Behavior Research Methods (2020) 52:1292–1312
https://doi.org/10.3758/s13428-019-01317-y

Using information-theoretic measures to characterize the structure of the writing system: the case of orthographic-phonological regularities in English

Noam Siegelman[1] · Devin M. Kearns[1,2] · Jay G. Rueckl[1,2,3]

Published online: 16 January 2020

Abstract
It is generally well accepted that proficient reading requires the assimilation of myriad statistical regularities present in the writing system, including in particular the correspondences between words' orthographic and phonological forms. There is considerably less agreement, however, as to how to quantify these regularities. Here we present a comprehensive approach for this quantification using tools from Information Theory. We start by providing a glossary of the relevant information-theoretic metrics, with simplified examples showing their potential in assessing orthographic-phonological regularities. We specifically highlight the flexibility of our approach in quantifying information under different contexts (i.e., context-independent and dependent readings) and in different types of mappings (e.g., orthography-to-phonology and phonology-to-orthography). Then, we use these information-theoretic measures to assess real-world orthographic-phonological regularities of 10,093 mono-syllabic English words and examine whether these measures predict inter-item variability in accuracy and response times using available large-scale datasets of naming and lexical decision tasks. Together, the analyses demonstrate how information-theoretical measures can be used to quantify orthographical-phonological correspondences, and show that they capture variance in reading performance that is not accounted for by existing measures. We discuss the similarities and differences between the current framework and previous approaches as well as future directions towards understanding how the statistical regularities embedded in a writing system impact reading and reading acquisition.

Keywords orthography-to-phonology transparency · print-speech correspondences · information theory · word recognition · reading

Over the last several decades, reading research has become increasingly grounded in the idea that proficient reading requires the assimilation of subtle statistical regularities present in the writing system. This statistical view of writing systems had an impact on virtually all sub-domains of reading research, including computational models (e.g., Harm & Seidenberg, 2004; Rueckl, Zevin, & Wolf VII, 2019), studies of reading acquisition (e.g., Arciuli, 2018; Steacy et al., 2018; Treiman & Kessler, 2006), cross-language research (e.g., Frost, 2012; Seidenberg, 2011; Seymour et al., 2003), research on adult online sentence processing (e.g., Fine & Florian Jaeger, 2013) and individual-differences studies investigating the predictors of proficient reading (e.g., Arciuli & Simpson, 2012; Frost, Siegelman, Narkiss, & Afek, 2013). In the basis of all of these studies is the notion that there are myriad statistical regularities in the written input to be picked up by the reader.

In this paper, we focus on the role of statistical regularities in the process by which readers convert an orthographic string to the spoken form it represents. Such phonological decoding ability is widely understood to be one of the fundamental skills underlying proficient reading. An extensive literature provides data supporting this idea, including both (a) observations of deficits in phonological decoding in populations with reading disabilities (e.g., Scarborough, 1998; Vellutino, Fletcher, Snowling, & Scanlon, 2004), and (b) correlations between mapping skills and overall reading proficiency within typically

✉ Noam Siegelman
noam.siegelman@yale.edu

[1] Haskins Laboratories, New Haven, CT 06511, USA

[2] Department of Educational Psychology, University of Connecticut, and Haskins Laboratories, Storrs, CT, USA

[3] Department of Psychological Sciences, University of Connecticut, and Haskins Laboratories, Storrs, CT, USA

developing samples (e.g., Kearns, Rogers, Koriakin, & Al Ghanem, 2016; Perfetti, Beck, Bell, & Hughes, 1987).

This skill is considered to be particularly important in writing systems with an opaque (or deep) orthography. In contrast to transparent languages such as Spanish or Finnish, which exhibit nearly one-to-one correspondences between graphemes and phonemes, opaque writing systems display substantial inconsistency in the mapping of letters to sounds (Frost, Katz, & Bentin, 1987). English exhibits such notable opacity, mostly apparent in vowel graphemes (e.g., the grapheme *ea,* which is pronounced differently in the words *bead, head,* and *steak*; the grapheme *i* in the words *mint, pint,* and *helium*).

But how can the degree of consistency (or inconsistency) in the mapping of an orthographic string to phonology be quantitatively assessed? What is the degree of consistency in a given word compared to another? Answering these questions is critical in order to assess one's skill of efficient mapping, to estimate the exact role of sensitivity to consistency in accounting for reading outcomes, and to accurately manipulate the degree of consistency of words or non-words in experimental paradigms (such as the Strain task, Strain, Patterson, & Seidenberg, 1995). Nevertheless, to date, there is no agreed-upon method to estimate consistency. Generally speaking, three approaches currently exist (and see Borleffs, Maassen, Lyytinen, & Zwarts, 2017, for a similar classification in the context of orthographic transparency across languages).

The first, the *regularity approach* (e.g. Baron & Strawson, 1976; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Forster & Chambers, 1973; Rastle & Coltheart, 1999) assumes that there is a set of rules dictating the regular correspondences between graphemes and phonemes (GPC rules; e.g., "*i* is regularly read as /ɪ/"). Note that grapheme is defined as a letter or a letter sequence that correspondences to a single phoneme (e.g., *e* in the word *bed*; *ea* in the word *head*). Per this approach, words that follow these rules are considered regular (e.g., *mint*) and words that do not are considered irregular (e.g., *pint*). Similarly, per this approach, a set of spelling rules specifies the regular spelling of each phoneme. It is thus a categorical classification system, where words are treated as either regular or irregular. Indeed, regularity was shown to impact word recognition measures, such as reaction time in lexical decision and naming tasks (see, e.g., Baron & Strawson, 1976; Glushko, 1979; Seidenberg, Waters, Barnes, & Tanenhaus, 1984; Stanovich & Bauer, 1978). Nevertheless, the use of a categorical division of words to regular and irregular was strongly criticized (e.g., Plaut et al., 1996; and see Glushko, 1979 for an earlier version of this criticism).

A second approach, the *statistical approach*, avoids a categorical distinction, and instead places each word's degree of consistency at a certain point over a continuum (Jared, McRae, & Seidenberg, 1990). Operationally, common measures of statistical consistency typically examine the ratio between the number of *friends* and *enemies* of that word. Consistency measures are most commonly used to capture the associations between orthography and phonology at the body-rime level: Thus, *body-rime consistency* quantifies the ratio of similar vs. different phonological realizations of a word body (e.g., *-ead* in the word *head*; e.g., Jared, McRae, & Seidenberg, 1990). Practically, to assess the body-rime consistency of *bead*, one would calculate the ratio between the number of friends where *-ead* is similarly pronounced as /ɛd/ (e.g., *dead, bread*, etc.) to the number of enemies where *-ead* is pronounced differently (e.g., *bead*). The consistency approach can be similarly used to quantify correspondences at a smaller grain size: Thus, *vowel consistency* focuses on friends vs. enemies at the grapheme-level (e.g., Chateau & Jared, 2003; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995). Practically, then, the vowel consistency of the pronunciation of *i* in the word *mint* (where *i* ➔ /ɪ/) is the ratio between the number of friends where *i* is pronounced as /ɪ/ (e.g., *bin*, *sing*, etc.) to the number of enemies where *i* is pronounced otherwise (e.g., *pint*). Generally, a large number of friends versus enemies constitutes a more consistent pronunciation at a given grain size, where a large number of enemies compared to friends reflects an inconsistent reading[1]. Importantly, studies employing the statistical approach show that the ease in which a pronunciation can be derived from an orthographic code varies continuously, reflecting sensitivity to the probabilities of the pairings between orthographic and phonological units. The two types of consistency measures – body-rime consistency and vowel consistency – exemplify an additional important advantage of the consistency approach, on top of its continuous measurement: measures of consistency can capture transparency at different grain sizes, either at the grapheme-to-phoneme in isolation, or in a broader context in which the grapheme appears. This is important because English and other opaque languages often have context-dependent grapheme-phoneme mappings, that is, the context in which a given grapheme appears impacts its pronunciation. Such sensitivity to context-dependent information is considered to play an important role in proficient reading in English (e.g., Steacy et al., 2018; Venezky, 1999; Ziegler & Goswami, 2005), and readers indeed display sensitivity to such context-dependent information. For example, *cheam* is almost read with *ea* ➔ /i/ (as in *beam*), but *chead* is sometimes read with *ea* ➔ /i/ and in other cases with *ea* ➔ /ɛ/ (as in *head*; Treiman, Kessler, & Bick, 2003). Lastly, note that consistency measures can also be used to quantify the mapping of the other

[1] Although consistency is typically computed as the ratio of friends to friends plus enemies, other formulations have sometimes been used. For example, Graves and colleagues (2010) defined consistency as the difference between the number of friends and enemies, and in Glushko's (1979) seminal work consistency was treated as a categorical measure: A word was considered consistent if it had no enemies and inconsistent otherwise.

direction, between phonology to orthography (i.e., spelling consistency; e.g., Treiman & Kessler, 2006).

A third way to quantify orthographic-phonological regularities, the *entropy approach,* focuses on the impact of uncertainty in the mapping between the two. Entropy, a term from Information Theory (Shannon, 1948), examines the extent of uncertainty over a distribution of events. A distribution of multiple equiprobable events implicates high uncertainty, whereas a distribution with a single event (that is thus fully predictable) implicates zero uncertainty (and see below for a formal definition and examples). In the context of orthography to phonology, the entropy approach does not consider the consistency of the specific pronunciation of a given grapheme in a word (e.g., in the word *mint*, to what extent is the mapping *i* ➔ /ɪ/ consistent), but rather the full distribution of possible phonemes that can be reflected by a given grapheme (or in the other direction – to calculate spelling entropy – the distribution of possible spellings of a given phoneme).

The main advantage of the entropy approach is therefore that by focusing on uncertainty it accounts for an additional facet that is not captured by the consistency approach. The difference between entropy and consistency was exemplified in the corpus analysis by Protopapas and Vlahou (2009) of Greek, which addressed entropy in the mapping from orthography-to-phonology and from phonology-to-orthography. To exemplify the difference between entropy and consistency, the authors describe two contrasting cases: the phoneme /g/, which is spelled as γκ in 85.5%, and as γγ in 14.5%; and the phoneme /ç/, which is spelled as χ in 85.0% of the times, as οι in 7.0% of the cases and as ι in 6.9% (and in other uncommon forms in the remaining cases). A consistency approach would consider the mapping of /g/ ➔ γκ and of /ç/ ➔ χ as similar (as both have a similar ratio of friends vs. enemies, ~ 85%/15%). An entropy approach, however, accounts for the difference in the distributions of the two spellings, and posits that the first mapping implicates less uncertainty, due to less random distribution over the possible spellings of the phoneme (and see mathematical explanation below). Note that in the context of cross-linguistic differences, this feature of entropy led some to use it as a proxy for a writing system's orthographic depth. For example, Borgwaldt, Hellwig, and De Groot (2004) compared the extent of uncertainty in the mapping of word-initial graphemes to phonemes across five languages, and later studies have used these entropy values as a measure for the extent of transparency vs. opaqueness of each writing system (Borgwaldt, Hellwig, & De Groot, 2005; Ziegler et al., 2010).

The goal of the current study is to offer a comprehensive information-theoretic approach for quantifying the structure of the mapping between an orthographic string and phonological form. Similarly to the entropy approach, our quantification takes into account the extent of uncertainty in the orthography-to-phonology mapping. Most importantly, however, our approach does not center on entropy alone, rather it uses additional information-theoretic measures to also consider the extent of (un)predictability of a phonological realization given an orthographic string (which is the typically the focus of the statistical consistency approach). Our approach thus merges insights from both the entropy and statistical consistency approach. Moreover, by using general information-theoretic terms, our approach highlights the similarity between orthographic-phonological decoding, other component processes of reading, and other aspects of language processing in general, all of which are affected by the degree of uncertainty and unpredictability embedded in the input. Thus, such effects have been documented in fields such as speech perception (e.g., Frank, 2013), speech production (e.g., Cohen Priva, 2015, 2017), syntactic processing (e.g., Hale, 2006; Linzen & Jaeger, 2015), morphological processing (e.g., Milin, Kuperman, Kostić, & Baayen, 2009) and sentence reading (e.g., Lowder, Choi, Ferreira, & Henderson, 2018; Smith & Levy, 2013).

