Effects of complexity and unpredictability on the learning of an artificial orthography

Xenia Schmalz, Claudio Mulatti, Gerd Schulte-Körne, Kristina Moll

PII: S0010-9452(22)00090-9

DOI: https://doi.org/10.1016/j.cortex.2022.03.014

Reference: CORTEX 3452

To appear in: *Cortex*

Received Date: 4 February 2022

Revised Date: 23 March 2022

Accepted Date: 26 March 2022

Please cite this article as: Schmalz X, Mulatti C, Schulte-Körne G, Moll K, Effects of complexity and unpredictability on the learning of an artificial orthography, *CORTEX*, https://doi.org/10.1016/j.cortex.2022.03.014.

Conceptualisation: XS, CM, KM & GSK
Methodology: XS, CM, & KM
Software: XS, CM
Formal analysis: XS
Writing - original draft: XS
Writing - review and editing: XS, CM, KM & CM
Supervision: XS & CM
Project administration: XS
Funding acquisition: XS & CM

Effects of complexity and unpredictability on the learning of an artificial orthography

Xenia Schmalz*[1,2], Claudio Mulatti[2,3], Gerd Schulte-Körne [1] & Kristina Moll[1]

[1] Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University Hospital, LMU Munich, Germany

[2] Department of Psychology and Cognitive Sciences, University of Trento, Italy

[3] Department of Developmental Psychology & Socialisation, University of Padova, Italy

* Corresponding author. Email: xenia.schmalz@med.uni-muenchen.de, phone number: +49 (0) 89 4400 55923, address: Pettenkoferstr. 8a, 80336 München, Germany.

Effects of complexity and unpredictability on the learning of an artificial orthography

Xenia Schmalz*[1,2], Claudio Mulatti[2,3], Gerd Schulte-Körne [1] & Kristina Moll[1]

[1] Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University Hospital, LMU Munich, Germany

[2] Department of Psychology and Cognitive Sciences, University of Trento, Italy

[3] Department of Developmental Psychology & Socialisation, University of Padova, Italy

Accepted for publication at Cortex

* Corresponding author. Email: xenia.schmalz@med.uni-muenchen.de, phone number: +49 (0) 89 4400 55923, address: Pettenkoferstr. 8a, 80336 München, Germany.

Abstract

Orthographies vary in complexity (the number of multi-letter grapheme-phoneme rules describing print-to-speech regularities) and unpredictability (the number of words which cannot be read correctly, even with at-ceiling knowledge of the rules). To assess how these constructs affect reading acquisition, we used an artificial orthography learning paradigm, where participants learn to read pseudowords written in unfamiliar symbols, and subsequently read aloud novel words written in the same symbols (generalisation). In three experiments (third experiment pre-registered), we manipulated the consistency of symbol-to-sound mappings: in the first inconsistent condition, vowel pronunciation depended on the subsequent letter (condition complexity), and in the second inconsistent condition, vowel pronunciation was unpredictable from the context (condition unpredictability). Across experiments, we found that pseudowords with inconsistent mappings are more difficult to learn than pseudowords with consistent mappings only, regardless of whether the inconsistency is due to complexity or unpredictability. Numerically, participants learning orthographies containing unpredictable correspondences seem to be less likely to form rules, either for simple or for complex correspondences. We propose that rule extraction and distributional learning happens simultaneously during reading acquisition: in a mathematical model, we show that distributional learning may lead to more complete knowledge than rule extraction for orthographies that are high in unpredictability.

Keywords: Artificial orthography; learning; sublexical processing; orthographic depth.

Effects of complexity and unpredictability on the learning of an artificial orthography

In order to become a good reader, a child first needs to learn and automatise the link between letters or letter clusters and their pronunciation (Blomert, 2011; Jackson & Coltheart, 2001). The efficiency of completing this learning task depends on the properties of the orthography in which a child is learning to read: Empirical work has shown that this process is particularly challenging for children learning to read in English, compared to children learning to read in other European orthographies (Aro & Wimmer, 2003; Frith, Wimmer, & Landerl, 1998; Landerl, 2000; Seymour, Aro, & Erskine, 2003).

The English orthography differs from other European orthographies on a linguistic construct called Orthographic Depth (Frost, Katz, & Bentin, 1987; Katz & Frost, 1992). In comparison to most European orthographies, the relationship between letters and sounds is less consistent in English. For example, the grapheme *a* has five possible pronunciations, as in the words “cat”, “watt”, “fall”, “spa”, and “nation”[1]. On the surface, this high degree of inconsistency (i.e., more than one possible pronunciation for a given orthographic unit) provides a likely explanation for the struggle of English readers: When a child is attempting to read, for example, the pseudoword *walp*, should the cognitive system activate the phoneme /æ/, /ɔ/, /oː/, /aː/ or /æɪ/ for the grapheme *a*?

Determining the mechanisms by which inconsistency affects learning performance is of importance if we want to understand what exact orthography-level characteristics influence the speed of reading acquisition. However, inconsistency is

[1] We define a grapheme as a letter or letter cluster that maps onto a single phoneme, and a phoneme as the smallest spoken unit that can distinguish between two word forms in a language.

also a relevant concept for any research areas which involve the learning of a quasi-regular system (Seidenberg & Plaut, 2014). Quasi-regular systems are characterised by correspondences which are mostly consistent, but include some degree of inconsistency. While we focus on the learning of print-to-speech correspondences, we consider it likely that complexity and unpredictability should have similar effects on learning performance across domains: the computational mechanism responsible for learning the print-to-speech correspondences in one's orthography or, more specifically, in our artificial orthography learning task is likely to be involved in other learning tasks.

Complexity and unpredictability: Two dimensions underlying inconsistency in deep orthographies

Consistency is not sufficient to explain cross-linguistic differences in children's ability to learn print-to-speech correspondences. Orthographies – the English orthography included – provide information about a grapheme's pronunciation that can be used as a cue to resolve inconsistencies (Venezky, 1970). For example, the two graphemes *c* in the English word "cicada" each have a different pronunciation. However, the pronunciation is predictable from the context, and can be described by the print-to-speech correspondence rules *c+[i]* → /s/ and *c+[a]* → /k/. These rules are complex, in the sense that more than one letter is involved in determining their pronunciation, but the pronunciation is nevertheless predictable, in the sense that someone who is familiar with the English orthography but has never read or heard the word "cicada" before will, with a high probability, give the correct pronunciation (Schmalz, Marinus, Coltheart, & Castles, 2015; Venezky, 1970).

