood</td>
</tr>
<tr>
<td>ba</td>
<td>baː</td>
<td>ball</td>
</tr>
<tr>
<td>blow</td>
<td>bləʊ</td>
<td>blown</td>
</tr>
<tr>
<td>brea</td>
<td>bɾe</td>
<td>bread</td>
</tr>
<tr>
<td>bu</td>
<td>buː</td>
<td>bush</td>
</tr>
<tr>
<td>ca</td>
<td>kɔ</td>
<td>call</td>
</tr>
</tbody>
</table>
much simpler rule, namely the rule for pronouncing plural “s” (ls, ns). Where the context determined vowel pronunciation, this occurred more often with codas in the form of a body unit. Thus, in terms of reducing the total description length, body units appeared to offer a more useful guide to inconsistent vowel pronunciation than head units. In fact, the only head units that reduced the total description length were those where “w” reliably distinguished between alternative pronunciations of graphemes “a” and “or.” The coda unit “dge” also reduced the total description length by influencing the pronunciation of the previous vowel: By including the final “e” with the “dg” grapheme, the preceding vowel is no longer lengthened by the final “e,” as it would be under the simplest grapheme hypothesis. For example, the correct pronunciation of “fridge” now corresponds to the briefest description according to the new hypothesis (f—ʃ, r—ɹ, i—ɪ, dge—dʒ), whereas previously the briefest description would have corresponded to an incorrect pronunciation (f—ʃ, r—ɹ, i*—aɪ, dg—dʒ). This pattern is true for all words in the children’s database that end in “dge.”

The findings from this last analysis show that larger, body-based units have a useful role to play in decoding English text, but that the simplest hypothesis about the data is dominated by grapheme-sized units. These results are consistent with empirical findings in which humans appear to base the pronunciation of the majority of nonwords, including those based on inconsistent word bodies (e.g., bive), on small unit correspondences rather than larger units (i.e., they choose a pronunciation that rhymes with “hive” rather than “give”) (Andrews & Scarratt, 1998). In the Andrews and Scarratt study, body-rime pronunciations were produced, but primarily only to nonwords derived from bodies that are never pronounced regularly in words (e.g., dask), especially when the nonword has more than one irregular neighbor. In fact, the best predictor of whether a word was assigned a regular pronunciation was the proportion of words in which the body was regularly pronounced. These data show that body-rime correspondences are primarily used when they provide a more probable cue to pronunciation than small units. We note that, in line with this empiri-

<table>
<thead>
<tr>
<th>Hypothesis Plus Additional Mappings</th>
<th>Total Description Length</th>
<th>Percentage of Target Pronunciations Output as Most Probable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simplest grapheme hypothesis</td>
<td>8,433</td>
<td>73.0</td>
</tr>
<tr>
<td>c (coda only) s</td>
<td>8,355</td>
<td>73.4</td>
</tr>
<tr>
<td>dge dʒ</td>
<td>8,338</td>
<td>73.3</td>
</tr>
<tr>
<td>wa wð</td>
<td>8,284</td>
<td>73.7</td>
</tr>
<tr>
<td>ce s</td>
<td>8,190</td>
<td>74.0</td>
</tr>
<tr>
<td>all zːl</td>
<td>8,146</td>
<td>74.4</td>
</tr>
<tr>
<td>ns nz</td>
<td>8,143</td>
<td>74.9</td>
</tr>
<tr>
<td>wor wɔː</td>
<td>8,127</td>
<td>75.2</td>
</tr>
<tr>
<td>ought zːt</td>
<td>8,132</td>
<td>75.4</td>
</tr>
<tr>
<td>ook ðk</td>
<td>8,125</td>
<td>76.0</td>
</tr>
<tr>
<td>as aːs</td>
<td>8,111</td>
<td>76.5</td>
</tr>
<tr>
<td>ls lz</td>
<td>8,106</td>
<td>76.9</td>
</tr>
</tbody>
</table>
6. Discussion

We have shown, with a series of analyses that focus on the relationship between print and sound in English, how the optimal choice of representational units in reading is driven by the change that comes about as the reading vocabulary increases in size. This development is explained by the application of a general simplicity principle, in which the simplest encoding of the data is sought. In the discussion that follows, we consider how such analyses can provide an important insight into the representational units that would be most useful for reading, and we consider the implications for selecting the most useful skills to teach beginning readers. We also relate our findings to theories of reading acquisition and to models of skilled reading. Finally, in light of the fact that the Simplicity Principle is a general principle and is not limited to decoding text, we discuss how it may be usefully generalized to other linguistic domains. We begin with a summary of the findings.