At a practical level, we use tools from Information Theory (Shannon, 1948), which defines uncertainty and unpredictability as the amount of information present in an input. Importantly, Information Theory's toolbox includes a diverse set of measures, which can be used to capture different aspects of the information distribution of an input. In the context of the present investigation, these tools enable us to assess not only the degree of uncertainty (i.e. *entropy*) in a distribution of possible pronunciations (e.g., Protopapas & Vlahou, 2009), but also the extent to which a given pronunciation is surprising (or unpredictable) given a grapheme (in information theory terms, the *surprisal* of an event), and the difference between the expected amount of information and the observed amount of information in a grapheme-to-phoneme correspondence (*information gain*). Our framework thus combines advantages of existing approaches in a theoretically motivated manner. As we will argue below, it also presents a flexible approach that can be used to quantify information under different contexts (i.e., context-independent and -dependent readings) and in different types of mappings (e.g., orthography to phonology and phonology to orthography).

Below we start by providing a glossary of the three relevant information-theoretical metrics – entropy, surprisal, and information gain. Then, we exemplify their use in assessing orthography-to-phonology mapping (as well as phonology-to-orthography) using a corpus of English monosyllabic printed words and their pronunciations. Next, we estimate the degree of transparency of the mapping between orthography-to-phonology of each of the words in the corpus, and use these word-level metrics to examine whether the information-theoretic measures capture variance in actual reading performance (using available large-scale datasets from naming and lexical decision tasks), while also comparing their predictive value to that of common consistency

measures. In a nutshell, our analyses demonstrate how information-theoretical measures can be used to assess the information distribution of a writing system, and show that these measures explain variance in reading behavior that is not accounted for by common consistency measures.

Quantifying information

In this section, we briefly review three key terms in Information Theory, surprisal, entropy, and information gain, which serve as the basis for the current investigation.

Surprisal

Surprisal (sometimes referred to as self-information) quantifies the amount of information carried by a single event given its probability. More predictable events are less surprising, and carry less novelty in terms of the information they offer. The common unit of surprisal is "bits of information". Mathematically, surprisal of an event *i* is defined as:

$$S_i = -\log_2 p(i) \quad (1)$$

Importantly, the probability of the event (i.e., p(i)) can either refer to its unconditional (marginal) probability, or to some conditional probability. Consider, for example, a hypothetical situation in which the grapheme *ea* has only two possible readings, either as the phoneme /i/ (as in the word *bead*), or as /ɛ/ (as in *head*; see also Table 1). Overall (unconditional of any other event), *ea* is read as /i/ in 60% of the cases, and as /ɛ/ in the remaining 40%. Given these probabilities, the pronunciation of the vowel in the word *head* is more surprising than in the word *bead*: *head* has a surprisal value of $-\log_2 0.4 = 1.32$ bits, versus $-\log_2 0.6 = 0.74$ bits in *bead* (see also Table 1). However, the conditional surprisal values of this mapping may be different. Assume that given the coda *-d*, the grapheme *ea* is more likely to be pronounced as /ɛ/, say in 70% of the cases (and as /i/ only in 30%). As a result, the surprisal of the vowel grapheme in *head*, conditioned on the coda, is lower than that of *bead*: $-\log_2 0.7 = 0.51$, versus $-\log_2 0.3 = 1.74$. Note that this toy example already demonstrates how notions of Information Theory can be used to assess information in different grain sizes, by focusing on either unconditional or conditional (i.e., context-independent or context-dependent) probabilities. It also shows the flexibility of this approach in accounting for the information structure of different parts of the input: Namely, conditional probabilities can be calculated over different parts of a word: coda, onset, following/preceding grapheme, etc. Another thing to note is that while these examples focus on the information in the mapping between graphemes to phonemes, surprisal (as other information-theoretic metrics) can also be applied to quantify the information in the opposite direction: that is, the mapping of phonemes onto graphemes.

Entropy

While surprisal concerns the information provided in a single observed event, entropy concerns the uncertainty present in a *distribution* of events. The more unpredictable events are in a distribution, the more information they carry. Hence, random distributions are characterized by high entropy, whereas highly skewed distributions (e.g., a distribution with a single highly probable event) are characterized by low entropy. The entropy of a distribution with only one possible event (i.e., a single event with $p = 1$) equals 0.

Mathematically, entropy is the *expected value* of the amount of information over a distribution of events. It is thus inherently linked to surprisal – surprisal quantifies the amount

Table 1 Calculation of information-theoretic measures for the EA grapheme

Measure Type	Calculation			
Unconditional	EA ➔ /i/ (p = .6)		EA ➔ / ɛ / (p = .4)	
Surprisal (S)	$-\log_2 0.6 = 0.74$		$-\log_2 0.4 = 1.32$	
(Shannon) Entropy (H)	$-(0.6 * \log_2 0.6 + 0.4 * \log_2 0.4) = 0.97$			
Information gain (H-S)	$0.97 - 0.74 = 0.23$		$0.97 - 1.32 = -0.35$	
Coda-conditional	-EAD (p = .4)		-EAT (p = .6)	
	EA ➔ /i/ (p = .3)	EA ➔ / ɛ / (p = .7)	EA ➔ /i/ (p = .8)	EA ➔ / ɛ / (p = .2)
Surprisal (S)	$-\log_2 0.3 = 1.74$	$-\log_2 0.7 = 0.51$	$-\log_2 0.8 = 0.32$	$-\log_2 0.1 = 2.32$
Conditional entropy (H)	$-(0.3 \times \log_2 0.3 + 0.7 \times \log_2 0.7) = 0.88$		$-(0.8 \times \log_2 0.8 + 0.2 \times \log_2 0.2) = 0.72$	
Information gain (H-S)	$0.88 - 1.74 = -0.86$	$0.88 - 0.51 = 0.37$	$0.72 - 0.32 = 0.40$	$0.72 - 2.32 = -1.60$
Markov entropy	$.4(0.3 \times \log_2 0.3 + 0.7 \times \log_2 0.7) + .6(0.8 \times \log_2 0.8 + 0.2 \times \log_2 0.2) = 0.78$			

of information of each single event, and entropy is the expected value of such surprisal values in some distribution of events. As a result, the unit of entropy is also bits of information. This link can also be seen in the formula used to calculate entropy, which shows that entropy is the sum of surprisal of each event weighted by its probability:

$$H = -\sum_i p(i) * \log_2 p(i) \quad (2)$$

Much like surprisal, entropy can be calculated over different distributions. For our purposes we again focus on unconditional entropy (sometimes referred to as *Shannon entropy)* as well as on conditional entropy. Concretely, we use unconditional entropy to calculate the overall uncertainty in the mapping of a vowel grapheme to phonemes, regardless of the context in which it appears. For example, in the toy example above (*e*a ➔/i/ in 60%, ea ➔ /ɛ/ in 40%), the unconditional entropy of the grapheme *ea* is $-(0.6 * \log_2 0.6 + 0.4 * \log_2 0.4) = 0.97$ bits (see also Table 1). Compare this value to the one resulting from calculating the entropy of the grapheme *ea* conditional on the coda *-d*: $-(0.3 * \log_2 0.3 + 0.7 * \log_2 0.7) = 0.88$ bits. In this hypothetical case, the conditional entropy of *ea* is lower given the coda *-d* than it is unconditionally. This is because the distribution given this coda is more skewed (i.e., less random) than the unconditional distribution.

This leads us to another measure: *Markov entropy.* This measure quantifies the *average* conditional entropy over a set of conditioning contexts. The Markov entropy of the grapheme-phoneme mapping given the coda would be the conditional entropy given each of these codas, weighted by their marginal probabilities. Mathematically, this is expressed by the formula:

$$H_{markov} = -\sum_i p(i) \sum_j p(j|i) * \log_2 p(j|i) \quad (3)$$

Conceptually, in the context of grapheme-to-phoneme mapping, this value examines the amount of uncertainty of some pronunciation given some part in the word. Thus, if some context (e.g., coda, onset, etc.) is highly predictive of a pronunciation, Markov entropy is low (low uncertainty). In contrast, if a context does not predict the phonological realization of a vowel grapheme, Markov entropy is high. The use of Shannon vs. Markov entropy, and of Markov entropy under different contexts, allow us to examine the overall uncertainty present in different grain sizes. As an example, consider the hypothetical case above (Table 1) where there are only two possible codas following the vowel *ea*: *–d* (e.g., *head*) and *–t* (e.g., *cheat*). Say that the coda *–d* is less frequent than *–t*: 0.4 vs. 0.6. The Markov entropy of the pronunciation of *ea* given the coda will be the average of the conditional entropy given the coda *–d* and that given *–t*, weighted by their marginal frequency (see calculation in the last row of Table 1).

Some general characteristics of entropy measures should be noted. First, we emphasize that similar to surprisal, entropy can be calculated to capture different aspects of the input: for example, it can be calculated over different distributions (unconditional and conditional), under different constraining contexts (e.g., coda vs. onset) and for different types of mappings (e.g., the uncertainty in the mapping of orthography-to-phonology, or that of phonology-to-orthography). Second, note that while the toy example deals with distributions that include only two possible events, entropy measures are not constrained to such cases and can be calculated on distributions with any number of events. This is of course necessary when dealing with correspondences, when the same grapheme can have multiple possible realizations (e.g., consider the grapheme *ea* in *heat, head, steak,* and *heart*). Third, and as noted above, entropy captures something that is typically not accounted for by looking at the probabilities of single events (either via surprisal measures, or via standard consistency measures). To illustrate, consider two hypothetical graphemes: *A* and *B*. *A* is pronounced as the phoneme /A_1/ in 70% of the cases, as /A_2/ in 15%, and as /A_3/ as 15%. In contrast, *B* is pronounced as /B_1/ in 70%, as /B_2/ in 25%, and as /B_3/ in 5%. The probability of the most common reading of each of the two graphemes is the same (and thus so is their surprisal value): both *A*➔/A_1/, and *B*➔/B_1/ has a probability of 70%. However, the grapheme *A* across its different pronunciations has a higher entropy than *B*. This is because the distribution of its possible readings is more random, containing higher uncertainty.

Information gain

Entropy and surprisal each capture a different aspect of the information structure of an input: surprisal quantifies the (un)predictability of a single event, while entropy quantifies the uncertainty across a distribution of events. In the context of orthography-to-phonology mapping, surprisal thus captures the extent of unpredictability of a given grapheme-to-phoneme correspondence (e.g., the mapping of *ea* ➔ /i/ in the word *bead*; either unconditionally or given context), and entropy captures the overall uncertainty in the distribution of possible pronunciations of a grapheme (again, either independently or conditional on context). *Information gain* is affected by both the unpredictability of a given event as well as the uncertainty of the full distribution of possible events.

Mathematically, *information gain* is simply entropy over a distribution minus the surprisal of the observed event, that is:

$$information\ gain_i = H - S_i \quad (4)$$

Conceptually, information gain thus quantifies the difference between the expected information value of an event (i.e., entropy) and the actual information provided by it (i.e.,

surprisal). It can be thus used to quantify the difference between the expected information of a grapheme given its possible pronunciations and the information provided by the actual grapheme-phoneme pairing. Importantly, similar to entropy and surprisal, information gain can be quantified over unconditional probabilities, or using conditional values (in different grain sizes). Note also that because information gain is the difference between entropy and surprisal, its unit is bits, too.