Compared to other European orthographies, English contains relatively many complex rules, but also many words where the pronunciation is unpredictable: in order to pronounce words with unpredictable pronunciations, the reader needs to know the whole word (Schmalz et al., 2015; van den Bosch, Content, Daelemans, & de Gelder, 1994). This can be illustrated with some minimal word pairs: for the words *gift* and *gist* the pronunciation of the letter *g* differs, even though both occur at the beginning of a word and before the letter *i*. As the grapheme *g* appears in almost identical contexts, sublexical knowledge, such as knowledge about the frequency of each possible pronunciation, will not suffice to give a correct pronunciation to both of these words. In contrast to words with complex but predictable print-to-speech-sound correspondences, a reader needs lexical knowledge to derive the correct pronunciation for both words (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001).

Thus, there are two theoretically different constructs which underlie inconsistency: a given grapheme may have more than one possible pronunciation, either due to context-sensitive rules, where surrounding letters systematically change its pronunciation, or due to unpredictability, where a change is unsystematic and whole-word knowledge is needed in order to assign a correct pronunciation. It is an open question to what extent these constructs have a different effect on the learning of the orthographic system (i.e., a system of correspondences between orthography and phonology).

On the theoretical level, it could be fruitful to study the learning of systems varying in complexity and unpredictability using connectionist models. In connectionist models of reading, the model is trained on a set of words. The inputs are the orthographic word forms, and the model is trained to output a phonological transcription (Perry, Ziegler, & Zorzi, 2007; Perry, Ziegler, & Zorzi, 2010; Plaut, McClelland, Seidenberg, & Patterson, 1996). The model's success is measured by its ability to generalise the learned knowledge of print-to-speech-sound correspondences to pseudowords or untrained words. Parallel Distributed Processing connectionist models distinguish between an orthography-to-phonology (OP) and an orthography-to-semantics-to-phonology (OSP) route (Plaut et al., 1996). In contrast to the sublexical pathway of the dual-route models, the OP route learns to rely on units larger than graphemes, which can even be whole words (e.g., *yacht* → /jɔt/). However, knowledge of whole words can only be acquired by previous exposure to the word. The division between the OP and OSP route is relevant for distinguishing the effects of complexity and unpredictability: the OP route should be able to represent complex correspondences, while unpredictability is defined as print-to-speech relationships that are not represented in the OP route and therefore require the involvement of the OSP route to yield a correct pronunciation (Schmalz et al., 2015). Thus, a connectionist model that has been trained on all English words except "wasp" and "yacht" is likely to give a correct pronunciation to the former (by extracting generalisable knowledge about the pronunciation of the grapheme *a* when it follows a *w*), but will give an incorrect pronunciation to the latter.

To date, there is little research directly comparing computational models' performance in orthographies differing in depth. To our knowledge, the only study using this approach compared how a connectionist network learned English versus German (Hutzler, Ziegler, Perry, Wimmer, & Zorzi, 2004). The authors were able to simulate the empirical finding of developmental cross-linguistic studies: at the beginning of reading acquisition (or training, in the case of the computational model), the difference between the two orthographies was greatest; it diminished as a function of the amount of training. However, even after learning performance reaches a plateau, generalisation accuracy remained lower for English than for German. Tentatively, we propose that the height of the plateau is likely to reflect differences in predictability: Even when all regularities are learned, there is a degree of uncertainty in how to pronounce a novel letter string. The steepness of the learning curve might differ as a function of complexity: If there are many correspondences to learn, it may slow down the speed of learning, as a larger amount of exposures will be needed for the model to learn the conditional probabilities of graphemes' pronunciations.

On an empirical level, in order to distinguish between the effect of complexity and unpredictability, an experimental approach is required to supplement the existing correlational data in children learning to read in different orthographies (Aro & Wimmer, 2003; Caravolas et al., 2012; Caravolas, Lervag, Defior, Malkova, & Hulme, 2013; Landerl et al., 2013; Moll et al., 2014; Seymour et al., 2003). First, to determine that a behavioural outcome is attributable specifically either to complexity or to unpredictability, one would need to find a set of orthographies that vary orthogonally on these two constructs. In general, complexity and unpredictability in European orthographies are correlated (Schmalz et al., 2015), thus, in the existing datasets, it is difficult to disentangle the independent effects of the two underlying constructs. To date, English has been identified as an orthography which is high in complexity and unpredictability; French is high in complexity, but low in unpredictability, and German and Dutch are relatively low both in complexity and unpredictability (Schmalz et al., 2015; van den Bosch et al., 1994). It is unclear whether any existing orthography is low in complexity, but high in unpredictability, which would be required for an orthogonal design.

Furthermore, existing orthographies differ from each other on characteristics which are theoretically unrelated to but often correlated with inconsistency, including syllable structure, average word length, lexical density, and morphological complexity. This makes it difficult to rule out the possibility that another language-level confound drives any observed cross-linguistic differences (Marinus, Nation, & de Jong, 2015). Thus, studies using existing languages should be used in conjunction with a more controlled experimental approach. Finally, cross-linguistic studies of natural reading development are time- and resource-consuming. Before resources are invested in such a project, smaller-scale experimental studies can help to formulate specific hypotheses.

In the current study, we aim to employ an experimental approach, using an artificial orthography learning task, to dissociate the effects of complexity and unpredictability on the learning of an orthographic system. In contrast to observational studies in children who are learning to read, artificial orthography learning studies have the advantage that they allow us to control item-level variables, as well as the pre-existing knowledge of the participants (as all participants have had no exposure to the orthography prior to participating in the experiment). While this limits the claims we can make about learning to read in children, it allows us to derive hypotheses that can be later tested in a more ecologically valid setting and across domains. Here, we focus on two questions: (1) The rate of learning, and (2) performance on a generalisation task, where participants have to generalise their knowledge of newly extracted regularities to a set of unfamiliar items built on the same principles as the training set (Bitan & Karni, 2003; Taylor, Plunkett, & Nation, 2011).

How may complexity affect learning processes?

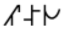
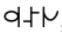
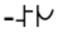
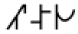
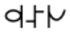
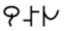
When learning a complex, but predictable system (such as the French orthography), knowing the complex rules should be sufficient to achieve ceiling accuracy on a generalisation task, involving the reading aloud of novel words or pseudowords: A French-speaking child who has learned the context-sensitive rules relating to the pronunciation of the grapheme *c* will pronounce the pseudowords *cilp* and *colp* with the initial phonemes /s/ and /k/, respectively.

In terms of the speed of learning, we expect that learning a system containing complex correspondences should take longer than when learning a system containing only simple correspondences. There are two reasons for this: First, the presence of complex correspondences increases the total numbers of correspondences that need to be learned (Asfaha, Kurvers, & Kroon, 2009). For example, in an orthography where the grapheme *g* is consistent (e.g., German), the child needs to extract and memorise only one grapheme-phoneme correspondence rule: *g* → /g/. In French, where the pronunciation of the grapheme *g* is inconsistent but predictable, the child needs to extract and memorise two rules: *g* → /g/, and *g+[e,i]* → /ʒ/.