6.1. Summary

The initial analysis found that overall, the shortest total description length was obtained for the grapheme hypothesis, with the head–coda and onset–body hypotheses the next longest, and the whole-word hypothesis longest of all. These results held for both children’s and adults’ text. Two tests of generalization suggested the grapheme hypothesis was not only the most preferable hypothesis, but also that it appeared able to generalize to new data more so than the other hypotheses. Analysis 2 showed that it was possible to further reduce the total description length of the onset–body and grapheme hypothesis to arrive at a more optimal hypothesis in each case. In so doing, the grapheme hypothesis still retained the shortest total description length, with very little reduction in the ability to describe the data compared to the onset–body hypothesis. Analysis 2 also revealed that the optimal onset–body hypothesis contained virtually no inconsistent mappings, yet 25% of mappings in the...
optimal grapheme hypothesis were inconsistent. Thus, at the grapheme level, alternative pronunciations appear to play an important role in decoding text, but much less so for the onset–body level. Finally, Analysis 2 yielded an ordered list of mappings, showing that simpler mappings (i.e., those that map single letter graphemes) play a more important role in the grapheme hypothesis than complex mappings (i.e., those that map three or four letter graphemes). Analysis 3 revealed different results when the size of the vocabulary was manipulated: The hypothesis with the least structure, the whole-word hypothesis, provided the shortest total description length for very small vocabularies; otherwise the grapheme hypothesis was preferred. Analysis 4 showed that the total description length and the ambiguity of some grapheme–phoneme mappings could be reduced by including contextual information for some mappings, either with position-specific constraints, or by introducing large-unit mappings. In summary, the results showed that overall the simplest encoding of the data resulted from the grapheme hypothesis. However, large units (in the form of whole-word mappings) were important early on when the reading vocabulary consisted of a small number of words, and again, with a larger vocabulary when head and body mappings were able to guide the pronunciation of inconsistent graphemes.

6.2. Implications for the choice of representational units in reading

As we have seen, and will discuss further below, much empirical research, from intervention to behavioral studies with both children and adults, is consistent with the pattern of results, supporting the assertion that simpler explanations are preferable.

The current findings offer new insights into implications of the claim that the representational units acquired in reading may be structured according to the demands of the vocabulary to be acquired, and as such seem to provide some support for the Psycholinguistic Grain Size (PGS) account of reading (Ziegler & Goswami, 2005) when viewed from a simplicity perspective. An important assumption behind PGS is that the consistency of the language drives the adaptation of the reading process and explains why readers progress more quickly for languages where the orthography is consistent. Thus, because English is inconsistent at the grapheme level, the reading system must pay attention to larger sublexical units to resolve those inconsistencies. However, the reading process must also resolve the granularity problem—that is, the larger the unit, the greater the number of units there are in learning to read.

The present results show the effects of the granularity problem for English text. Results from Analysis 1 show that although the large units describe the data better than small units, because there are so many of them, the cost of describing the mappings is never outweighed by the brevity with which they describe the data. Consistent with a body of other research, the results showed greater inconsistency at the grapheme level than any other level. However, some alternative pronunciations for inconsistent graphemes were found to “earn their keep” in the search for the simplest grapheme hypothesis in Analysis 2. Analysis 4 provided support for PGS in that the results demonstrated an advantage for the simplest grapheme hypothesis when it was augmented with larger units. The role played by these units seemed to be to resolve ambiguity in pronouncing some inconsistent (vowel) graphemes. The
optimal (simplest) solution to inconsistency appears therefore to involve recruiting other larger units, as postulated by PGS.

We note evidence supporting PGS exists at a within-language level. The choice of the optimal strategy depends on the vocabulary to be acquired, both in terms of the particular words within a vocabulary and the size of that vocabulary. Analysis 3 showed that for small vocabularies there is not enough evidence of sublexical structure to warrant the formation of such hypotheses and so those hypotheses are not preferred. Rather, the simplest hypothesis in this case is a whole-word one. However, this result changes as the vocabulary size grows, because the effects of the granularity problem quickly become apparent for a whole-word hypothesis, and more evidence of the sublexical structure also becomes available. The structure is then best described by the relationship that exists between graphemes and phonemes. However, the structure of the mapping between print and sound continues to change as evidence becomes available to reveal reliable large-unit cues to the pronunciation of inconsistent graphemes. There is a further shift, therefore, from a grapheme only hypothesis, to one that includes a small number of larger-units. Thus, the unfolding structure of the vocabulary drives a shift in choice from whole words, to graphemes, and back to include large-unit mappings again. The choice of optimal strategy therefore depends in part on the evidence for different units within different subsets of the same language. Another language with less inconsistency may well favor an alternative hypothesis composed entirely of small-unit mappings.