To illustrate, consider the hypothetical example above, regarding the unconditional mappings of *ea* (see also Table 1). Given a grapheme *ea* a reader expects 0.97 bits of information (the unconditional entropy value calculated over the distribution *ea* ➔/i/ in 60%, *ea* ➔/ɛ/ in 40%; i.e., $-(0.6 \times \log_2 0.6 + 0.4 \times \log_2 0.4)$). Let's assume that the reader now encounters the word *head*. In this case, the surprisal of the grapheme-phoneme mapping of the vowel is $-\log_2 0.4 = 1.32$. The unconditional information gain of the word *head* is therefore negative: the actual observed information is higher (i.e., the event is more unpredictable) than the expected value: $0.97 - 1.32 = -0.35$ bits. In contrast, the unconditional information gain of the word *bead* is positive: the actual information (i.e., $-\log_2 0.6 = 0.74$) is lower than the expected, and equals $0.97 - 0.74 = 0.23$ bits. Importantly, similar to entropy and surprisal, information gain can be applied not only to unconditional (context-independent) probabilities, but also to conditional (context-dependent) events. To compute conditional information gain, we use the difference between the conditional entropy over some context and the conditional surprisal. For example, to calculate the coda-conditional information gain of the word *head*, we take the conditional entropy of *ea* given the coda *-d*, and subtract the conditional surprisal of *ea*➔/ɛ/ given *-d*. Thus, in the hypothetical example above (where given coda *-d*, *ea*➔/i/ in 30%, *ea*➔/ɛ/ in 70%), the information gain of *head* is now positive: the conditional entropy is 0.88 bits (i.e., $-(0.3 \times \log_2 0.3 + 0.7 \times \log_2 0.7)$), the conditional surprisal is 0.51 bits (i.e., $-\log_2 0.7$), and thus the information gain is $0.88 - 0.51 = 0.37$ bits. Conceptually, this means that given the coda *-d* the pronunciation /ɛ/ is more predictable than the average event. Note that in cases where entropy and surprisal are identical (i.e., when the expected information equals the actual information), information gain is 0.

Information in orthography to phonology: real-world examples

In the previous section, we used simplified examples to explicate how information-theoretic constructs can be used to characterize the structure of the orthography-phonology mapping. In this section, we apply these constructs to a large-scale corpus of English words.

Corpus description

Our estimation of the information in English words is based on an analysis of words from the Unisyn database (Fitt, 2001) which contains 117,625 English words and their pronunciations. Unisyn was used instead of CELEX (Baayen, Piepenbrock, & Van Rijn, 1995) for two reasons. First, Unisyn contains more words and their pronunciations than CELEX. Second, Unisyn contains Perl scripts to adapt its "accent-independent keyword lexicon" to a wide variety of accents, whereas CELEX pronunciations are given only in Received Pronunciation. For this study, the General American pronunciations were used because the behavioral data were from the United States-based English Lexicon Project (ELP; Balota et al., 2007). Note that all analyses below are based on a subset of the Unisyn corpus containing only mono-syllabic words, with a total of 10,093 word types. Also note that this set of items includes some proper names and acronyms. In addition, the database includes words with apostrophes (e.g., can't), but the apostrophe was not included in the list of graphemes in a word (e.g., in *can't* the list included the graphemes *c*, *a*, *n*, and *t*)[2].

Next, we constructed a program to match the graphemes in each word with each phoneme in its pronunciation. This program operated using two data sources. One was the Unisyn database already described. The other was a master list of grapheme-phoneme correspondences (GPCs). The GPCs were constructed by the research team first by connecting a single phoneme with graphemes of multiple lengths. For example, the phoneme /i/ was associated with the graphemes *e* (*median*), *ey* (*key*), *ea* (*team*), *ee* (*meet*), *ey* (*key*), and *i* (*quiche*), among others. As much as possible, consonant phonemes were associated with graphemes containing only consonant letters. For example, the /ʤ/ sound was associated with *g* in *strange*—not *ge*—despite that the *E* is a marker that the *G* is pronounced /ʤ/. This also applied to cases of *gu* (e.g., in *guest*, *guide*, and *guy*) where the *U* was marked as silent. Similarly, vowel phonemes were usually associated with graphemes containing only vowel letters. For example, in the word *half*, the vowel phoneme (/æ/) was associated with *a* in rather than with the sequence *al* despite the presence of additional words with the *al* pattern where *l* has no sound (e.g., *calm, palm*).

[2] This decision was made since the English Lexicon Project – which was used for the validation of our measures – also includes such words. To make sure that the decision to keep words with apostrophes in the database did not affect our results significantly, we also estimated the information-theoretic measures based on a sub-set of the corpus that did not include words with apostrophes. Then, we examined the correlation of the information-theoretic computed on the corpus with and without words with apostrophes over the remaining words (i.e. words without apostrophes; n=9289 words). The observed correlations for all information-theoretic measures (surprisal, entropy, and information gain; unconditional, coda-conditional and onset-conditional) were near perfect, ranging between r = 0.977 to r = 0.999. Thus, the decision regarding the inclusion (or exclusion) of words with apostrophe seem to only have a negligible effect on the estimation of the information-theoretic measures.

There was one notable exception to the vowel-phoneme vowel-letter principle, where rhotacized vowels were concerned. In part, this was because the rhotacized /ɝ/ phoneme is considered a single phoneme. In other cases, the vowel and *R* made separate sounds but are so commonly associated with each other they were treated as a single grapheme (e.g., /ɑɹ/ for *ar* in *car*; /eɪɚ/ for *-are* in *care*; aɪɚ for *-ire* in *fire,* /ɔɹ/ for *or* in *for*). There was one similar consonant case, *qu,* where the two letters are almost always pronounced with the two-phoneme unit /kw/.

In almost all cases, graphemes consisted only of letters that appeared consecutively. There was one exception, the vowel-consonant-*E* graphemes. These graphemes consist of a vowel letter (i.e., *A, E, I, O, U,* or *Y*), a single consonant letter, and a single *E* (e.g., in *make, mete, mite, note, cute,* or *type*). So, [*aeiou*]-*e* was considered a grapheme (e.g., *a-e*) and so on. Also note that the letters *w* and *y* could be classified as either consonants or vowels: They were classified as a consonant letter when they corresponded to a consonant phoneme, and as a vowel when they corresponded to a vowel sound.

Treating vowel-*R* and vowel-consonant-*E* patterns as single GPCs did increase the consistency of GPCs relative to body-rime units. For example, by creating the *ar*➔/ɑɹ/ pattern, there were fewer instances of *a*➔/ɑ/ than if the *A* and *R* were treated as separate graphemes (e.g., *a*➔/ɑ/ and *r*➔/ɹ/. As a result, the coding system might have slightly over-estimated words' GPC consistency.

The program operated by trying every GPC in the master list against every letter and sound in the word. For example, for the word *get,* coded /gEt/ in X-SAMPA, the program would examine the letters, moving from left-to-right, so *g* first. Then it would try to find a GPC that contained a *g* letter and tried to locate one that contained one of the sounds in /gEt/. It would locate g=/g/ and code those together. It would repeat this process for the remaining letters. It was designed to do this for GPCs with multiple letters (e.g., tch=/C/) and multiple phonemes (e.g., x=/ks/). In general, one of the underlying principles of the GPC program was a minimization of the number of graphemes. Thus, in cases where a consistent larger grapheme could be extracted, the program tended to do so (e.g., *qu* in the word *quick*). The decoded versions of the words can be found in the Supplementary Material "https://osf.io/kfme8/", and can also be examined using the Phinder program available at https://phinder.devinkearns.org.

In what follows, we use this corpus to assess the information in the mapping between orthography and phonology (and between phonology and orthography) in English. We do so by examining surprisal, entropy and information gain of English vowels in different contexts (unconditional, coda-conditional, and onset-conditional). These analyses are meant to provide intuitions regarding how these metrics capture the information distribution of letter-to-sound correspondences in English.

Surprisal

Figure 1 presents examples for three (grapheme-to-phoneme) surprisal measures: unconditional, conditional on coda, and conditional on onset, for three words containing the grapheme *ea*: *bead*, *head*, and *health*. To emphasize, these values are based on the real-world probabilities of *ea* readings, based on the corpus. Note that in terms of unconditional surprisal, *bead* has a lower surprisal (i.e., higher predictability) than *head* and *health*. This is because the pronunciation *ea*➔/i/ is more common than *ea*à/ɛ/ across the full corpus. Calculations conditional on the coda, in contrast, produce higher surprisal (more unpredictability) for *bead* compared to *head*. This means that given the coda *-d*, the reading /i/ is less common than /ɛ/. Note also that *health* has an even lower conditional surprisal value than *head*, suggesting that the mapping of *ea* to /ɛ/ has a higher probability given the coda *-lth* than the coda *-d*. Note also that in these two words, coda-conditional surprisal values are lower than onset surprisal. This suggests that the codas in these two words (*-d* and *-lth*) are more predictive of the mapping *ea*➔/ɛ/ compared to their onset, *h-*. As will be shown in the next section, this is in fact a representative feature of the full corpus.

Entropy

To exemplify the use of entropy, we turn to an analysis of the uncertainty in vowel pronunciations given different contexts (independent of context, coda-dependent, and onset-dependent) across all monosyllabic English words in the corpus. Figure 2 presents the entropy in grapheme-to-phoneme mapping of all vowel graphemes that appear in more than 100 out of the 10,093 word types in the corpus (that is, in at least ~1% of words). Note that this figure includes the unconditional (Shannon) entropy of each vowel, as well as the Markov entropy (mean conditional uncertainty) given the coda, as well as the onset. Several observations can be made from these results. First, in English, there is more uncertainty in unconditional, versus conditional, reading of vowel graphemes. In other words, the context in which a vowel appears (either coda or onset) reduces the uncertainty regarding its pronunciation. Second, in most cases, the coda is more predictive (i.e., reduces uncertainty to a larger extent) compared to the onset. Both of these findings are consistent with earlier investigations showing that the context in which a vowel grapheme appears – and the coda of monosyllabic words in particular – is a reliable cue for its pronunciation (Aronoff & Koch, 1996; Kessler & Treiman, 2001; Stanback, 1992). Next, to examine this issue more closely, we calculated the overall Markov entropy conditional on coda vs. onset (averaged across all vowel graphemes, weighted by their marginal probability), and found that indeed coda Markov entropy is lower than onset Markov entropy: 0.25 versus 0.37 bits. Both of these values are smaller than the overall Shannon entropy

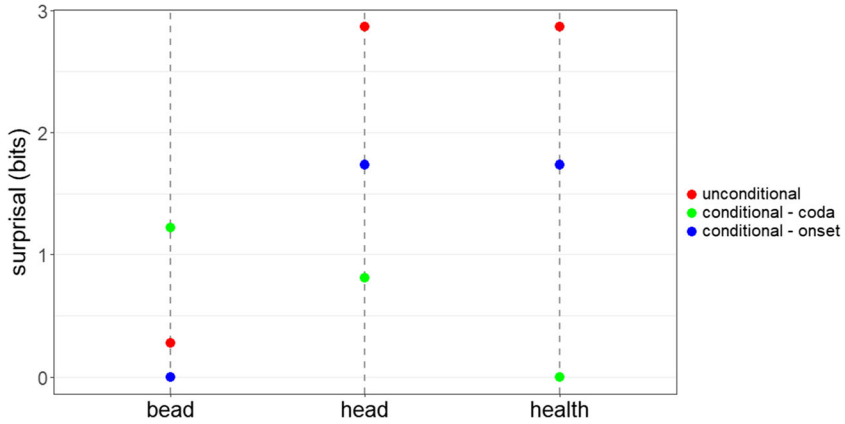


Fig. 1 Examples for surprisal values: unconditional, conditional on coda, and conditional on onset, for three words containing the grapheme *ea*

(across all vowel graphemes): 0.69 bits. Third, and importantly, Fig. 2 shows that vowels differ from each other in their uncertainty in general and, moreover, that uncertainty varies in the three grain sizes (or in other words: the uncertainty of different vowels depends on whether and how it is conditionalized). While outside the scope of this paper, this information can be used to assess to ease or difficulty in acquiring the orthography-to-phonology mapping of a given grapheme and to assess the extent to which a reliance on context helps in deciphering different graphemes.