Second, even once such rules are learned, the cognitive system may activate two possible pronunciations, one of which would need to be repressed by the cognitive system once the context-sensitive rule is applied. This would lead to slower naming latencies and occasional mispronunciations (Rastle & Coltheart, 1998; Rey, Jacobs, Schmidt-Weigand, & Ziegler, 1998; Rey & Schiller, 2005). Thus, a higher degree of proficiency might be needed before the cognitive system activates the pronunciation corresponding to a multi-letter cluster automatically and causes minimal interference with the single-letter pronunciations.

In summary, simple rules in an orthography should be easier to learn than complex rules. However, once proficiency is obtained, the knowledge of complex rules should be generalisable to novel words, resulting in high pseudoword reading accuracy.

How may unpredictability affect learning processes?

Unpredictability should have an effect both on the accuracy in a generalisation task and on the rate of learning during training. Complete unpredictability is rare in a linguistic setting, because two different outputs for a given input are, in practice, never perfectly equiprobable. However, unpredictability can be created in an experimental setting, by assigning pronunciations to a set of symbols in such a way that the pronunciation is not predictable from context. For example, one could create a mini-orthography, with a training set consisting only of the two pseudoword pronunciations  and , where the body  is pronounced as /æm/ in the first case, and as /ɔm/ in the second case. This way, in the training material, there is no statistical regularity that could provide participants with a cue as to the correct pronunciation, as long as the first consonant is not predictive of the pronunciation. Consequently, both possible pronunciations should occur with an equal probability in a generalisation task: if, after being trained on  → /læm/ and  → /nɔm/, participants are presented with the novel pseudoword (provided they know the correspondence → /t/), it is of interest if they are equally likely to give the pronunciations /tæm/ and /tɔm/, reflecting the statistics of the input material.

In terms of the speed of learning, we may expect a qualitatively different learning strategy when the system contains unpredictable print-to-speech correspondences (as in the English orthography). The presence of unpredictable correspondences may impair the system's ability to extract the regularities. This may lead to a whole-word rote memorisation approach during training (Bitan & Karni, 2003). There is a larger number of pseudowords than letters, meaning that there will be more items to memorise when the cognitive system works with whole pseudowords compared to sublexical units. This should lead to an overall increase in the amount of exposure needed for participants to learn the training set.

A related question is how the presence of unpredictable correspondences affects the learning of predictable correspondences within the same system. If the presence of unpredictable correspondences changes the cognitive process towards rote memorisation, this should also have a detrimental effect on the system's efficiency in extracting the existing regularities. Establishing whether this is the case has practical implications. The well-established finding that English-speaking children lag behind their European peers in terms of the speed of reading acquisition is generally attributed to orthographic depth. However, if we find support for the above prediction, the combination of unpredictability and complexity might be the detrimental factor, where the unpredictable correspondences would prevent the cognitive system from learning the existing (complex) regularities.

Pilot data

The question of how complexity and unpredictability affect the learning of orthography-phonology mappings can be addressed with an artificial orthography learning experiment (Bitan & Karni, 2003; Taylor et al., 2011; Yoncheva, Wise, & McCandliss, 2015). Here, participants learn the spoken and written form of a set of pseudowords, which are written in unfamiliar symbols, in a training phase. In a subsequent generalisation phase, participants are presented with untrained pseudowords written with the same symbols, and need to generalise their knowledge of the symbol-sound correspondences. As pilot data, we present the results from two experiments on the effect of complexity and unpredictability, one using a within-subject design, and one with a between-subject design.

Pilot Experiment 1: Within-subject design

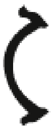
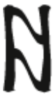
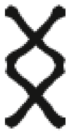
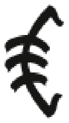
In the current experiment, the aim was to compare the learning of consistent, complex-predictable, and unpredictable correspondences in a within-subject design. Originally, the data were collected as two experiments (“whole-word” and “phonics” conditions): the only difference between them was a pre-exposure phase for the “phonics” condition, where the participants were exposed to the symbols, individually, and the sounds that each symbol corresponds to, prior to learning the training set. As the results did not differ across conditions, we collapse across them here to maximise the sample size (for a similar null-result on the behavioural level, see Bitan & Karni, 2003).

Methods

Participants. The participants were native Italian speakers. We aimed to test 16 participants per condition (i.e., 32 across the two experiments). Participants were trained until they achieved an accuracy rate of the training material of >70% (see the “Procedure” section below). In case a participant could not reach this threshold within 10 repetitions of the training block, the testing session was terminated and the participant was not included in any of the analyses. Twenty-six and 23 participants were tested for the “whole-word” and “phonics” conditions, respectively, in order to obtain 16 complete data sets. Thus, 17 participants had to be excluded.

Materials. The artificial language consisted of a set of three-phoneme (CVC) pseudowords, which were made up of the consonants /b/, /d/, /f/, /g/, /k/, /m/, /n/, /p/, /s/, /t/, /v/, and /z/, and the vowels /a/, /ɛ/, /e/, /i/, /o/, and /u/. The pseudowords were spelled with ancient Hungarian runes (Taylor et al., 2011). For the consonants, there was a consistent, one-to-one mapping between phonemes and symbols. The consistency manipulation was introduced for the vowels’ pronunciations (see Table 1).

Table 1: Consistent, complex and unpredictable vowel symbols and their correspondences, as used in Pilot Experiment 1

Character	Pronunciation	Condition
	/a/	Consistent
	/ɛ/	Consistent
	/e/ (before /f/, /k/ or /m/) /i/ (before /d/, /t/ or /z/)	Complex-Predictable
	/o/ /u/	Unpredictable

There were four vowel symbols. Two of them mapped consistently onto the same sound (/a/ and /ɛ/, respectively). The pronunciation of the “complex-predictable” symbol was determined by the subsequent consonant (/e/ before /f/, /k/, /m/, and /i/ before /d/, /t/, /z/). The pronunciation of the “unpredictable” symbol was counter-balanced, such that it could map onto two different pronunciations, which occurred equally often, both in total and across consonantal contexts: for example, the body -OZ occurred in two pseudowords in the training set, once with the pronunciation /oz/ and once with the pronunciation /uz/.

The training set consisted of 36 pseudowords in total, and 12 per condition. The generalisation set consisted of 36 pseudowords which did not occur during the training phase but were created from the same symbols by exchanging the onset consonant. Again, there were 12 items per condition: In the consistent condition, 6 items each contained one of the two consistent vowel symbols. In the complex-predictable and unpredictable conditions, each of the two vowel symbols occurred 6 times. A full list of items is downloadable from here: https://osf.io/z8d72/.