The idea that vocabulary growth drives a preference for finer-grained representations has also been suggested in the spoken domain (Metsala & Walley, 1998). In Metsala and Walley’s Lexical Restructuring Model, fine-grained phonemic representations of spoken words are thought to develop in order to distinguish between words in increasingly crowded phonological neighbourhoods. For example, recognizing the word “big” will be hard if the child already knows bag, bug, bib, bit, dig, etc., unless finer-grained (phonemic) representations are established. Our findings suggest that the growth of a reading vocabulary would encourage the use of finer-grained representations for reasons of efficiency and economy in cognitive resources rather than to overcome immediate performance constraints. However, both accounts assume that the cognitive system adapts to the emerging structure of the data with which it is faced in response to the need to accurately (and succinctly) capture that data.

6.3. Implications for instruction

It is generally accepted that English children need phonological awareness and an understanding of the mapping from orthography to phonology before they can master reading. However, although much effort has gone into determining how best to deliver such instruction (e.g., Wyse & Goswami, 2008), less research has focused on identifying those mappings that would be most useful in the decoding process. In the present paper, we aimed to identify the mappings that, according to the Simplicity Principle, should be preferred when faced with the task of decoding English text. In this section, we consider the potential relevance of the findings to reading instruction.
6.3.1. Sight vocabulary

Beginning sight vocabulary are words taught by rote/recognition without any analysis of letter-sound correspondences. This often occurs very early in instruction, either before or alongside more formal decoding methods. The current results posit a role for learning sight vocabulary only under certain conditions, namely until there is sufficient structure in the data to warrant an alternative, more complex hypothesis. For example, Analysis 3 showed that learning words by sight was more preferable when the vocabulary was small, approximately 50 words or less. As the vocabulary increased in size beyond 50 words, a hypothesis that was able to capture the evident structure embodied by the reliable relationship between letters and sounds was preferred, and the whole-word hypothesis became akin to the task of learning a telephone directory (Share, 1995). The present results complement other recent research that shows sight vocabulary is best limited to around 100 words for other reasons, namely that the 100 most frequent words account for about 50% of all word tokens (J. Solity, E. McNab, & J. I. Vousden, unpublished data; Solity & Vousden, 2009; Stuart et al., 2003; Vousden, 2008). Many of the most frequent words contain inconsistent mappings (e.g., was, said, of, etc.): therefore, these very salient words are less likely to be rapidly acquired through the use of phonics, but acquired with relative ease through sight vocabulary (Shapiro & Solity, 2008; Solity & Shapiro, 2008). The results from Analysis 3 (Fig. 5A–C) show the extent to which irregular words are less well characterized (and presumably harder to learn) by grapheme-sized units. These words would be best taught as sight vocabulary in order that beginning readers could decode the texts in which they appeared.

However, this evidence does not suggest that instruction should follow discrete stages of whole-word learning followed by phonics instruction. Recent behavioral data have shown that arbitrary mappings between print and sound (i.e., a pure whole-word type of approach) are much harder to learn than mappings that contain some relationship (Bowman & Treiman, 2008; Ellefson et al., 2009). There is evidence to suggest that learning a sight vocabulary is in fact aided by some letter-sound knowledge (Rack, Hulme, Snowling, & Wightman, 1994; Stuart, Masterson, & Dixon, 2000). Thus, although the current results from Analysis 3 may imply a discrete change in preference from sight vocabulary to phonics, in light of the fact that a much larger vocabulary is eventually acquired, and behavioral evidence showing an early benefit for phonic input, it is much more likely that the different strategy preferences should overlap.

6.3.2. Phonics

The main results from all the above analyses suggest a central role is played by the knowledge of grapheme–phoneme mappings when decoding English text. These results reinforce the large body of research that shows that phonological awareness and knowledge of spelling-to-sound knowledge is fundamental in determining reading outcome (Hulme et al., 1998; Wyse & Goswami, 2008). Results from Analysis 2 show that the optimal number of grapheme–phoneme mappings is much less than the total number within English. In general, grapheme–phoneme mappings generalize better to new text than do other sized units. The present results show that within an optimal set, some mappings are more useful than others. In terms of deciding what to teach, the analyses presented here provide a theoretical rationale for why
some mappings should be introduced early, and others later. In general, single letter mappings are more useful than larger, more complex mappings; more specifically, mappings near the top of Appendix S1 are more useful and should be introduced before those further down the table. The results do not imply that all the grapheme–phoneme mappings identified in the simplest grapheme hypothesis should be taught explicitly. Indeed, it is possible that at least some children are able to use some grapheme–phoneme knowledge to infer further mappings for themselves (Share, 1995; Stuart, Masterson, Dixon, & Quinlan, 1999). Rather, the results provide a guide to the order that mappings could be taught initially. In the behavioral study reported earlier by Shapiro and Solity (2008), children who were taught 100 high-frequency words at a sight level alongside 64 GPCs and only learned the most frequently occurring phoneme for any given grapheme had better reading outcomes than those in comparison schools. Thus, knowledge of a considerably reduced number of GPCs may be sufficient to bootstrap reading acquisition, at least for some children.