Information gain

We next calculated the information gain (difference between entropy and surprisal) of all mono-syllabic words in the corpus. We calculated both unconditional information gain values as well as coda-conditional and onset-conditional values. Note that across all words, there was a correlation between unconditional and conditional values (unconditional and coda-conditional correlation: $r = 0.51$; unconditional and onset-conditional correlation: $r = 0.67$). This is expected: events that have some level of predictability independent of context have in many cases a similar level of predictability given a context (e.g., the pronunciation of *ee* is nearly always /i/ both given a context as well as unconditionally, and as a result information gain across grain sizes is similar). Importantly, however, this correlation is not perfect, and in many cases unconditional and conditional values differ substantially. Figure 3 presents some examples of unconditional and coda-conditional information gain values of words calculated using real-world probabilities. Note the difference between the words *head* and *bead*. Unconditionally, *bead* has a positive information gain, while *head* has a negative gain. When conditioned on the coda, however, the opposite pattern is observed: *head* has a higher information gain compared to *bead.* The word *health* has a negative information gain unconditionally, but a near-0 value conditional on the coda. This means that given the coda *-lth* the actual information provided by *ea*➔/ɛ/ (i.e., surprisal of *ea*➔/ɛ/ given the coda *–lth*) is similar to the expected value (i.e., entropy of the grapheme *ea* given *–th*). The other two points in this graph represents information gain values of two famous examples for grapheme-phoneme irregularity in English: *mint* versus *pint*. Note how *mint* has positive information gain values (conditionally and unconditionally), whereas *pint* presents a strong negative deviation from expected information levels, in both dimensions[3]. This is expected given the difference in predictability of the pronunciation of these two words.

Behavioral impact

So far, we introduced terms from information theory which we hypothesize are relevant for the mapping of orthography onto phonology, and exemplified how these can be used to assess the real-world information structure of a writing system. We now turn to examine whether they indeed impact reading performance, and how their explanatory power compares to typical measures of orthography-to-phonology consistency.

To do so, in the following we use information-theoretic metrics to predict inter-item variance in behavioral data. Most of our investigation focuses on the naming portion of the English Lexicon Project (ELP; Balota et al., 2007), which includes data of 40,481 English words, collected on large samples of native English university students ($N = 444$ in the naming task). From the 40,481 words in ELP, we focus here only on mono-syllabic words that also exist in our corpus: 5713 words overall. We examine how well information-theoretic measures predict between-item variance in response latencies and accuracy. In the first section, we simply examine whether the information-theoretic measures indeed have an effect on behavior (controlling for general measures:

[3] in fact, from the 10,093 words in the corpus, *pint* had an unconditional and coda-conditional information gains in the 3rd and 1st percentile, respectively.

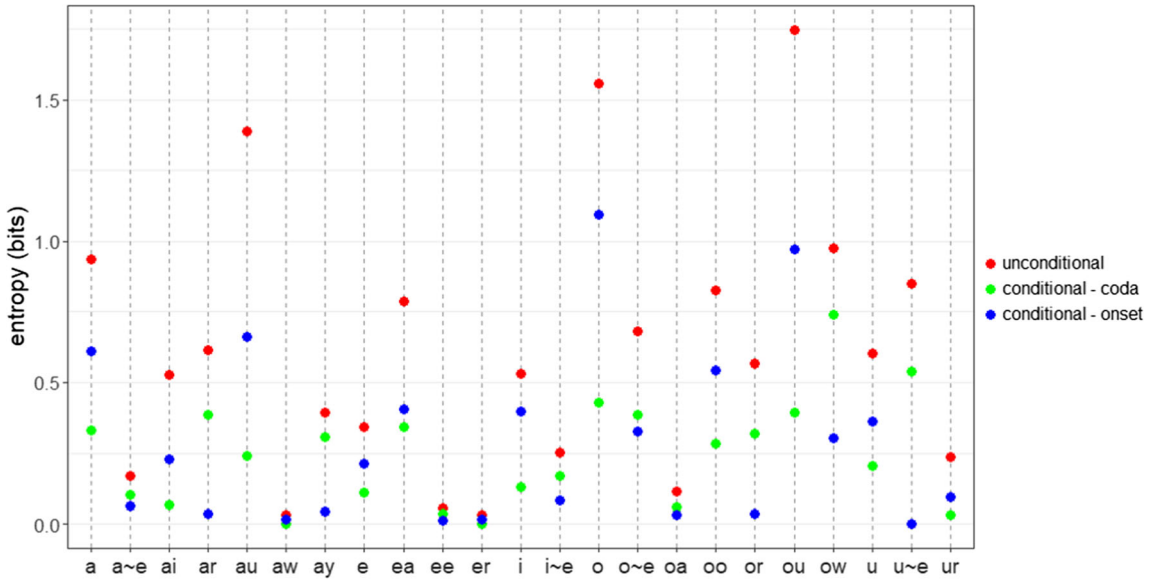


Fig. 2 Entropy of vowel graphemes (orthography to phonology): unconditional, conditional on coda, and conditional on onset entropy. This plot includes all vowel graphemes that appear in at least 100 words in the corpus

frequency, length, and articulation parameters). In the second section, we compare the information-theoretic measures to standard measures of orthography-to-phonology consistency. In the third section, we demonstrate the flexibility of the current approach by further exploring the relevance of information-theoretic measures to reading behavior in other contexts (investigating lexical decision behavior, and comparing information in orthography-to-phonology mapping to phonology-to-orthography). To preview our findings, we show that not only the information-theoretic measures in the three grain sizes indeed impact reading, they do so above and beyond existing consistency measures.

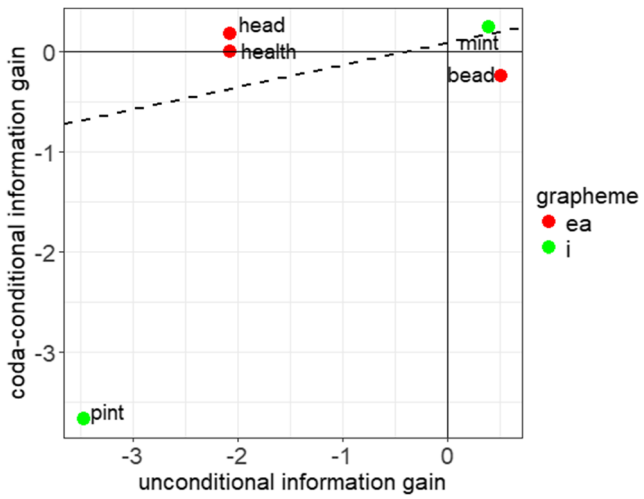


Fig. 3 Examples for information gain in the real-world examples. The *x*-axis shows unconditional values and the *y*-axis shows coda-conditional values. *Points in red* are words with the grapheme *ea*; *points in green* with *i*. The *dashed line* shows the overall linear relation between coda-conditional and unconditional values based on all mono-syllabic words in the corpus

Information-theoretic measures as predictors of behavior

As a starting point, we examined whether surprisal in orthography-to-phonology mapping accounts for cross-item variance in RT and accuracy. The basic prediction here is that words with more surprising (i.e., unpredictable) mapping will be characterized by longer RTs and higher error rates. To examine this prediction, we ran six multiple regression models. Each model included surprisal in one grain size (unconditional, coda-conditional, or onset-conditional) as a predictor. Three models were run with mean word accuracy as a dependent variable, and the remaining three models had mean word log-transformed RT as a dependent variable. All models also included word length (in graphemes) and log-transformed word frequency (estimated from the HAL corpus, which includes 131M English tokens, Burgess & Livesay, 1998), as control variables. We also included dummy-coded variables to control for the place and manner of articulation of the first consonant in a word (see Yap & Balota, 2009).

As can be seen in Table 2, surprisal indeed accounts for variance in behavioral outcomes. As predicted, higher surprisal is associated with longer RTs and lower accuracy rates. Significant effects were observed in all three grain sizes: unconditional, coda-conditional, and onset-conditional (even though conditional values had generally smaller effects in comparison to unconditional values).

Table 2 Effect of surprisal on word naming RT and accuracy in ELP data

Model	DV	Predictor	β	SE	*t*	*p*	R^2 (%)[a]
1	*Acc.*	*Unconditional surprisal*	– 0.012	0.001	– 14.533	< .001	3.2
		Log Freq.	0.013	0.001	28.270	< .001	12.0
		Word Length	0.010	0.001	9.058	< .001	1.2
		Articulation[b]					0.6
2	*Log-RT*	*Unconditional surprisal*	0.011	0.001	9.896	< .001	1.2
		Log Freq.	– 0.017	0.001	– 27.476	< .001	9.6
		Word Length	0.0024	0.002	1.528	.13	0.03
		Articulation[b]					14.6
3	*Acc.*	*Coda-conditional surprisal*	– 0.012	0.002	– 7.539	< .001	0.9
		Log Freq.	0.012	0.001	26.916	< .001	11.1
		Word Length	0.009	0.001	7.628	< .001	0.9
		Articulation[b]					0.6
4	*Log-RT*	*Coda-conditional surprisal*	0.014	0.002	6.509	< .001	0.6
		Log Freq.	– 0.016	0.001	26.817	< .001	9.2
		Word Length	0.004	0.002	2.515	.02	0.1
		Articulation[b]					14.2
5	*Acc.*	*Onset-conditional surprisal*	– 0.016	0.001	– 12.910	< .001	2.5
		Log Freq.	0.013	0.001	27.878	< .001	11.8
		Word Length	0.011	0.001	9.233	< .001	1.3
		Articulation[b]					0.7
6	*Log-RT*	*Onset-conditional surprisal*	0.009	0.007	5.696	< .001	0.4
		Log Freq.	– 0.016	0.001	– 26.790	< .001	9.2
		Word Length	0.0025	0.002	1.571	.12	0.03
		Articulation[b]					14.5

Note: Acc. = Accuracy; DV = dependent variable; Log-RT = Log-transformed response time; Log Freq. = Log-transformed frequency; SE = standard error

[a] R^2 values for each predictor are the difference between the R^2 of a model without this predictor and that of a full model that includes it

[b] Articulation parameters do not have estimates, SE, and *t*/*p*-values because these are a set of 15 dummy-coded variables (reflecting manner/place of articulation of first consonant)

Importantly, while surprisal does account for some variance in behavior, we hypothesized that it would only have limited explanatory power. This is due to the fact that surprisal only takes into account the probability of a given grapheme-to-phoneme correspondence, ignoring the uncertainty implicated by the grapheme overall, across all of its possible realizations. In fact, an additional set of six regression models revealed significant effects of entropy – unconditional, coda-conditional, or onset-conditional, on naming behavior (in five out of six models; see Table 3).

So far, analyses revealed that both the uncertainty in the pronunciation of a grapheme (entropy, Table 3 above) as well as the actual unpredictability in its actual reading (surprisal, Table 2) have an effect on reading behavior. We next focus on the *information gain* of a word – the difference between entropy and surprisal, and its impact on reading outcomes. Because information gain quantifies the distance between the expected and observed information, it potentially serves as a single measure that can account for the surprisal of a grapheme-phoneme event, as well as the uncertainty implicated by a grapheme.

To examine the predictive value of information gain, we ran an additional six multiple regression models (three on log-transformed RT, three on accuracy) to examine the effect of information gain on naming. Each model included one of the information gain measures: unconditional information gain, coda-conditional information gain, and onset-conditional information gain. Similar to the models above, all models also included word length, log-transformed word frequency and dummy-coded articulation variables as control variables. As can be seen in Table 4, in all six models, there was a highly significant effect of the information gain on the dependent variable. This suggests that information gain indeed impacts naming behavior, where high information gain (in each of the three grain sizes) is predictive of faster response latencies and higher accuracy. These results are also visually depicted in Fig. 4, which shows the raw effect of information gain measures on mean RT and accuracy.