Procedure. The procedure was closely based on Taylor et al. (2011). Participants were tested individually in sessions lasting up to 1.5 hours. Item presentation for all experimental tasks was controlled with the software DMDX (Forster & Forster, 2003). The experimental session consisted of the following tasks:

(0) Exposure to the symbol-sound correspondences (for the "Phonics" condition only): Participants were presented with each symbol for 5 seconds and heard the pronunciation simultaneously. The order of presentation was randomised. Participants were instructed to repeat the phoneme, and to try and remember the symbol. Each symbol was presented only once, including the inconsistent symbols: we presented the complex-predictable symbol with the phoneme /e/ and the unpredictable symbol with the phoneme /o/. As explained below, this manipulation does not seem to have been strong enough to achieve any group differences at the behavioural level.

(1) Exposure to the training set: Both groups of participants saw each of the training items for five seconds, and simultaneously heard the pronunciation. They were instructed to repeat each pseudoword, and to try and remember its spelling. Each pseudoword was presented only once, and the order of presentation was random.

(2) Training phase: Here, the participants saw each training item, one at a time, and were instructed to read it aloud. It was stressed that they should guess, even if they felt like they had no idea. Self-corrections were allowed: if the participants gave more than one response to a given item, the final response was scored. The

items were presented in random order. The item remained on the screen for 5 seconds, and at the end of the 5-second interval, it was always followed by the correct pronunciation. The experimenter immediately scored the accuracy of each response (i.e., the match between the participant's response and the feedback response at the end of each trial): responses where all symbols were pronounced correctly were scored as 1, responses with any mistakes were scored as 0. After the participants had completed the block, the experimenter calculated the accuracy rate. If the accuracy rate was below 70%, the whole training block was repeated, meaning that the participants read all of the training items again, until accuracy exceeded 70% or for a maximum of 10 repetitions. This means that the number of exposures was equal for all items for a given participant, though it varied across participants. In addition to the overall accuracy, we transcribed the responses, for each participant, from the last repetition of the training block.

(3) Generalisation phase: Participants were presented with items which were created from the same symbols as the training set but did not occur during the training phase. These items were created by exchanging the onset symbol of the training items. The items were presented in random order, for five seconds or until the voice-key was triggered. The participants were told that they would see a set of new words from the orthography that they had learned, and that they would need to read aloud each word. They were encouraged to guess if they were unsure. Each item was shown only once, and the participants' responses were transcribed offline.

Results

As mentioned above, we collapsed across data from the “whole-word” and “phonics” conditions. Data for the two conditions, separately, can be found here: https://osf.io/z8d72/.

Participants took between 2 and 7 runs through the training set until they reached the threshold of >70% of the overall accuracy. In addition, as mentioned above, 17 participants did not reach the threshold even after 10 repetitions of the training set, and were excluded from all analyses. At the last block, the accuracy rate across item conditions was 84.4% (by-participant SD = 9.4%) for the consistent items, 66.1% (SD = 11.8%) for the complex items, and 66.1% (SD = 14.3%) for the unpredictable items. An ANOVA on the participant-level data showed a significant effect of item type on accuracy, $F(2,62) = 18.06$, $p < 0.0001$. Pairwise post-hoc *t*-tests with Bonferroni correction showed that the accuracy for complex items differed from the accuracy for consistent items, $p < 0.0001$, Cohen’s $d = 1.7$, and accuracy for unpredictable items differed from the accuracy of consistent items, $p < 0.0001$, $d = 1.5$, but there was no significant difference between the accuracy of the complex and unpredictable items, $p > 0.9$, $d < 0.1$. These results reflect higher accuracy for consistent than inconsistent items.

For the generalisation phase, we first scored the accuracy of the responses. At first pass, we used a lenient marking criterion: for the complex and unpredictable conditions, we marked both plausible pronunciations as correct. This means that, for the complex condition, we counted the vowel responses /i/ and /e/ as correct, even if they did not correspond to the context-sensitive rule. In this scoring system, the accuracy for the three conditions was 77.9% (SD = 21.4) for the consistent condition, 77.3% (SD = 23.7%) for the complex condition, and 76.0% (SD = 20.3) for the unpredictable condition. A three-way ANOVA did not show a significant difference between the three item types, $F(2,62) = 0.05$, $p > 0.9$.

The theoretically interesting questions relate to the vowel responses that the participants provided for the complex and unpredictable conditions. In the complex condition, it is of interest whether the participants learned the context-sensitive rule, that the vowel was determined by the subsequent consonant letter. To address this question, we first excluded all responses that were counted as errors in the above analysis (i.e., all responses that contained either consonant errors or vowel phonemes that never corresponded to the vowel symbol), which left us with 77.3% of the “plausible” responses from the complex condition. From these, we calculated the percentage of responses that were also context-appropriate. The mean percentage of context-appropriate responses was 48.7% (SD = 18.7). As chance level is at 50%, this suggests that, on the group level, the participants did not learn the context-sensitive rule. To assess whether any individual participants learned the context-sensitive rule, we used the binomial distribution to calculate the probability of each participant’s number of successes out of their total number of plausible responses, under the null hypothesis of no learning (i.e., that accuracy = 50%). On the individual level, performance was significantly above chance for 4 out of the 32 participants, $p < 0.05$, which may be interpreted as evidence for learning.

For the unpredictable condition, it is of interest whether participants’ responses reflect the statistical distribution of the items encountered during training. In this case, each participant should give approximately 50% of the two different vowel responses across the items (assuming that the pre-exposure phase for the phonological condition did not bias the participants towards a default pronunciation). Alternatively, participants may “over-generalise” in the sense of deducing a single rule for a default pronunciation and applying this to all instances of the symbol. To quantify whether a given participant (X) tended to use a default response or if they gave a mixture of responses, we used the concept of entropy (H) from information theory. Entropy measures the amount of information contained in a signal: a consistent preference for one vowel pronunciation over the other would provide a strong signal, leading to a low entropy value (H(X) → 0). An equal split carries little information, and would result in high values of entropy (H(X) → 1). We used the binary entropy function,

$$\mathrm{H}(X) = -p \ \log_2 p - (1-p)\log_2(1 - p)$$

to calculate the entropy of each participant’s plausible responses for the unpredictable condition. p is the observed proportion of /o/ responses for a given participant (after excluding incorrect responses).

Across participants, entropy varied between H = 0 and H = 0.99, with an average = 0.28 and SD = 0.34. Eighteen out of the 32 participants had an entropy of 0, reflecting that they consistently used the same pronunciation for the vowel symbol. In the phonics condition, the average entropy was numerically higher (mean = 0.38, SD = 0.34) than in the whole-word condition (mean = 0.17, SD = 0.32), Welch’s *t*(30) = 1.71, *p* = 0.1. Participants in the phonological condition were exposed, in the pre-exposure phase, to the inconsistent symbol with the pronunciation /o/; however, the participants in this condition were approximately evenly divided into those that tended to give the /o/ pronunciation (N = 9, average entropy = 0.35), and those that gave the /u/ pronunciation (N = 7, average entropy = 0.41). The numerically greater entropy and the even distribution of responses in the phonological pre-exposure condition suggest that the pre-exposure manipulation did not substantially bias the participants’ generalisation responses – a stronger manipulation, such as a training phase where participants would need to memorise the symbol-sound correspondences may have been needed to achieve an effect on the behavioural level.