It is an open question as to how much explicit instruction is required before self-teaching can occur, or when semantic input can fill the gaps (Ricketts, Nation, & Bishop, 2007). Analysis 4 clearly shows a role for larger units, and that phonic-based instruction need not be restricted solely to grapheme-sized units—when chosen appropriately, larger units offer valuable guidance on the pronunciation of inconsistent graphemes. However, at the point at which it may be useful to teach larger units, children are likely to be decoding text with polysyllabic as well as monosyllabic words. In this case, it may be preferable to teach alternative mappings appropriate to polysyllabic text. These mappings might take the form of frequent word endings (e.g., -tion), or prefixes and suffixes. Whether these mappings would add to, or replace, heads, rimes, and codas, as part of an optimal representation, remains an empirical question. Again, the change in the vocabulary to be acquired may lead to shifts of optimal strategy.

6.3.3. Summary of implications for instruction

Taken together, the analyses presented here offer a theoretically based explanation for why teaching mainly grapheme-sized letter-sound correspondences should be more beneficial than teaching other sized correspondences, consistent with current empirical evidence. The analyses go further and offer guidance on which grapheme-sized correspondences should confer greatest benefit, and in this sense make a unique contribution to the literature. The implications for reading instruction are several. Learning a sight vocabulary (whole-word approach) is best limited to a small number of words (50–100), and it may be more beneficial for irregular words. However, learning a carefully chosen set of grapheme–phoneme correspondences is also clearly beneficial, from fairly small vocabularies upwards. Therefore, it would seem wise to introduce grapheme–phoneme correspondences from the outset of reading instruction, starting with the most useful correspondences first. Learning some large-unit correspondences confers potential benefits, but only when restricted to those units where the large majority of words containing the unit share its (irregular) large unit pronunciation.

There is a wealth of evidence consistent with the general implication that learning grapheme–phoneme correspondences is a highly effective method of learning to read.
Furthermore, several studies have shown improved performance for programs that incorporate a small sight vocabulary alongside small-unit phonics instruction (Shapiro & Solity, 2008; Solity & Shapiro, 2008; Vellutino & Scanlon, 1986). However, there has been a dearth of research on which grapheme–phoneme correspondences confer the greatest benefits for learning to read. In this respect, we note there is considerable variation in the number of correspondences that are taught across popular reading programs. For example, Jolly Phonics (Lloyd, 1992) teaches 64–70, whereas THRASS (Davies & Ritchie, 2003) teaches over 100, and little in the way of theoretical motivation for the selection of those correspondences. A comparison of the potential and real impact of teaching the (simplest) grapheme–phoneme correspondences presented here with other popular programs would indeed be instructive, and it is the subject of ongoing research (Vousden, Chater, Ellefson, & Solity, 2010). We expect that the number and exact choice of correspondences taught will affect outcomes.

6.4. Relation to theories of reading acquisition

One aim of the current article was to apply a general cognitive principle—the Simplicity Principle—to explore the representations that best serve reading acquisition, rather than to test a particular theory of reading acquisition, or model of skilled reading. However, the results of the analyses do bear on current theories of reading acquisition regarding the units that support the links from orthography to phonology.

According to the large-units first view, children initially decode words according to large (body) units. Evidence consistent with this view comes from training studies in which beginning readers are able to decode words primed or segmented by body units with greater ease than when other units are invoked (Bruck & Treiman, 1992; Goswami, 1986). In contrast, the small-units first view assumes early decoding is based on grapheme-sized units, based on evidence from studies that observe an increase in the ability to detect and use body-units with development (Duncan, Seymour, & Hill, 2000). Finally, the flexible unit size model assumes both unit types are potentially available early in development (Stuart et al., 1999) and are applied flexibly according to task demands (G. D. A. Brown & Deavers, 1999). Our results sit well with the flexible unit size model because the simplest hypothesis concerning representations encompassed both small and large unit sizes, although they suggest that, where task demands are set by the requirement to read a large vocabulary, large units will be of less use than small units.