Table 3 Effect of entropy on word naming RT and accuracy in ELP data

Model	DV	Predictor	β	SE	*t*	*p*	R^2 (%)[a]
1	*Acc.*	*Unconditional Entropy*	– 0.012	0.0021	– 5.804	< .001	0.5
		Log Freq.	0.012	0.0005	26.479	< .001	10.9
		Word Length	0.010	0.0012	8.775	< .001	1.2
		Articulation[b]					0.6
2	*Log-RT*	*Unconditional Entropy*	0.001	0.0029	0.440	.661	< 0.01
		Log Freq.	– 0.016	0.0006	– 26.242	< .001	8.8
		Word Length	0.003	0.0016	1.899	.058	0.1
		Articulation[b]					14.2
3	*Acc.*	*Coda-conditional Entropy*	– 0.014	0.0026	– 5.344	< .001	0.4
		Log-freq	0.012	0.0005	26.675	< .001	11.0
		Word length	0.009	0.0012	7.414	< .001	0.9
		Articulation[b]					0.6
4	*Log-RT*	*Coda-Conditional Entropy*	0.014	0.0035	4.097	< .001	0.2
		Log Freq.	– 0.016	0.0006	– 26.554	< .001	9.0
		Word Length	0.004	0.0016	2.535	.011	0.1
		Articulation[b]					14.2
5	*Acc.*	*Onset-conditional Entropy*	– 0.013	0.0021	– 6.374	< .001	0.6
		Log Freq.	0.012	0.0005	26.744	< .001	11.1
		Word Length	0.010	0.0012	8.872	< .001	1.2
		Articulation[b]					0.7
6	*Log-RT*	*Onset-conditional Entropy*	0.006	0.0028	2.145	.032	< 0.1
		Log Freq.	– 0.016	0.0006	– 25.987	< .001	8.7
		Word Length	0.003	0.0016	2.138	.032	0.1
		Articulation[b]					13.6

Note: Acc. = Accuracy; DV = dependent variable; Log-RT = Log-transformed response time; Log Freq. = Log-transformed frequency; SE = standard error.

[a] R^2 values for each predictor are the difference between the R^2 of a model without this predictor and that of a full model that includes it.

[b] Articulation parameters do not have estimates, SE, and *t*/*p*-values because these are a set of 15 dummy-coded variables (reflecting manner/place of articulation of first consonant).

Comparison of information-theoretic metrics and other consistency measures

Next, we asked how our information-theoretic metrics compare to standard consistency measures. In this section we thus investigate the correlations between information-theoretic measures and standard consistency measures, and examine whether information-theoretic measures add to the predictive power of typical measures. Given the results above, we examine both surprisal and information gain in the three grain sizes. It should be noted that various possible consistency measures exist. For simplicity, we focus here on two common measures. The first, *vowel consistency*, examines the number of pronunciations of a given vowel grapheme in the same manner as in a given word, across all other words in the corpus (i.e., the vowel consistency value of the word *head* is the ratio of *ea* pronounced as /ɛ/ out of all words in the corpus with the vowel grapheme *ea*; Chateau & Jared, 2003; Treiman et al., 1995). The second, *body-rime consistency*, examines consistency at the body-level, looking at the number of similar pronunciations of a given body in the same manner as in a given word (e.g., the body-rime consistency of the word *head* is the percent of *ead* ➔/ɛd/; e.g., Cortese & Simpson, 2000; Jared, McRae, & Seidenberg, 1990; Ziegler, Stone, & Jacobs, 1997b). Note that we calculated vowel consistency and body-rime consistency based on types rather than on tokens (i.e. frequency did not play a role as both frequent and infrequent words were counted once in all calculations). This was done to provide a closer parallel to our information-theoretic measures which were calculated on types, too.

The correlations between surprisal and information gain at the three grain sizes and the two existing consistency measures are shown in Table 5. As can be seen, the various measures, including the information-theoretic measures and typical consistency measures, are generally correlated with each other. Particularly high correlations are observed between surprisal values and consistency measures at the same grain size (i.e., unconditional surprisal and vowel consistency, r = –

Table 4 Effect of information gain (IG) on word naming RT and accuracy in ELP data

Model	DV	Predictor	β	SE	*t*	*p*	R^2 (%)[a]
1	*Acc.*	*Unconditional IG*	0.012	0.001	13.392	< .001	2.7
		Log Freq.	0.013	0.001	28.028	< .001	11.9
		Word Length	0.010	0.001	8.572	< .001	1.1
		Articulation[b]					0.6
2	*Log-RT*	*Unconditional IG*	– 0.012	0.001	-10.613	< .001	1.4
		Log Freq.	– 0.017	0.001	-27.573	< .001	9.6
		Word Length	0.003	0.002	1.834	.067	< 0.1
		Articulation[b]					14.5
3	*Acc.*	*Coda-conditional IG*	0.011	0.002	5.301	< .001	0.4
		Log-freq	0.012	0.001	26.462	< .001	10.8
		Word length	0.010	0.001	8.329	< .001	1.1
4	*Log-RT*	*Coda-Conditional IG*	– 0.013	0.003	– 4.978	< .001	0.3
		Log Freq.	– 0.016	0.001	– 26.468	< .001	8.9
		Word Length	0.003	0.002	1.922	.055	0.1
		Articulation[b]					14.3
5	*Acc.*	*Onset-conditional IG*	0.017	0.002	11.323	< .001	2.0
		Log Freq.	0.012	0.001	27.333	< .001	11.4
		Word Length	0.099	0.001	8.528	< .001	1.1
		Articulation[b]					0.6
6	*Log-RT*	*Onset-conditional IG*	– 0.018	0.002	– 8.654	< .001	1.0
		Log Freq.	– 0.016	0.001	– 27.024	< .001	9.2
		Word Length	0.003	0.002	1.838	.066	< 0.1
		Articulation[b]					14.1

Note: Acc. = Accuracy; DV = dependent variable; Log-RT = Log-transformed response time; Log Freq. = Log-transformed frequency; SE = standard error

[a] R^2 values for each predictor are the difference between the R^2 of a model without this predictor and that of a full model that includes it

[b] Articulation parameters do not have estimates, SE, and *t*/*p*-values because these are a set of 15 dummy-coded variables (reflecting manner/place of articulation of first consonant)

0.86; coda-conditional surprisal and body-rime consistency, $r = -0.9$). Correlations are lower between vowel/body-rime consistency and information gain measures, but still positive, and sometimes high (*r* ranging from 0.2 to 0.72 across different grain sizes). This is again expected: on the one hand information gain measures tap into parts that are not captured by typical measures (specifically, grapheme entropy), but they are also affected by surprisal, which is tightly related to consistency measures. Overall, though, in many cases correlations are far from perfect (note in particular the low correlations between measures focusing on different grain sizes, e.g., body-rime consistency and onset-conditional information gain). This led us to examine how information-theoretic measures fare in relation to typical consistency measures in terms of their predictive power. Specifically, we asked whether information-theoretic measures would have an effect on naming performance beyond that of standard consistency measures.

Our analytic strategy for this comparison was as follows. First, we ran two benchmark models – one with (log-transformed) RT as a DV, and another with accuracy. These two base models had similar predictors: the various control variables (log-frequency, length, and articulation predictors) as well as the two standard consistency measures (vowel and body-rime consistency). Then, we ran models that in addition to these predictors included information-theoretic measures: either surprisal or information gain measures. We then conducted F-tests comparing the fit of the full model (including the surprisal/information metrics) to that of the base model (using the *anova()* function in R).

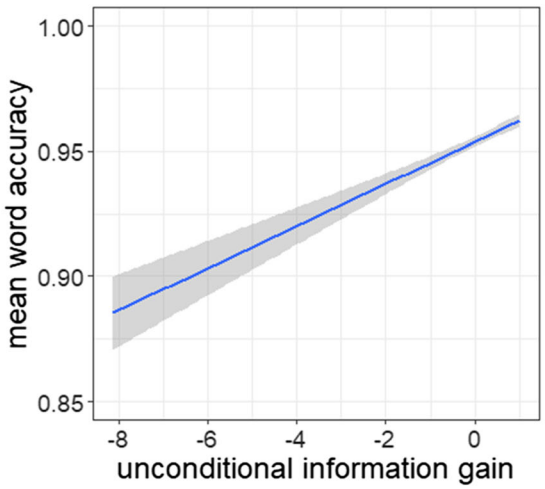
Table 6 presents the result of this comparison. As can be seen, all models that included the information-theoretic values were characterized by a significant improvement in model fit compared to the base models. This is true for both surprisal and information-gain measures. It is important to note that this

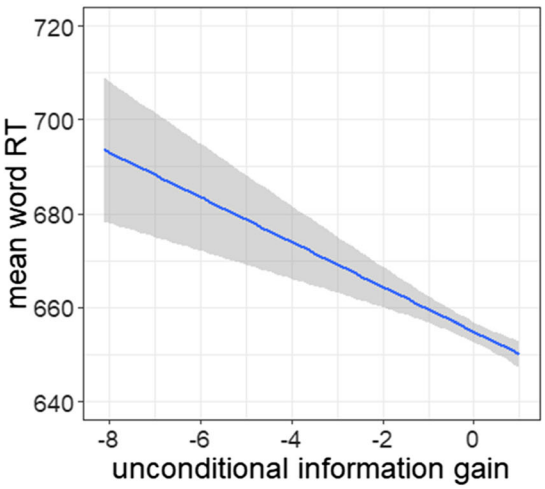
Fig. 4 Effects of information gain on mean accuracy (*left*) and RT (*right*) in naming data. The *top row* shows effects of unconditional information gain, the *middle row* of coda-conditional information gain, and the *bottom row* of onset-conditional information gain

Accuracy

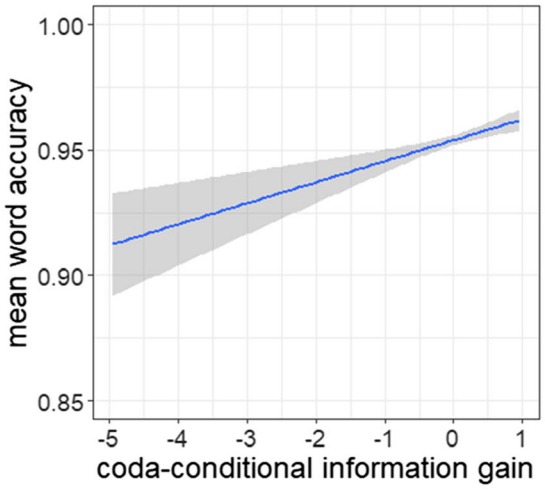
Reaction Time

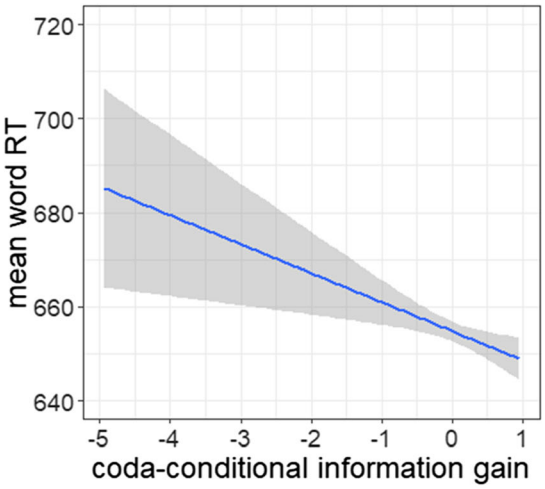
Unconditional IG:



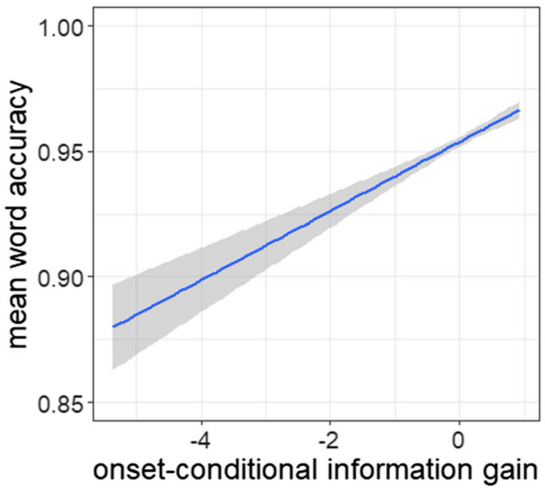


Coda-conditional IG:





Onset-conditional IG:



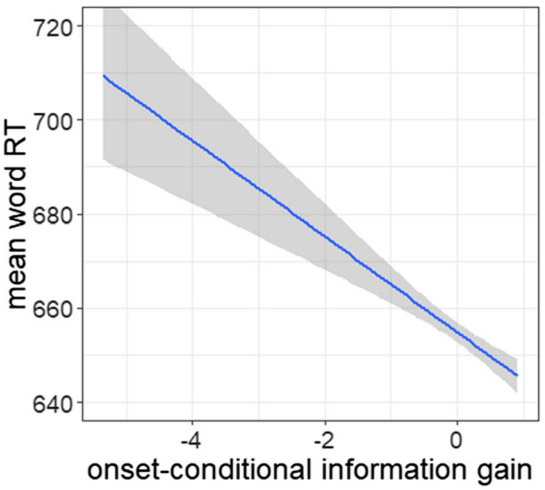


Table 5 Correlations between information-theoretic measures and typical consistency measure

Measure	2	3	4	5	6	7	8
1. Unconditional surprisal	.63	.76	.93	.47	.62	.86	.60
2. Coda-conditional surprisal	—	.29	.6	.79	.19	.55	.90
3. Onset-conditional surprisal		—	.69	.17	.81	.66	.30
4. Unconditional Information gain			—	.51	.67	.72	.53
5. Coda-conditional information gain				—	.1	.37	.56
6. Onset-conditional information gain					—	.49	.20
7. Vowel consistency						—	.55
8. Body-rime consistency							—

Note: For readability, values appear here in absolute values. Actual correlations between surprisal measures and information gain/consistency measures are negative due to opposite scales

improvement does not stem only from the addition of onset-related information (which is not accounted for by vowel/body-rime consistency measures): a significant improvement was observed also when comparing the base models to models with only unconditional and coda-conditional information-theoretic metrics.