Discussion

In the first pilot experiment, we trained 32 adults on an artificial orthography, and subsequently tested their ability to generalise their freshly learned knowledge to a set of unfamiliar items which were written in the same symbol. The consistency of the vowel's pronunciation was manipulated: a given vowel either corresponded consistently to the same pronunciation (consistent), or it corresponded to two different pronunciations in such a way that the pronunciation was predictable from the subsequent letter (inconsistent-complex-predictable), or it could map onto two different pronunciations which varied unsystematically (inconsistent-unpredictable).

Several findings are of interest. First, a substantial number of participants (34.7%) were not able to finish the training phase, even after 10 repetitions of the items. The high non-completion rate limits the generalisability of our results, as it is not clear in what ways the participants who did not learn the orthography differed systematically from those who did complete all tasks. The high non-completion rate is in contrast to a previous study on which we based our design (Taylor et al., 2011), where all participants learned 36 training items to >70% accuracy within 6 runs through the training block. Second, most participants in our experiment were not able to learn the context-sensitive rule: only 4 of our participants were significantly above chance in assigning a context-appropriate vowel pronunciation, and on the group level, accuracy was numerically below chance. This is also in contrast to the previous study by Taylor et al. (2011): Here, the authors found that participants were able to provide correct responses to generalisation items with context-sensitive rules (though it was not tested how many participants were above chance at assigning the correct vowel pronunciation).

The addition of an unpredictable condition is a new manipulation which was not used by Taylor et al. (2011). We found that participants showed a preference for one pronunciation over the other, despite the training set containing an equal amount of exposure for each. This is in line with a rule-based account of print-to-speech correspondence representation (Coltheart et al., 2001): rather than reflecting the statistical distribution of the input set, participants form a symbol-to-letter correspondence which does not seem to represent the uncertainty in the input.

The two results that differ across the current experiment and that of Taylor et al. (2011) warrant further discussion. Our design was very similar to this earlier study, in terms of the number of items, types of symbols used, the manipulation of consistent and inconsistent symbol-sound correspondences, type and order of the tasks, and the training procedure. The difference in the non-completion rate seems substantial (0% in Taylor et al., 35% in our study), and we drew different conclusions about whether or not participants learn context-sensitive rules. There are two potential explanations for the different results: First, our study was conducted in Italian native speakers, while the study of Taylor et al. (2011) was conducted with English native speakers. English and Italian are considered to be on opposite ends of the orthographic depth continuum (Seymour et al., 2003; Tabossi & Laghi, 1992). These orthographies especially differ in that vowels are very consistent in Italian, while they are the main source of inconsistency in English (Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995). It is possible that the cognitive systems of Italian native speakers are less well equipped to deal with inconsistent print-to-speech correspondences. As in the Italian orthography, the pronunciation of some consonants is context-dependent while the pronunciation of vowels is perfectly consistent (with the exception of isolated loanwords), the participants may have been biased (consciously or subconsciously) towards picking up higher-level regularities when it comes to the consonant pronunciation, at the cost of extracting regularities about the vowel pronunciation.

A second difference between the two experiments was the introduction of an “unpredictable” condition in our study. An alternative explanation is therefore that adding an unpredictable component may affect overall learning performance (Kempe & MacWhinney, 1998; Tamminen, Davis, & Rastle, 2015; Wonnacott, 2011). This is in line with the finding of our high non-completion rate. Our participants also struggled to learn the context-sensitive rule: we hypothesise that introducing an unpredictable component specifically affects the learning of complex rules. Participants may form (consciously or unconsciously) hypotheses about how the orthographic system works. If they construct a general hypothesis that a symbol’s pronunciation may depend on the subsequent vowel, this hypothesis becomes disproved once they encounter an instance of an unpredictable correspondence, where the same symbol has different pronunciations, even when followed by the same consonant.

These two explanations are not mutually exclusive. The role of the characteristics of the participants’ native orthographies is of both theoretical and practical interest, as it would show how learning of new material is entrenched into pre-existing knowledge (Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018), and would have implications for second language learning. For the current purposes, we are interested in the latter explanation, as it would suggest that learning an orthography becomes more difficult when an unpredictable component is introduced.

This would directly explain why learning to read in English is more challenging than learning to read in other European orthographies. A follow-up experiment is needed, where we directly compare learning and generalisation performance in an artificial orthography containing complex correspondences, and one containing complex and unpredictable correspondences.

Pilot Experiment 2: Between-subject design

In a second pilot experiment, we created two different orthographies: one contained consistent and complex correspondences, and the other contained consistent and unpredictable correspondences.

Methods

Participants. The participants were native Italian speakers. None had participated in the previous experiment. Again, we aimed to have 16 participants per condition ("consistent + complex" and "consistent + unpredictable"), and we excluded participants if they did not reach the pre-determined threshold of >70% accuracy after 10 repetitions of the training block. In the complex condition, 24 participants were tested until we obtained 16 complete datasets, and in the unpredictable condition, 23 participants. Thus, 15 participants had to be excluded.

Materials. The items were CVC pseudowords, which consisted of the consonants /b/, /d/, /f/, /g/, /k/, /m/, /n/, /p/, /s/, /t/, /v/, and /z/, and the vowels /a/, /ɛ/, /o/, and /u/, and were spelled with ancient Hungarian runes. The two training sets (complex and unpredictable) contained 24 items each. For the "unpredictable" condition, these were identical to the Consistent and Unpredictable conditions from Pilot Experiment 1. For the "complex" condition, we changed the pronunciations of the vowels in the "unpredictable" condition: the participants were presented to the same visual stimuli in both conditions, but in the complex condition the auditory stimuli were changed in such a way that the vowel symbol was pronounced as /o/ before /z/, /f/, and /m/, and as /u/ before /d/, /t/, and /k/. This allowed us to use the same visual stimuli across condition during the generalisation phase.

The generalisation phase consisted of 24 items: 12 with consistent vowel pronunciations (/a/ or /ɛ/) and 12 items with inconsistent vowel pronunciations, where the pronunciation was predictable for participants who had been exposed to the "complex-predictable" training set, and unpredictable to participants who had been exposed to the "unpredictable" training set.

Procedure. The procedure was identical to Pilot Experiment 1. There was no pre-exposure phase.

Results

In both conditions, participants took between 4 and 10 runs through the training items until they reached >70% accuracy. The average number of run-throughs was 6.6 (SD = 1.2) for the complex-predictable condition, and 6.3 (SD = 1.9) for the unpredictable condition. The accuracy at the last repetition of the practice block is summarised in Table 2.