Our results offer insight into models in which early reading development is dominated by something akin to whole-word associations, followed by the increasing development of the alphabetic principle in which letter(s) become associated with phonemes (Ehri, 1992; Perfetti, 1992). Early in development, when vocabulary size is small, few systematic relations between letters and words exist, yet it is relatively easy to associate existing whole-word phonological representations with written forms. However, these associations are arbitrary, do not integrate with the statistical structure embodied in larger vocabularies, and are hard to learn when many items must be distinguished (i.e., for larger vocabularies). In contrast, once the vocabulary becomes large enough to exhibit systematic relations between letter(s) and sounds, this structure offers the most efficient method of representing...
the link between orthography and phonology, and hence provides a driving force for the development of small unit (phoneme) awareness. Thus, the Simplicity Principle offers an independently motivated potential explanation for such data.

6.5. Relation to models of skilled reading

The results from the current analyses can also be compared to current instantiations of models of skilled word recognition, in particular to the way in which the links between orthography and phonology are represented. Typically, such models are less concerned with the relative complexity of the (sublexical) representations within the model compared to the fit to the data and, therefore, often capture the relations between orthography and phonology at multiple unit sizes in an unconstrained manner (Perry et al., 2007; Plaut, McClelland, Seidenberg, & Patterson, 1996; Zorzi et al., 1998). The representations they embody do not align simply with the analyses presented here. However, the current results imply that models in which there is a direct association between predominantly graphemic and phonemic units (e.g., Perry et al., 2007) will exhibit better generalization than those where representations are more aligned to body units (e.g., Zorzi et al., 1998). We note that some recent models of reading find that structuring the input to represent graphemes and/or pretraining the models on select grapheme–phoneme mappings improves the model fit to the data (Hutzler et al., 2004; Perry et al., 2007; Powell et al., 2006). This effectively strengthens the contribution from grapheme–phoneme mappings relative to that from other unit sizes, consistent with the results of the simplicity analyses presented here. Furthermore, models in which this association is represented statistically (in the sense that the same graphemes are linked differentially to multiple phonemes depending on the strength of their association) and flexibly (in that some large unit mappings are also accommodated where appropriate) rather than as a symbolic rule system like the DRC model (Coltheart et al., 1993), will capture the data more efficiently. Indeed, one of the many disputed points between proponents of statistically based associative models (Plaut et al., 1996), and supporters of symbolic rule-based models (Coltheart et al., 1993) concerns the merits of a model that instantiates a simple principle that accounts for a large amount of data versus a more complex model that provides a closer match to behavioral data (Woollams, Lambon Ralph, Plaut, & Patterson, 2010). The results presented here fit best with models such as the CDP+ model (Perry et al., 2007), in which a statistical relationship, predominantly between graphemes and phonemes, exists.

6.6. Relation to other aspects of language acquisition

The approach reported here, of employing the Simplicity Principle to choose between different forms of spelling-to-sound mappings, is not limited to the specific case of reading acquisition. Many other linguistic domains are similarly characterized by a quasi-regular stable structure, where irregularities and exceptions coexist within a largely regular system. The scope for generalizing the current approach to spelling is obvious, but the Simplicity Principle may equally be generalized to other domains, such as the acquisition of English past tense forms. Indeed, the Simplicity Principle has been applied widely in various other
linguistic domains, such as phonology (Ellison, 1992; Goldsmith, 2002; Goldsmith & Riggle, 2010), speech segmentation (Brent & Cartwright, 1996), morphology (Goldsmith, 2001), learning syntax in the absence of negative evidence (Dowman, 2000; Onnis et al., 2002; Perfors et al., 2006), and to the broader topic of language evolution (Brighton & Kirby, 2001; Roberts, Onnis, & Chater, 2005).

For example, in a related analysis, using a Bayesian model selection method that is formally very closely related to the simplicity principle, Perfors et al. (2006) demonstrated why hierarchical grammars fit a corpus of child-directed speech better than nonhierarchical grammars. Perfors et al. compared linear (structure independent) and hierarchical (structure dependent) grammars for their ability to fit different levels of a corpus that contained increasingly less frequent and more complex sentence forms. Like the whole-word hypothesis above, where evidence is sparse in the input (for the smaller, simple levels), the linear grammar fit the data better due to the fact that it is simpler—the hierarchical grammar was overly complicated for the level of evidence in the data. However, like the grapheme hypothesis, as the input became more diverse and numerous, the hierarchical grammar fit better because of the abstraction it offered: Additions to the data required few additions to the hierarchical grammar, whereas a linear grammar made as many additions to the grammar as there were additions to the data. Thus, the hierarchical grammar was able to capture sufficient regularity in the data to allow it to generalize to novel data, just like the grapheme hypothesis above.

It appears likely that other areas, such as spelling, could bear fruit from a simplicity-type analysis. Of course, the mappings that are optimal for reading may differ from those that are optimal for spelling. However, identifying mappings that are optimal for both or just one task may have useful pedagogical applications.