Interestingly, we found that not only information-theoretic measures account for variance beyond typical measures, but also that the consistency measures have predictive value beyond that of the information-theoretic measures. This was revealed by parallel analysis examining the improvement in fit when adding typical consistency measures (vowel and body-rime consistency) to base models that include the information theoretic measures (surprisal or information gain in the three grain sizes; see Table 7). Together, the results of these two sets of models suggest that typical consistency measures and our information-theoretic measures have at least some unique (non-overlapping) predictive value. We return to this point in the General Discussion.

Additional investigations

Effects of orthography-to-phonology transparency on lexical decision

The effects of the information-theoretic measures on behavior are not confined to tasks where reading aloud is required. Table 8 shows the results of 6 regression models, examining the impact of information gain on accuracy and log-transformed RT in the lexical decision portion of the ELP[4]. As can be seen, information gain still has significant effects on lexical decision response latencies and errors (while controlling for word frequency and length) in five out of six models, albeit with smaller effect sizes compared to the naming data.

Orthography-to-phonology versus phonology-to-orthography

An additional question is whether naming behavior is affected only by the consistency in the mapping between orthography and phonology, or also by that in the other direction – between phonology-to-orthography (e.g., the mapping between the sound /ɛ/ and the vowel grapheme *ea*; see, e.g., Lacruz & Folk, 2004; Ziegler, Petrova, & Ferrand, 2008). As noted above, one of the advantages of the information-theoretic framework is in its flexibility in capturing information across different types of mappings. We thus repeated the estimation of the information-theoretic measures described above, but this time using probabilities of mappings from phonology-to-orthography. As might be expected, positive correlations between orthography-to-phonology and phonology-to-orthography information gain were observed in each of the three grain sizes: unconditional: $r = 0.6$; coda-conditional: $r = 0.46$; onset-conditional: $r = 0.37$: That is, words that are more transparent in one mapping direction are also generally more transparent in the other direction.

We next examined the predictive value of the phonology-to-orthography measures on naming performance. Our analytic strategy here was similar to that above comparing the information-theoretic metrics to standard consistency measures (Tables 6 and 7). We first ran six baseline models (with either accuracy or log-RT as DV) including log-frequency, word length, and the articulation parameters, as well as an orthography-to-phonology information gain measure in one grain size (unconditional, coda-conditional, or onset-conditional). We then compared these baseline models to models that included all of these parameters, plus an information gain measure of phonology-to-orthography (in the same grain size). This examined whether the structure of the phonology-to-orthography mapping explains naming behavior above and beyond orthography-to-phonology transparency. The results are presented in Table 9. In all six model comparisons, we found a significant improvement in model fit when adding the phonology to orthography information gain measure. This suggests that indeed the mapping between phonology-to-orthography accounts for additional variance that is not accounted for by orthography to phonology alone.

General discussion

In this paper, we propose a new framework for assessing the transparency of the mapping between orthography to

[4] The focus here on information gain measures, and not surprisal, is simply due to conciseness considerations. Further analyses revealed that running the same models reported in this section with surprisal values instead of information gain showed a similar pattern of results with significant effects in all six models. This is true also for the next section on phonology-to-orthography vs. orthography-to-phonology effects.

Table 6 Added predictive value of information-theoretic measures over typical consistency measures for word naming accuracy and RT

DV	Model with information-theoretic predictors	F^a	p	ΔR^2 (%)
Accuracy	Base model + All surprisal measures	32.793	< .001	1.5
	Base model + Unconditional surprisal + Coda-conditional surprisal	46.473	< .001	1.4
Log-RT	Base model + All surprisal measures	25.399	< .001	1.0
	Base model + Unconditional surprisal + Coda-conditional surprisal	34.003	< .001	0.9
Accuracy	Base model + All IG measures	19.012	< .001	0.9
	Base model + Unconditional IG + Coda-conditional IG	22.837	< .001	0.7
Log-RT	Base model + All IG measures	20.799	< .001	0.8
	Base model + Unconditional IG + Coda-conditional IG	26.964	< .001	0.7

Note: DV: dependent variable; IG: information gain; RT: reaction time. Base model predictors are vowel consistency, body-rime consistency, log-transformed frequency, length, and the set of 15 dichotomous articulation characteristics.

[a] *df* for all models with three additional predictors compared to base models are 3, 5693 and *df* for all models with two additional predictors are 2, 5694

phonology. Our approach is rooted in the mathematical tools provided by Information Theory, specifically relying on three information-theoretical notions. First, we assessed the *surprisal* of a grapheme-to-phoneme correspondence, as a measure for the extent of unpredictability in a mapping. Second, we calculated the *entropy* of a grapheme, to quantify the uncertainty implicated by it. Last, we looked at the *information gain* of a grapheme-to-phoneme correspondence, quantifying the difference between the expected amount of information in a grapheme and the information provided by the actual grapheme-to-phoneme mapping (i.e., the difference between entropy and surprisal). To reiterate, by assessing both surprisal and entropy our approach accounts for the unpredictability of the mapping between a given grapheme and phoneme as well as the overall uncertainty implicated by a grapheme. We first demonstrated how these measures can be used to quantify different aspects of the English writing system and its transparency. Then, we showed that these measures are related to reading behavior, accounting for inter-item variability in naming and word recognition more broadly. Throughout these analyses, the flexibility of the current approach was demonstrated, by examining transparency in different grain sizes (i.e., context independent, coda-dependent and onset dependent), and by quantifying the amount of information not only

Table 7 Added predictive value of typical measures to information-theoretic measures for word naming accuracy and RT

DV	Predictors in base model	Model with information-theoretic predictors	F^a	p	ΔR^2 (%)
Accuracy	Controls, surprisal measures	Base model + Vowel consistency, Body-rime consistency	20.354	< .001	0.7
Log-RT	Controls, surprisal measures	Base model + Vowel consistency, Body-rime consistency	27.817	< .001	0.7
Accuracy	Controls, information gain measures	Base model + Vowel consistency, Body-rime consistency	14.731	< .001	0.4
Log-RT	Controls, information gain measures	Base model + Vowel consistency, Body-rime consistency	15.440	< .001	0.4

Note: DV: dependent variable; RT: reaction time. Control predictors are log-transformed frequency, length, and the set of 15 dichotomous articulation characteristics

[a] *df* for all models with two additional predictors are 2, 5694

Table 8 Effect of information gain (IG) on mean word RT and accuracy in ELP lexical decision

Model	DV	Predictor	β	SE	*t*	*p*	R^2 (%)[a]
1	*Acc.*	*Unconditional IG*	0.011	0.0017	6.303	< 0.0001	0.5%
		Log Freq.	0.045	0.0009	48.807	< 0.0001	29.2%
		Word length	0.049	0.0022	22.446	< 0.0001	6.2%
2	*Log-RT*	*Unconditional IG*	– 0.010	0.0016	– 6.189	< 0.0001	0.4%
		Log Freq.	– 0.036	0.0007	– 53.602	< 0.0001	33.0%
		Word length	– 0.010	0.0016	– 5.918	< 0.0001	0.4%
3	*Acc.*	*Coda-conditional IG*	0.004	0.0039	0.976	0.329	0.01%
		Log Freq.	0.044	0.0009	48.233	< 0.0001	28.7%
		Word length	0.049	0.0022	22.374	< 0.0001	6.2%
4	*Log-RT*	*Coda-conditional IG*	– 0.009	0.0029	– 3.170	0.002	0.1%
		Log Freq.	– 0.036	0.0007	– 53.202	< 0.0001	32.7%
		Word length	– 0.010	0.0016	– 5.902	< 0.0001	0.4%
5	*Acc.*	*Onset-conditional IG*	0.020	0.0031	6.442	< 0.0001	0.5%
		Log Freq.	0.044	0.0009	48.768	< 0.0001	29.2%
		Word length	0.050	0.0022	22.593	< 0.0001	6.3%
6	*Log-RT*	*Onset-conditional IG*	– 0.011	0.0023	– 4.785	< 0.0001	0.3%
		Log Freq.	– 0.036	0.0007	– 53.386	< 0.0001	32.8%
		Word length	– 0.011	0.0016	– 6.020	< 0.0001	0.4%

Note: Acc. = Accuracy; DV = dependent variable; Log-RT = Log-transformed response time; Log. Freq. = Log-transformed frequency; SE = standard error

[a] R^2 values for each predictor are the difference between the R^2 of a model without this predictor and that of a full model that includes it

in the mapping of orthography-to-phonology, but also in the opposite direction from phonology-to-orthography.

The motivation behind this study was to offer a comprehensive account of print-speech transparency, which can be flexibly used to capture regularities in different mappings and considering both the predictability of a grapheme-phoneme correspondence as well as the overall uncertainty (entropy) regarding a full distribution of possible pronunciations. Indeed, we demonstrated that our approach diverges from standard consistency measures. Interestingly, the analyses revealed that the information-theoretical measures and the standard measures of consistency – vowel consistency and body-rime consistency – account for non-overlapping variance in naming and lexical decision behavior (see Tables 6 and 7

Table 9 Added predictive value of phonology-to-orthography measures to orthography-to-phonology measures

DV	Predictors in base model	Model with information-theoretic predictors	F^a	*p*	ΔR^2 (%)
Accuracy	Controls, O2P unconditional IG	Base model + P2O unconditional IG	20.448	< 0.001	0.4%
Log-RT	Controls, O2P unconditional IG	Base model + P2O unconditional IG	27.366	< 0.001	0.5%
Accuracy	Controls, O2P coda-conditional IG	Base model + P2O coda-conditional IG	20.448	< 0.001	0.4%
Log-RT	Controls, O2P coda-conditional IG	Base model + P2O coda-conditional IG	31.008	< 0.001	0.5%
Accuracy	Controls, O2P onset-conditional IG	Base model + P2O onset-conditional IG	4.687	0.030	0.1%
Log-RT	Controls, O2P onset-conditional IG	Base model + P2O onset-conditional IG	14.218	< 0.001	0.2%

Note: DV = dependent variable; Log-RT = Log-transformed response time; IG: information gain; O2P: orthography to phonology; P2O: phonology to orthography; RT: reaction time; SE: standard error. Control predictors are log-transformed frequency, length, and the set of 15 dichotomous articulation characteristics

[a] *df* for all models are 1, 4240

above). To understand the reasons behind this finding, it is important to underline the similarities and differences between the information-theoretical and standard measures. One inherent difference between information-theoretical and consistency measures is that the former set of measures capture an aspect that the later by-definition do not: the overall uncertainty over the possible pronunciations of a grapheme, assessed by entropy. For example, standard measures would only consider the probability of *i* ➔ /ɪ/ to assess the consistency of the word *mint*, whereas entropy (and information gain) considers the predictability of this correspondence as well as the full distribution of possible phonemes given *i*. Thus, graphemes with many possible pronunciations are generally characterized by higher entropy, even when the probability of many of these pronunciations are low and might not affect standard measures of consistency. More broadly, and as exemplified above, two grapheme-phoneme correspondences with an equal probability might still diverge in their entropy. It seems that the effect of entropy on behavior (see Table 3, above) leads to at least some of the added predictive value of the information gain measures (which are a function of both entropy and surprisal) beyond standard consistency measures.