Table 2: Results from the training phases of Pilot Experiment 2: Last block accuracy. Percentage (SD)

Condition	**Complex-predictable participant group**	**Unpredictable participant group**
Consistent items	80.7% (10.0)	80.2% (10.0)
Inconsistent items	73.4% (9.2)	71.9% (10.0)

We conducted a 2x2 ANOVA on the participant-level data, with condition (complex versus unpredictable) as a between-subject manipulation and item type (consistent versus inconsistent) as a within-subject manipulation, and their interaction. We obtained a main effect of item type, *F*(1,30) = 9.0, *p* = 0.0054, but no significant effect of participant condition, *F*(1,30) = 0.2, *p* = 0.6340. There was also no significant interaction between item type and participant condition, *F*(1,30) = 0.04, *p* = 0.8428. The effect of item type reflected a lower accuracy for inconsistent compared to consistent items (Cohen's *d* = 0.83).

The accuracy for the generalisation phase (the percentages of plausible responses, i.e., using a lenient marking criterion) is summarised in Table 3. We conducted an analogous ANOVA to the training data. Here, we obtained a significant effect of participant condition, *F*(1,30) = 4.9, *p* = 0.0338, but no significant effect of item type, *F*(1,30) = 1.1, *p* = 0.3110, and no significant interaction, *F*(1,30) = 0.2, *p* = 0.7330. The effect of participant condition reflects higher accuracy for the unpredictable compared to the complex-predictable training groups, as shown in Table 3.

Table 3: Results from the generalisation (test) phase of Pilot Experiment 2. Percentage of plausible responses (SD)

Condition	**Complex-predictable participant group**	**Unpredictable participant group**
Consistent items	54.7% (20.0)	68.8% (18.3)
Inconsistent items	57.0% (29.5)	72.9% (15.1)

For the participants who learned the complex-predictable condition, we counted the number of correct responses in the inconsistent condition, where the vowel pronunciations were in line with the context-sensitive rule. We calculated the percentage of correct responses out of the number of plausible responses. The average percentage of correct out of plausible responses was 65.1% (SD = 18.7). On the group level, this was significantly above chance, $t(15) = 3.2$, $p = 0.0028$ (one-sided *t*-test). Fourteen out of the 16 participants gave more correct (context-appropriate) than incorrect-but-plausible responses. Four of the participants were significantly above chance, $p < 0.05$.

For the unpredictable condition, we calculated the entropy value, as for Pilot Experiment 1, to quantify the degree to which participants tended towards giving a consistent vowel pronunciation for all items, or whether their responses reflected the statistical distribution of the training set. The average entropy was 0.9 (SD = 0.25). One participant had an entropy value of 0; the second-lowest value was 0.8; the high entropy value shows that most participants' responses were very heterogeneous.

Discussion

Both Pilot Experiment 1 and Pilot Experiment 2 showed that learning consistent correspondences is easier than learning inconsistent correspondences (as evidenced by a difference in accuracy in the last training block). This suggests that inconsistency, regardless of the underlying reason for why a particular correspondence is inconsistent, affects the speed of learning. In the generalisation phase, when we consider the number of plausible vowel pronunciations, we did not find any significant group differences in either experiment. As the participants gave plausible responses to inconsistent symbols, this suggests that participants learn the phonemes that the symbol can, in principle, correspond to.

Numerically, it does not seem to be the case that unpredictable correspondences are harder to process than complex-predictable correspondences. In both experiments, accuracy during the last training block was lower for inconsistent than consistent items, but the results for items with complex-predictable versus unpredictable correspondences were very similar. In Pilot Experiment 1, we did not find a significant difference in generalisation performance across the two inconsistent conditions; for Pilot Experiment 2, the participants in the unpredictable condition even performed better than the participants in the complex-predictable condition.

There are also some differences in the results across the two pilot experiments. First, in relation to the context-sensitive rule in the complex-predictable condition, learning seems to have occurred on the group level in Pilot Experiment 2 but not Pilot Experiment 1 (where performance was numerically below chance). On the individual level, it is difficult to draw conclusions: Numerically, a smaller percentage of participants were above chance in Pilot Experiment 1 (4/32, i.e., 12.5%) than Pilot Experiment 2 (4/16, i.e., 25%). We cannot confirm this numerical trend with a significance test, however, because these numbers are not directly comparable: the error rate (i.e., percentage of implausible pronunciations) was relatively high for Pilot Experiment 2, leading to a smaller number of valid trials, meaning the accuracy threshold to achieve significantly above-chance performance is higher than for Pilot Experiment 1. Numerically, though, the findings are in line with our expectations: When the orthography contains unpredictable as well as complex correspondences (as in Pilot Experiment 1), the complex correspondences are more difficult to learn than when there are no unpredictable correspondences (as in Pilot Experiment 2).

Second, while the participants in Pilot Experiment 1 tended to provide the same vowel pronunciation for unpredictable items, we did not observe this trend in Pilot Experiment 2: Most participants had a high entropy value, reflecting a close to 50-50 distribution of the vowel responses. This suggests that participants may have learned the statistical distribution of the input and used a probabilistic procedure to provide the output. This across-experiment difference is unexpected. It could suggest that participants are less likely to learn the statistical distribution of an input when there is a higher degree of systematicity: rather than extracting the statistical properties, the cognitive system may be focussed on trying to extract a rule that would describe the pronunciation of an unpredictable symbol.

Overall, the accuracy in the generalisation phase was lower in Pilot Experiment 2 than Pilot Experiment 1. Also, the non-completion rate was comparable across the two experiments, even though the item set was smaller in the second (24 pseudowords) than the first (36 pseudowords) experiment. In the generalisation phase of Pilot Experiment 2, contrary to our predictions, participants who learned the unpredictable orthography performed better than participants who learned the complex-predictable orthography, both for the consistent and inconsistent items. We have no straight-forward explanation for this set of results. One speculative explanation could be that participant characteristics such as motivation can influence their performance in this task. As we did not match participants across groups, there may have been some unanticipated differences, by chance, for example in the average levels of motivation of the participants. Such individual differences become problematic for drawing confident conclusions from a between-subject experiment such as ours: due to the small number of participants, some of the results may be driven by participant characteristics rather than our manipulations.

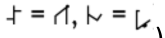
Pre-registered experiment: Combined within-between design

In Pilot Experiments 1 and 2 we set out to explore the way in which complexity and unpredictability of symbol-sound correspondences affect participants' learning of an artificial orthography. While some results were in line with theoretical expectations, others were unexpected and may have been driven by differences in participants characteristics. Thus, a follow-up study is needed to verify if the theoretically expected results replicate in a better-controlled setting. The follow-up study is pre-registered: the approved Stage 1 submission can be found here: https://osf.io/c4bvp/.