7. Conclusion

Over the last few decades, much attention has been given to identifying the skills that make good readers. Significantly less research has focused on identifying which spelling-to-sound mappings yield the greatest potential outcome in terms of decoding English text—that is, to understanding the structure of the mappings that the child must acquire, and how that structure changes as the number of words in the child’s reading vocabulary increases. In this article, we have applied a general cognitive principle—the Simplicity Principle—to ask what representations would be contained in an optimal reading system. In doing so, the Simplicity Principle states only that the hypothesis that provides the briefest total description of the data is to be preferred. The briefest encoding of the data was found for a hypothesis that contained a range of unit types (optimal hypotheses that were restricted solely to one type or another were not as concise), but it was composed mainly of grapheme–phoneme mappings. While strategy choice is likely to change depending on the specific vocabulary, another general finding was that within the optimal hypothesis some mappings are more useful than others.

The current findings should be informative for the design of reading programs because they provide a theoretically motivated rationale for choosing which mappings to teach early.
or later. Although it may be intuitively clear that a basic understanding of letter sounds is an essential skill for decoding purposes, English contains hundreds of letter combinations to represent phonemes. Our findings provide a motivation for choosing which of those will be most helpful for subsequent decoding. It is, of course, important to identify appropriate methods to introduce such material to a beginning reader (Wyse & Goswami, 2008). However, the main finding here is that the appropriate choice of materials the children are taught will maximize the possible outcome as far as decoding English text is concerned and may be worthy evidence for future intervention-based research to consider.

We have shown how applying the Simplicity Principle to reading provides a rich domain for enhancing our understanding of existing behavioral patterns. In applying this general cognitive principle to reading, we have demonstrated the potential gains to be made by those in the field of reading instruction, but we have also provided additional support for an already wide-ranging cognitive theory by demonstrating the breadth of data that match the results of the analyses. The results show the detailed picture of predictions to be revealed from considering the structure of the language to be read, and they suggest a fuller understanding of the representational units in reading for any language can only be enhanced by considering such data.

Notes

1. Often in the literature, the term consistency is restricted to refer to the consistency of the word body: A body that is always pronounced the same way is consistent; if it has more than one possible pronunciation it is not. Here, we use the term consistency in a more general sense to refer to whether any given orthographic unit (e.g., body or grapheme) has one or more possible pronunciations.

2. The current assignment of words to vocabulary sets uses written frequency as a proxy for age of reading acquisition. However, it appears that the specific order in which words are acquired does not seem to be critical to the results; when words are randomly assigned to vocabulary sets (so that they are unlikely to resemble any typical order of acquisition) similar results are obtained.

Acknowledgments

This research was supported by grant RES000222448 from the Economic and Social Research Council (UK), and grant F/215/AY from the Leverhulme Trust. This work was carried out while Janet Vousden was at the Department of Psychology, University of Warwick; Michelle Ellefson was at the Learning, Research and Development Center, University of Pittsburgh, and subsequently at the Department of Psychology, Virginia Commonwealth University; and Nick Chater was at the Department of Psychology, University of Warwick, and subsequently at the Department of Perceptual, Cognitive and Brain Sciences, University College London.
References


**Appendix**

We used MDL to provide a quantifiable measure of simplicity (L) for each hypothesis by computing the length of the code needed to specify both the hypothesis (the spelling-to-sound mappings) and the data (the pronunciations of words) under the current hypothesis:

\[
L = L(H) + L(D|H)
\]

For the whole-word hypothesis, \(L(H)\)—the whole-word mappings—is simply the length of code needed to describe the data themselves. Therefore, it was not necessary to calculate \(L(D|H)\) for the whole-word hypothesis.

In order to compute \(L(H)\), the length of the code needed to specify all the information concerning spelling-to-sound mappings that is used to re-create the data (i.e., produce pronunciations) must be calculated. Pronunciations are produced by applying spelling-to-sound mappings and, with the exception of the whole-word hypothesis, are assigned a probability based on the frequencies of the mappings used. The probability value that the
pronunciations are assigned is used to determine how well the mappings in question describe, or re-create, the data. \( L(H) \) can therefore be broken down further into the length of code needed to specify the precise orthographic to phonemic translations (the mappings) and the frequency with which they occur:

\[
L(H) = L(\text{mappings}) + L(\text{mapping frequencies})
\]

In summary, the length of code needed to describe the whole-word hypothesis is:

\[
L = L(H)
\]

and for all other hypotheses, there are three components to be calculated to measure the total code length:

\[
L = L(\text{mappings}) + L(\text{mapping frequencies}) + L(D|H)
\]

This allows us to compare different explanations of the same data by finding the total description length of each hypothesis, or representation—shorter description lengths indicate simpler explanations.