Moreover, taking entropy (and not only predictability) into account leads also to differences between the approaches in how they assess the transparency of words with very rare vowel graphemes or bodies ("hermit words"). Consider as an example the word *laugh*. Given the rareness of the body *–augh,* which appears only in this word, the body-rime consistency of this word is perfect – it has one friend (the word itself) and no enemies, resulting in a body-rime consistency of 1. This will also be reflected in the surprisal estimate of the word, which will be minimal because p(i) = 1, log(p(i)) = 0, hence no surprisal, or maximal predictability. In contrast, the information gain of this word will not be set to the maximum level of transparency. This is because in hermit words the entropy of the vowel grapheme or body is also zero (the entropy of a distribution with only one option with a probability of 1), and since surprisal is equal to zero the information gain is zero, too. In other words, while consistency (or surprisal) measures consider hermit words as fully predictable, information gain measures assess them as neutral in terms of their transparency.

In addition to these differences between *information gain* and consistency measures, two differences between *surprisal* and consistency measures should also be noted. These measures are highly similar as both are calculated based on the probability of some grapheme-to-phoneme correspondence at a given grain size. Thus, at the vowel level, vowel consistency and unconditional surprisal are both calculated according to the probability of a given grapheme-to-phoneme correspondence. The only difference between these two measures is the use of a different scale: Consistency is calculated over raw probabilities, while surprisal uses log-transformed values. The results above raise the possibility that the psychological effect of grapheme-to-phoneme correspondences may be linear as well as log-linear, or, alternatively, that it follows a different non-linear function that resembles some combination of linear and log-linear effects. This issue can be addressed by future research, potentially by using subtler statistical analysis (e.g., random forest analysis, Matsuki, Kuperman, & Van Dyke, 2016) combined with designs with stimuli in values that are maximally informative in distinguishing between linear and log-linear effects, which fall at different regions on the probability distribution (and see, e.g., Smith & Levy, 2013, for a related discussion in the context of word predictability).

Last, there is also another, subtler, difference between body-rime consistency measure and its information-theoretic parallel, coda-dependent surprisal. Our coda-dependent measure (as well as the onset-related measure) examines the transparency of the vowel grapheme given the coda as a constraining context. In contrast, body-rime consistency concerns the transparency of the mapping of the full unit, the coda, and is therefore also affected by the pronunciation of the consonants. As a result, a word like *lease* would be considered inconsistent using standard body-rime consistency (as *lease* does not rhyme with words such as *please* or *tease*), whereas coda-dependent surprisal would consider *lease* transparent because the vowel *ea* is pronounced as /i/, as expected given the coda *-se.* Because inconsistency in consonants is relatively rare in English, this difference is limited to a relatively small number of words but may nevertheless be the source of some of the differences between these measures. Note also that in other languages, which have substantial degree of irregularity in the pronunciation of consonants, accounting for this difference is even more important. This seemingly small difference points to a fundamental theoretical question: What are the basic units of the mapping between orthography to phonology — vowels or codas/onsets. At a practical level, the question is what orthographic units should serve as the basis for the calculation of the different information-theoretic measures (e.g., whether to estimate the body-level transparency of the word *mint* by examining the surprisal/entropy of the grapheme *i* given the coda *–nt*, or of the body *–int* as a whole).

One other point that deserves emphasis is that all of our analyses examined whether differences in print-speech transparency are related to inter-item variability in reading (i.e., whether words that are more transparent are read more easily). To do so, our analyses focused on aggregated data from a megastudy (the ELP), examining mean response latencies and accuracy of a large number of words, aggregated across a sample of highly skilled readers. As such, our results do not speak to two important issues. First, it is still an open question how the current results would generalize to in lab studies: that is, whether the information-theoretic measures can capture significant variance also in smaller-scale studies (see Balota, Yap, Hutchison, & Cortese, 2012 for discussion of the

differences between megastudies and more traditional factorial experiments). A second issue is that our analyses do not speak to the question of how does reliance on orthography-to-phonology account for individual-differences in reading (cf., e.g., Steacy et al., 2018). In general, analyses conducted using data from mega-studies, such as the ELP, only concerns the average behavior of the skilled reader: In the current case, demonstrating that on average skilled readers recognize faster and more accurately words with more transparent O-P mapping. These findings, however, overlook the extensive inter-individual variability in reading behavior, and does not necessarily mean that the same measures can be used to account for meaningful variability across individuals (see Andrews, 2012 for discussion). To address this question, future studies should examine the degree to which each individual is impacted by letter to sound transparency and examine its correlation with overall reading proficiency skills. The current information-theoretical metrics can be useful to this aim because they can quantify different aspects of transparency (e.g., surprisal vs. entropy) at different grain sizes (e.g., context dependent/independent) to which individuals may be differentially sensitive. In this context, it is also important to note that the extent to which the reliance on orthography-to-phonology mapping accounts for individual differences in reading proficiency may be less pronounced in highly proficient readers (the subjects in the ELP dataset used here), who are likely to be very efficient in computing phonology and who therefore may show decreased consistency effects in word reading compared to younger readers (Sprenger-Charolles, Siegel, Béchennec, & Serniclaes, 2003; but see, Ziegler, Bertrand, Lété, & Grainger, 2014). Future studies can use the measures provided here to examine these associations across development, accounting for the different aspects of the information distribution between orthography and phonology.

Another general finding stemming from the current investigation is the significant effect of phonology-to-orthography transparency on visual word recognition. Whether such feedback effects indeed exist was a matter of much debate in the reading literature for over two decades (see, e.g., Chiarello, Vaden, & Eckert, 2018; Lacruz & Folk, 2004; Lee, Hsu, Chang, Chen, & Chao, 2015; Perry, 2003; Stone, Vanhoy, & Van Orden, 1997; Ziegler, Montant, & Jacobs, 1997a; Ziegler et al., 2008). As noted above, one of the advantages of the current approach is its flexibility to easily capture the information distribution of mappings in various directions. Our information-theoretic approach, combined with the use of large databases, revealed evidence for the existence of phonology-to-orthography feedback effect on word naming performance (cf. Kessler, Treiman, & Mullennix, 2008, who did not find a feedback consistency effect in an analysis of a large-scale naming database using standard consistency measures). Nonetheless, note that these effects were somewhat weaker than the feedforward effects of orthography-to-phonology, as reflected in their smaller effect sizes.

We also wish to comment on some other open questions that are beyond the scope of the current study. The first is the expansion of the current approach to multi-syllabic words. As a first step, we opted to limit our current investigation to mono-syllabic words. This was done because in such words the definition of conditioning contexts that are relevant for the mapping of vowel graphemes to phonology is relatively straight-forward, as mono-syllabic words are built from a structure of onset-vowel-coda. Defining the relevant conditioning contexts in multisyllabic words is a task for future studies (and see Chateau & Jared, 2003; Yap & Balota, 2009 for extended discussion and promising avenues). Moreover, assessing the transparency of multi-syllabic words requires a theory regarding how to combine multiple information values to a single value (e.g., what is the overall information value of a word with one vowel that incur a high amount of information and another with a low amount of information, such as *island*?). Second, and as noted above, our investigation focused only on the pronunciation of vowels. This was due to the structure of English, where the majority of orthography-to-phonology opaqueness is concentrated in vowels. It is important to note however that the information-theoretical approach is not limited to vowels (or to consonants) and can be easily extended to the full mapping of graphemes to phonemes in a writing system. Third, here we only examined a single writing system – English. This was done due to practical reasons, namely, the availability of a corpus of English written words and their pronunciations. Future cross-linguistic studies are left with the task of comparing the information distributions of orthography-to-phonology of different writing systems. Indeed, previous studies used some information-theoretical notions to compare orthographic depth across languages, looking at entropy of restricted parts of the input as a basis for their calculation (e.g., letter-to-sound and sound-to-letter mappings of word onsets, Borgwaldt et al., 2004, 2005; Ziegler et al., 2010). The metrics presented here can help extend these studies and cross-linguistically compare orthography-to-phonology transparency in a more comprehensive manner, taking into account different aspects of the written input.

On a broader note, it is important to note that one of the unique advantages of the information-theoretic approach is that it is not limited to assessing regularities only in the context of print-speech correspondences. Needless to say, print-speech correspondences are only one source of information among multiple additional cues that readers can rely on in the task of visual word recognition (e.g., the consistency of the

mapping of orthography-to-semantic, Marelli & Amenta, 2018)[5], cues that are joined by additional sources of information in continuous text reading (in particular word predictability, Staub, 2015). As reviewed in the introduction, information-theoretic measures are used throughout all domains of spoken and written language, including among others morphological, syntactic, and orthographic regularities (and see Mollica & Piantadosi, 2019 for an estimation of the total linguistic information held by language users across all domains). This presents an opportunity to estimate the impact of the information distribution across different levels of the written input on reading, their relative weighing, and their possible interactions. The current paper shows that indeed print-speech correspondences, and their impact on single word recognition, can be captured using information-theoretic measures. This can serve as a springboard for addressing theoretical questions regarding how the degree to which print-speech correspondences, along with other regularities that are present in the written input (regularities between print and morphology, e.g., Milin et al., 2009; semantics, e.g., Montemurro, 2014; syntax, e.g., Linzen & Jaeger, 2015, etc.) impact reading, and whether they interact in any ways (e.g., whether the impact of print-speech regularities is more pronounced when other regularities are absent, such as in orthographically unpredictable words). We believe that the generalizable tools of information-theory can prove to be extremely useful in pursuing these critical questions.

To conclude, the current work provides a powerful and flexible set of tools to quantify the information provided by the input that readers are exposed to, based on the regularities between words' orthographic and phonological forms. It joins efforts across various sub-domains of language research that use related measures to capture the information distribution of different elements in the written and spoken input. We believe that exact quantification of the information present in the linguistic input is a critical step in understanding how linguistic phenomena, including literacy acquisition and reading, are shaped by the regularities that are available to learners and users of the language.

Acknowledgements This work was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Awards P01HD070837, P20HD091013, P01HD001994, and 5R37HD090153-02. Noam Siegelman is a Rothschild Yad-Hanadiv post-doctoral fellow. We wish to thank Mark van den Bunt for his helpful comments.

Supplemental Material The full list of the 10,093 monosyllabic words, their GPC coding, and their information-theoretic measures of orthographic-phonological regularities (entropy, surprisal and information gain; unconditional and conditional) is available at: https://osf.io/kfme8/

Open practices statement The full list of words, their GPC coding, and their information-theoretic measures is available as Supplementary Material https://osf.io/kfme8/. The full phonological corpus is also available at https://phinder.devinkearns.org.

[5] As a side note, we examined the predictive power of the information-theoretic measures of orthography-to-phonology jointly with those reported by Marelli and Amenta (2018) that quantifies orthography-to-semantic consistency. To do so, we repeated our basic analyses, presented in Tables 2 to 4, but this time adding to the regression models the orthography-to-semantics measures. The predictive power of the orthography-to-phonology measures did not change by this addition, with the information-theoretic measures still predicting accuracy and RT in naming. In parallel, as reported by Marelli and Amenta (2018), the measures of orthography-to-semantic consistency still accounted for significant variance in naming response latencies. This suggests that the two measures capture non-overlapping variance, and that the two sources of information are available to readers in parallel.

References

Andrews, S. (2012). Individual differences in skilled visual word recogindividual differences in skilled visual word recognition and reading. In J. S. Adelman (Ed.), *Visual word recognition, volume 2* (pp. 151–172). Hove, UK: Psychology Press.