In this pre-registered experiment, we performed a mixed within- and between-subject comparison. For the between-subject comparison, we took care that participants were matched, between conditions, on their learning ability, which should also reflect their overall level of motivation and attention. The between-subject contrast allowed us to compare the learning of complex rules under different circumstances: in the presence or absence of unpredictable correspondences. If the same participants learned two different orthographies in a completely within-subject design, the first orthography may shape the participants' expectations of the second orthography.

To match the participants across conditions, first, all participants learned a mini-artificial-orthography with consistent correspondences only. We also recorded their ages. Each participant was then assigned to one of the two experimental orthographies by pair-wise matching, such that the average learning ability and age were kept similar across the two conditions. We aimed to recruit a larger number of participants than in the pilot experiments (see "Participants" section below).

Furthermore, we increased the number of items in the generalisation phase, which should give more stable estimates of participants' accuracy across conditions. This is difficult with Italian native speakers: Italian only has six vowel phonemes (or 5 in some dialects, which do not distinguish between /e/ and /ɛ/), which makes it difficult to create a large number of phonotactically legal (i.e., pronounceable) phoneme strings from a combinatorial perspective. We therefore conducted the experiment in German, which has a larger number of phonemes.

We aimed to address an additional shortcoming of Pilot Experiments 1 and 2. As participants may have prior knowledge or associations which may make some symbol-sound correspondences easier to learn for a given participant, it is advisable to create two parallel versions of the orthography in an artificial orthography learning task (Taylor et al., 2011). We did not do this for the pilot experiments. In the proposed study, there were two parallel versions of each of the two orthographies, which were counterbalanced across participants. While the phonology was identical in these orthographies, different symbols corresponded to each phoneme (upper-case versus lower-case letters in the BACS font, which are distinct visually, e.g., ).

For the experimental task, each participant learned one of two orthographies: The first contained consistent, complex-predictable and unpredictable correspondences (akin to Pilot Experiment 1). The second contained consistent correspondences, and two sets of complex-predictable correspondences. This allowed us to directly compare the performance on the consistent and first set of complex-predictable training and generalisation items, as these were identical across the conditions. In the generalisation set, we included an additional condition: Pseudowords with unfamiliar bodies (-*VC* clusters), where the inconsistent vowels were followed by consonants which did not occur in the final position during the training set. This provides an “unpredictable” condition both for participants who learned the consistent+complex+complex orthography, and for participants who learned the consistent+complex+unpredictable orthography.

Participants

The participants were staff or students at universities in southern Germany, who received course credit or vouchers for their participation. They were tested in individual sessions lasting up to 2 hours. We aimed to obtain at least 25 complete, pairwise-matched datasets for each condition, i.e., 50 participants who pass the training phase. Due to the COVID-19 pandemic, we were unable to recruit a sufficient number of participants within the timeframe covered by the funding for this project. Thus, we terminated data collection when we had 22 participants in the consistent+complex+complex condition, and 17 in the consistent+complex+unpredictable condition. One participant learning the complex orthography did not pass the training, as did one participant from the unpredictable condition.

We analysed the data with Bayes Factors (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Originally, we aimed to collect data until we found evidence for the main hypotheses, as described below: we planned to continue recruiting participants until the evidence became clear for each of the tests (BF > 6 or BF < 1/6). However, the sample size was limited by practical considerations (see above).

Methods

Experimental learning task

We created an artificial orthography. All items had a CVC structure, and all phonemes occurred in the German phonology. The symbols used for this orthography were taken from the BACS font (Vidal, Content, & Chetail, 2017): This is a set of characters, designed specifically for experiments using artificial orthographies, to be similar to letters in terms of visual complexity.

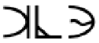
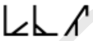
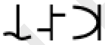
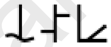
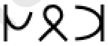
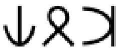
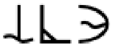
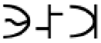
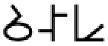
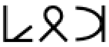
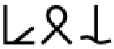
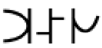
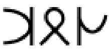
The consonants that made up this language are /f/, /ʃ/, /z/ or /s/[2], /ts/, /k/, /l/, /m/, and /p/. The vowels were /a:/, /o:/, /u:/, /ɛ:/, /e:/, /ø:/, and /y:/, and the diphthongs /aɪ/, /aʊ/ and /ɔʏ/ (which were represented in our orthography as single symbols). Each consonant consistently mapped onto a single symbol. The vowel phonemes /e:/ and /ø:/ also had consistent pronunciations. In the complex orthography, the four symbols corresponding to /a:/, /o:/, /u:/ and /ɛ:/ in front of /f/ or /ʃ/ were pronounced as /aɪ/, /ɔʏ/, /y:/ and /aʊ/, respectively, if succeeded by /k/ or /p/. This means that we have the context-sensitive rule X[/f/, /ʃ/] → P_1 and X[/k/, /p/] → P_2, where X corresponds to one of the four vowel letters and P1 and P2 to two possible pronunciations. In the unpredictable orthography, the pronunciation of the two symbols corresponding to /u:/ or /y:/, and /ɛ:/ or /aʊ/, was counterbalanced in such a way that it was not predictable from the subsequent consonant. Thus, the complex-predictable orthography had two consistent vowel symbols, and four inconsistent vowel symbols which had complex-predictable symbol-sound correspondences (consistent+complex+complex). The unpredictable orthography also had two consistent vowel symbols, and two inconsistent symbols with a complex-predictable link to phonology, and two inconsistent symbols with an unpredictable link to phonology (consistent+complex+unpredictable). The conditions are described in Table 4.

Each training set consisted of 40 training items. Both orthographies contained eight items with consistent pronunciations: each of the two consistent vowels occurred four times. For the inconsistent conditions, the four inconsistent vowels were

[2] In Standard German, word-final phonemes are always unvoiced; to reflect this phonotactic constraint, the symbol S is pronounced as /z/ at the beginning of words, and /s/ at the end of words. However, the pronunciation differs across dialects: in Southern Germany, /z/ often changes to /s/. Thus, we scored both the /s/ and the /z/ pronunciations as correct.

embedded into orthographic bodies. The bodies were thus made up of the inconsistent vowel symbols and four consonant symbols. Each of the bodies occurred twice during the training set, paired with different consonants at the onset. Thus, each inconsistent vowel symbol was presented eight times: four times with one pronunciation and four times with another pronunciation. This means that all vowel-related symbol-sound correspondences occurred four times in the training set.

Table 4: Description of the training conditions for the two orthographies and three item types.