**Description Length for Mappings—\( L(\text{mappings}) \)**

A comprehensive list of mappings and their frequencies was obtained by parsing each word in a database into its constituent mappings according to each hypothesis (i.e., the word bread would be parsed by the head–coda hypothesis as brea—/bre/, d—/d/; as br—/br/, ead—/ed/ by the onset–body hypothesis; and b—/b/, r—/r/, ea—/e/, d—/d/ by the grapheme hypothesis) and simply summing the frequency of each mapping. The mappings themselves can be listed, and the length, in bits, of this list can again be calculated using Shannon’s (1948) noiseless coding theorem:

\[
\text{length} = \log_2(1/p)
\]

where \( p \) indicates the probability of an event occurring. Mappings of any unit size (e.g., bed—/bed/, on—/on/, ea—/i:/) can be listed as a string of letters (e, a) followed by a space and then a phonemic string (/i:/) and finally an end of line (eol) symbol. So the length in bits, can be calculated as

\[
\log_2(1/p(e)) + \log_2(1/p(a)) + \log_2(1/p(\text{space})) + \log_2(1/p(/i:/)) + \log_2(1/p(\text{eol}))
\]

Each letter, phoneme, and end-of-line symbol is associated with an equal probability.

**Description Length for Mapping Frequencies—\( L(\text{mapping frequencies}) \)**

The mapping frequencies form part of each hypothesis (with the exception of the whole-word) as they are used to calculate the probability of each pronunciation and therefore the description length, in bits, of encoding the frequencies must be calculated. The description length of the mapping frequencies would require a large number of bits if actual frequencies
were used, but it can be minimized by assigning numbers to different frequency ranges and substituting actual frequencies with the appropriate number. Frequency ranges, or bins, are defined by taking logs, to the base x, for each frequency and rounding up to the nearest integer. Using base 10 as an example, a frequency of 1 would be assigned the number 0, frequencies in the range 2–10 would be assigned the number 1, frequencies in the range 11–100 would be assigned the number 2, frequencies in the range 101–1,000 would be assigned the number 3, and so on. Following this procedure results in replacing each frequency value with a number, some of which will be more frequent than others. Each number can then be encoded using Shannon’s (1948) noiseless coding theorem (or more accurately, using adaptive encoding, which for large N approximates Shannon’s method), by associating each number with its probability of occurrence:

\[ \text{length} = \log_2(1/p) \]

The choice of base to use will affect both the description length of the hypothesis (larger bases will entail shorter lengths to describe the frequencies as there are fewer bins) and the description of the data when encoded by the hypothesis (smaller bases entail shorter lengths as the mapping probabilities, on which the pronunciations are based, will be more accurate), and is chosen to minimize the overall description length. Thus, for each hypothesis under consideration, simulations were run to calculate total description length for a range of different base values in order that the shortest description length can be found. The raw type frequencies for the grapheme–phoneme mappings used throughout these analyses can be found by inspecting Data S1 and Data S2.

**Description Length for Data—L(D|H)**

The description of the data given the hypothesis (i.e., the fit) can be measured by listing how well the mappings predict the pronunciation of each word. The general strategy is to produce all possible pronunciations for each word (given the mappings within each hypothesis) and list them in rank order of probability (based on the frequencies of the mappings). The target pronunciation for each word can then be simply represented by a number (or symbol)—its rank order in this list, instead of the actual pronunciation. The description of the data, given the hypothesis, is then reduced to describing a mapping from orthography to a simple rank. Again, the code length to describe the ranks is based on Shannon’s (1948) noiseless coding theorem, by associating each rank with it probability of occurrence:

\[ \text{Length} = \log_2(1/p(\text{rank})) \]

Where \(p(\text{rank})\) simply reflects the proportion of outputs where the target pronunciation is given with rank = 1, rank = 2, and so on. The more often pronunciations for different words are assigned the same rank, the shorter the code length will be to describe that rank. The length in bits, for the data (e.g., on -> Zn) will then be:
\[
\log_2(1/p(o)) + \log_2(1/p(n)) + \log_2(1/p(\text{space})) \\
+ \log_2(1/p(\text{rank }= 1)) + \log_2(1/p(\text{eol}))
\]

Hence, the more often the hypothesis provides the target pronunciation as rank 1 for each word, the shorter the code length will be to describe the data.