Arciuli, J. (2018). Reading as statistical learning. *Language, Speech, and Hearing Services in Schools*, *49*(3), 634–643. https://doi.org/10.1044/2018_lshss-stlt1-17-0135

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science*, *36*(2), 286–304. https://doi.org/10.1111/j.1551-6709.2011.01200.x

Aronoff, M., & Koch, E. (1996). Context-sensitive regularities in english vowel spelling. *Reading and Writing*. https://doi.org/10.1007/BF00420278

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1995). *The CELEX Lexical Database*. *Philadelphia Linguistics Data Consortium University of Pennsylvania*. https://doi.org/10.1016/j.bbrc.2013.10.120

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The english lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual Word Recognition Volume 1: Models and Methods, Orthography, and Phonology* (pp. 90–115). Hove, UK: Psychology Press. https://doi.org/10.4324/9780203107010

Baron, J., & Strawson, C. (1976). Use of orthographic and word-specific knowledge in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(3), 386–393. https://doi.org/10.1037/0096-1523.2.3.386

Borgwaldt, S. R., Hellwig, F. M., & de Groot, A. M. B. (2004). Word-initial entropy in five languages: Letter to sound, and sound to letter. *Written Language & Literacy*, *7*(2), 165–184. https://doi.org/10.1075/wll.7.2.03bor

Borgwaldt, S. R., Hellwig, F. M., & De Groot, A. M. B. (2005). Onset entropy matters - Letter-to-phoneme mappings in seven languages. *Reading and Writing*, *18*(3), 211–229. https://doi.org/10.1007/s11145-005-3001-9

Borleffs, E., Maassen, B. A. M., Lyytinen, H., & Zwarts, F. (2017). Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: a narrative review. *Reading and Writing*, *30*(8), 1617–1638. https://doi.org/10.1007/s11145-017-9741-5

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from

Kučera and Francis. *Behavior Research Methods, Instruments, and Computers*, *30*(2), 272–277. https://doi.org/10.3758/BF03200655

Chateau, D., & Jared, D. (2003). Spelling-sound consistency effects in disyllabic word naming. *Journal of Memory and Language*, *48*(2), 255–280. https://doi.org/10.1016/S0749-596X(02)00521-1

Chiarello, C., Vaden, K. I., & Eckert, M. A. (2018). Orthographic influence on spoken word identification: Behavioral and fMRI evidence. *Neuropsychologia*, *111*, 103–111. https://doi.org/10.1016/j.neuropsychologia.2018.01.032

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, *6*(2), 243–278. https://doi.org/10.1515/lp-2015-0008

Cohen Priva, U. (2017). Not so fast: Fast speech correlates with lower lexical and structural information. *Cognition*, *160*, 27–34. https://doi.org/10.1016/j.cognition.2016.12.002

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256. https://doi.org/10.1037/0033-295X.108.1.204

Cortese, M. J., & Simpson, G. B. (2000). Regularity effects in word naming: What are they? *Memory and Cognition*, *28*(8), 1269–1276. https://doi.org/10.3758/BF03211827

Fine, A. B., & Florian Jaeger, T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, *37*(3), 578–591. https://doi.org/10.1111/cogs.12022

Fitt, S. (2001). *Unisyn Lexicon Release (Version 1.3) [Datafile and codebook]*. Edinburgh, Scotland: Centre for Speech Technology Research at the University of Edinburgh.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, *12*(6), 627–635. https://doi.org/10.1016/S0022-5371(73)80042-8

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, *5*(3), 475–494. https://doi.org/10.1111/tops.12025

Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences*, *35*(5), 263–279. https://doi.org/10.1017/S0140525X11001841

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(1), 104–115. https://doi.org/10.1037/0096-1523.13.1.104

Frost, R., Siegelman, N., Narkiss, A., & Afek, L. (2013). What predicts successful literacy acquisition in a second language? *Psychological Science*, *24*(7), 1243–1252. https://doi.org/10.1177/0956797612472207

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(4), 674–691. https://doi.org/10.1037/0096-1523.5.4.674

Graves, W. W., Desai, R., Humphries, C., Seidenberg, M. S., & Binder, J. R. (2010). Neural systems for reading aloud: A multiparametric approach. *Cerebral Cortex*, *20*(8), 1799–1815. https://doi.org/10.1093/cercor/bhp245

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4), 643–672. https://doi.org/10.1207/s15516709cog0000_64

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720. https://doi.org/10.1037/0033-295X.111.3.662

Jared, D., McRae, K., & Seidenberg, M. S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language*, *29*(6), 687–715. https://doi.org/10.1016/0749-596X(90)90044-Z

Kearns, D. M., Rogers, H. J., Koriakin, T., & Al Ghanem, R. (2016). Semantic and phonological ability to adjust recoding: A unique correlate of word reading skill? *Scientific Studies of Reading*, *20*(6), 455–470. https://doi.org/10.1080/10888438.2016.1217865

Kessler, B., & Treiman, R. (2001). Relationships between sounds and letters in English monosyllables. *Journal of Memory and Language*, *44*(4), 592–617. https://doi.org/10.1006/jmla.2000.2745

Kessler, B., Treiman, R., & Mullennix, J. (2008). Feedback-consistency effects in single-word reading. In E. L. Grigorenko & A. J. Naples (Eds.), *Single-Word Reading: Behavioral and Biological Perspectives*. New York, NY: Erlbaum. https://doi.org/10.4324/9780203810064

Lacruz, I., & Folk, J. R. (2004). Feedforward and feedback consistency effects for high- and low-frequency words in lexical decision and naming. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *57*(7), 1261–1284. https://doi.org/10.1080/02724980343000756

Lee, C. Y., Hsu, C. H., Chang, Y. N., Chen, W. F., & Chao, P. C. (2015). The feedback consistency effect in Chinese character recognition: Evidence from a psycholinguistic norm. *Language and Linguistics*, *16*(4), 535–554. https://doi.org/10.1177/1606822X15583238

Linzen, T., & Jaeger, T. F. (2015). Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. *Cognitive Science*, *40*(6), 1382–1411. https://doi.org/10.1111/cogs.12274

Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical Predictability During Natural Reading: Effects of Surprisal and Entropy Reduction. *Cognitive Science*, *42*(4), 1166–1183. https://doi.org/10.1111/cogs.12597

Marelli, M., & Amenta, S. (2018). A database of orthography-semantics consistency (OSC) estimates for 15,017 English words. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1017-8

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading*. https://doi.org/10.1080/10888438.2015.1107073

Milin, P., Kuperman, V., Kostić, A., & Baayen, H. R. (2009). Paradigms bit by bit: An information-theoretic approach to the processing of inflection and derivation. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: Form and acquisition* (pp. 214–252). Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199547548.003.0010

Mollica, F., & Piantadosi, S. T. (2019). Humans store about 1.5 megabytes of information during language acquisition. *Royal Society Open Science*, *6*(3), 181393. https://doi.org/10.1098/rsos.181393

Montemurro, M. A. (2014). Quantifying the information in the long-range order of words: Semantic structures and universal linguistic constraints. *Cortex*, *55*, 5–16. https://doi.org/10.1016/J.CORTEX.2013.08.008

Perfetti, C. A., Beck, I., Bell, L. C., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. *Merrill-Palmer Quarterly*, *33*, 283–319.

Perry, C. (2003). A phoneme-grapheme feedback consistency effect. *Psychonomic Bulletin and Review*, *10*(2), 392–397. https://doi.org/10.3758/BF03196497

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115. https://doi.org/10.1037/0033-295X.103.1.56

Protopapas, A., & Vlahou, E. L. (2009). A comparative quantitative analysis of Greek orthographic transparency. *Behavior Research Methods*, *41*(4), 991–1008. https://doi.org/10.3758/BRM.41.4.991

Rastle, K., & Coltheart, M. (1999). Serial and strategic effects in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(2), 482–503. https://doi.org/10.1037/0096-1523.25.2.482

Rueckl, J. G., Zevin, J. D., & Wolf VII, H. (2019). Using computational techniques to model and better understand developmental word-

reading disorders (i.e., dyslexia). In J. Washington & D. Compton (Eds.), *Dyslexia: Revisiting Etiology, Diagnosis, Treatment, and Policy*. Baltimore, MD: Brookes Publishing Co.

Scarborough, H. S. (1998). Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness, verbal memory, rapid naming, and IQ. *Annals of Dyslexia*, *48*(1), 115–136. https://doi.org/10.1007/s11881-998-0006-5

Seidenberg, M. S. (2011). Reading in different writing system: One architecture, multiple solutions. In *Dyslexia across language: Orthography and the gene-brain-behavior link*.

Seidenberg, M. S., Waters, G. S., Barnes, M. A., & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior*, *23*(3), 383–404. https://doi.org/10.1016/S0022-5371(84)90270-6

Seymour, P. H. K., Aro, M., Erskine, J. M., Wimmer, H., Leybaert, J., Elbro, C., … Olofsson, Å. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*(2), 143–174. https://doi.org/10.1348/000712603321661859

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Sprenger-Charolles, L., Siegel, L. S., Béchennec, D., & Serniclaes, W. (2003). Development of phonological and orthographic processing in reading aloud, in silent reading, and in spelling: A four-year longitudinal study. *Journal of Experimental Child Psychology*. https://doi.org/10.1016/S0022-0965(03)00024-9

Stanback, M. L. (1992). Syllable and rime patterns for teaching reading: Analysis of a frequency-based vocabulary of 17,602 words. *Annals of Dyslexia*. https://doi.org/10.1007/BF02654946

Stanovich, K. E., & Bauer, D. W. (1978). Experiments on the spelling-to-sound regularity effect in word recognition. *Memory & Cognition*, *6*(4), 410–415. https://doi.org/10.3758/BF03197473

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Linguistics and Language Compass*, *9*(8), 311–327. https://doi.org/10.1111/lnc3.12151

Steacy, L. M., Compton, D. L., Petscher, Y., Elliott, J. D., Smith, K., Rueckl, J. G., … Pugh, K. R. (2018). Development and prediction of context-dependent vowel pronunciation in elementary readers. *Scientific Studies of Reading*, *23*(1). https://doi.org/10.1080/10888438.2018.1466303

Stone, G. O., Vanhoy, M., & Van Orden, G. C. (1997). Perception is a two-way street: Feedforward and feedback phonology in visual word recognition. *Journal of Memory and Language*, *36*(3), 337–359. https://doi.org/10.1006/jmla.1996.2487

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1140–1154. https://doi.org/10.1037/0278-7393.21.5.1140

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

Treiman, R., Kessler, B., & Bick, S. (2003). Influence of consonantal context on the pronunciation of vowels: A comparison of human readers and computational models. *Cognition*, *88*(1), 49–78. https://doi.org/10.1016/S0010-0277(03)00003-9

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General*, *124*(2), 107–136. https://doi.org/10.1037/0096-3445.124.2.107

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *45*(1), 2–40. https://doi.org/10.1046/j.0021-9630.2003.00305.x

Venezky, R. L. (1999). *The American Way of Spelling: The Structure and Origins of American English Orthography*. Guilford Press.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language*, *60*(4), 502–529. https://doi.org/10.1016/j.jml.2009.02.001

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, *131*(1), 3–29. https://doi.org/10.1037/0033-2909.131.1.3

Ziegler, Johannes C., Bertrand, D., Lété, B., & Grainger, J. (2014). Orthographic and phonological contributions to reading development: Tracking developmental trajectories using masked priming. *Developmental Psychology*, *50*(4), 1026–1036. https://doi.org/10.1037/a0035187

Ziegler, Johannes C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faísca, L., … Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science*, *21*(4), 551–559. https://doi.org/10.1177/0956797610363406

Ziegler, Johannes C., Montant, M., & Jacobs, A. M. (1997a). The feedback consistency effect in lexical decision and naming. *Journal of Memory and Language*, *37*(4), 533–554. https://doi.org/10.1006/jmla.1997.2525

Ziegler, Johannes C., Petrova, A., & Ferrand, L. (2008). Feedback Consistency Effects in Visual and Auditory Word Recognition: Where Do We Stand After More Than a Decade? *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*(3), 643–661. https://doi.org/10.1037/0278-7393.34.3.643

Ziegler, Johannes C., Stone, G. O., & Jacobs, A. M. (1997b). What is the pronunciation for -ough and the spelling for /u/? A database for computing feedforward and feedback consistency in English. *Behavior Research Methods, Instruments, & Computers*, *29*(4), 600–618. https://doi.org/10.3758/BF03210615