If the hypothesis is unable to offer the target pronunciation for a word at all (this result occurs for later simulations in which mappings are removed from hypotheses to reduce the total description length), then the target output is described by taking the most probable pronunciation, given the available mappings, and calculating a correction cost. The correction cost, derived by calculating the number of bits necessary to change the most probable pronunciation into the target one, is based on Levenshtein distance (Levenshtein, 1966). In brief, the less closely the most probable pronunciation matches the target one, the higher the correction cost, as each change requires \(\log_2(1/p(\text{any one phoneme}))\) bits.

### Producing pronunciations

To produce all possible pronunciations of a given word, each word is first parsed into constituent parts, according to the unit size under consideration. Head–coda and onset–rime unit sizes yield only one parse. For example, given the word *bread*, a head–coda parse gives *brea.d* and an onset–rime parse yields *br.ead*, because there is only one way of splitting the word into either head–coda or onset–rime. However, for grapheme units, there are many ways to parse a word, not all of which will correspond to available grapheme–phoneme mappings. For example, *bread* could theoretically be parsed many ways: *b.read*, *brea.d*, *br.ead*, *bre.ad*, *b.re.ad*, *b.rea.d*, *b.r.ea.d*, *b.r.ea.d*, *b.r.e.ad*, *b.r.ea.d*, *b.re.a.d*, *b.re.a.d*, *b.r.ea.d*, *b.r.e.ad*, *b.r.e.a.d*, *b.re.a.d*, *b.r.e.a.d*, but only the parses *b.r.ea.d* and *b.r.ea.d* correspond to available grapheme–phoneme mappings. For each parse, the relevant mappings are applied combinatorially and each pronunciation assigned an overall probability according to the product of the individual mapping probabilities. Individual mapping probabilities are based on the (type) frequency with which orthographic representations are associated with phonological representations. This frequency measures a directionless association between orthography and phonology (the sonograph, see Spencer, 2009), and contrasts with directional measures such as conditional grapheme–phoneme probabilities, or token-based measures that weight grapheme–phoneme counts by word frequency. Sonograph frequency has been shown to be a better predictor of reading accuracy in young British readers than either directional or token-based measures (Spencer, 2010). For the MDL analyses presented here, probabilities based on type rather than token frequency are appropriate because they provide a more accurate guide to pronunciation of the context-independent orthographic representations that are used to redescribe the data. Unless the context in which any given representation occurs is encoded, then the use of token frequencies could distort the likelihood with which orthographic representations are commonly pronounced across multiple contexts, which would be contrary to the goal of a child learning to read. In order to reduce the description length of encoding the frequencies (see above), mapping frequencies are represented by a number corresponding to the fre-
frequency range to which they belong (by taking the log of the actual frequency). The only information available concerning mapping frequency is obtained by computing the inverse of the log of the number that the mapping was assigned when originally taking logs to encode the frequency. To continue using base 10 as an example, if a mapping had been assigned number 3, then the frequency used in calculating pronunciations would be taken to be $10^3 = 1000$. A small amount of Gaussian noise is added to each pronunciation probability to avoid different pronunciations occurring with equal probability (this result happens occasionally because the encoded mapping frequencies have restricted resolution due to taking logs), and average code lengths for each hypothesis were based on at least 100 simulations. Pronunciations are filtered so those that violate the phonotactics of English (i.e., those that either contain disallowed sequences of phonemes, such as /gn/, or contain phonemes in disallowed positions, such as starting a word with /ŋ/) are not considered. The list of pronunciations is arranged in descending order of probability, and the target pronunciation assigned a rank.

Continuing the example using bread, the head–coda hypothesis uses the parse brea.d and mappings brea—/bre/, brea—/bri/, brea—/breI/, and d—/d/ to produce three outputs (in descending order of probability): /bred/, /bri:d/, and /bretd/. The onset–rime model applies mappings br -> /br/, ead -> /ed/, and ead -> /i:d/ to produce two outputs from the parse br.ead: /bred/ and /bri:d/. The GPC model produces many outputs from the parse b.r.e.a.d: For example, /bred/, /brea:d/, /breəd/, ..., /braɪəd/, /briːd/, ..., all of which form phonotactically improbable monosyllables; and produces eight outputs from the parse b.r.ea.d: /briːd/, /bred/, /bretd/, /briːt/, /bret/, /bretI/, /briːd/, and /briːt/, one of which (/braɪt/) is phonotactically improbable. The target pronunciation is described by ranks of 1, 1, and 2, respectively, for each hypothesis.

**Supporting Information**

Additional Supporting Information may be found in the online version of this article on Wiley Online Library:

- **Appendix S1.** Ranked list of grapheme–phoneme mappings, in descending order of contribution to the brevity of the total description length.
- **Data S1.** Grapheme–phoneme frequencies in monosyllabic British English.
- **Data S2.** Conversion of keyboard phonemes to IPA.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.