Condition		**Example item spelling**	**Example item pronunciation**	
Training	**Number of items**		**Complex orthography**	**Unpredictable orthography**
Consistent	8		/fe:ts/	/fe:ts/
	8		/pe:l/	/pe:l/
Complex	8		/ka:f/	/ka:f/
	8		/kaɪp/	/kaɪp/
Complex / Unpredictable	8		/mɛ:f/	/maʊf/
	8		/ʃɛ:f/	/ʃɛ:f/
Generalisation			**Correct pronunciation(s)**	
Consistent			/ke:ts/	/ke:ts/
Complex			/tsa:f/	/tsa:f/
			/zaɪp/	/zaɪp/
Complex / Unpredictable			/pɛ:f/	/pɛ:f/ or /paʊf/
			/paʊk/	/pɛ:f/ or /paʊk/
Unpredictable			/fa:m/ or /faɪm/	/fa:m/ or /faɪm/
			/fɛ:m/ or /faʊm/	/fɛ:m/ or /faʊm/

Note: Pink shading = consistent, green shading = complex-predictable, blue shading = unpredictable.

Note that, in the consistent+complex+complex condition, participants had an additional advantage for learning the complex rules, namely that they were presented with two more items containing complex rules, which may better equip them to grasp the concept of context-sensitivity in the artificial orthography. If it is indeed the case that the presence of many context-sensitive rules facilitates the extraction of a single context-sensitive rule, this would be in line with our explanation of why the presence of unpredictable rules may impede the extraction of context-sensitive rules (see Discussion section of Pilot Experiment 1): the presence of unpredictable correspondences may serve to disprove hypotheses that participants create about the role of context in determining a vowel's pronunciation. The flip side would be that the presence of more context-sensitive rules could confirm participants' hypotheses. Thus, doubling the number of context-sensitive rules maximises the manipulation.

The generalisation set was identical for the two orthographies. It contained 96 items in total: 24 items with consistent vowel symbols, and 3 x 24 items with inconsistent symbols. From the inconsistent items, 24 items contained symbol-sound correspondences that should be predictable by complex rules for all participants. A further 24 items contained correspondences that should be predictable by complex rule for the participants who learned the consistent+complex+complex orthography, and unpredictable for the participants who learned the consistent+complex+unpredictable orthography. The final set of 24 items contained items with vowel symbols which occurred in the inconsistent condition for all participants, but the vowel symbol were followed by a consonant that never occurred after the vowel symbol, thus removing the context necessary for applying a context-sensitive rule and rendering the pronunciation of these items unpredictable. The conditions are described in Table 4, and all items can be downloaded from https://osf.io/z8d72.

The procedure, including the exposure, training and generalisation phase, were identical to the procedure described for the pilot experiments.

Matching task

Before the participants learned the orthography for the experimental task, we trained them on an easier artificial orthography. This task served to ensure that participants were matched on learning performance across conditions, which we assumed would capture both their cognitive capability and motivation. The easy orthography consisted of four consonants /b/, /j/, /g/, and /n/, and three vowels, /o:/, /u:/, and /y:/. All pseudowords in the orthography had a CVCV structure. Each phoneme consistently mapped onto a single symbol. The symbols were taken from the BACS font, but do not occur in the experimental learning task.

The exposure and training phase was as described for the experimental task, however, we set the number of runs through the training phase to 3. The dependent variable, on which participants were matched across conditions, was the percentage of accurate responses in the last repetition of the block. In a different study, we found a wide range of scores across participants using this dependent variable (average accuracy = 45.5%, SD = 41.8), and high correlations with the number of repetitions needed to reach threshold accuracy in a different artificial orthography learning task, $r(18) = -0.77$, $p < 0.001$ (Schmalz, Schulte-Körne, de Simone, & Moll, 2021). There was no generalisation phase. The accuracy was scored directly by the experimenter. The experimenter assigned the participants to one of the two experimental orthographies, based on the percentage of accurate responses. Participants were matched pairwise on their performance on this task (accuracy, ±2 items) and their age (±3 years). The experimental procedure is further described in Figure 1.

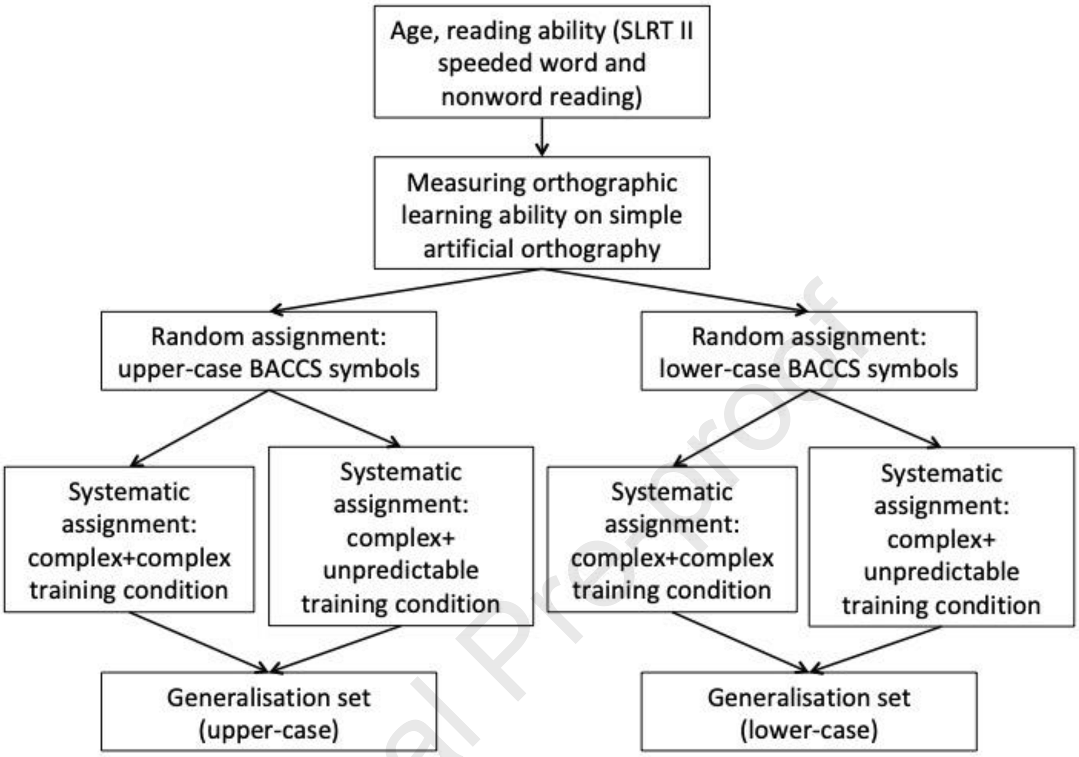

Age, reading ability (SLRT II speeded word and nonword reading)
Measuring orthographic learning ability on simple artificial orthography
Random assignment: upper-case BACCS symbols
Random assignment: lower-case BACCS symbols
Systematic assignment: complex+complex training condition
Systematic assignment: complex+ unpredictable training condition
Systematic assignment: complex+complex training condition
Systematic assignment: complex+ unpredictable training condition
Generalisation set (upper-case)
Generalisation set (lower-case)

Figure 1: The sequence of tasks in the proposed experiment, and the between-subject conditions. Note: The condition assignment involved systematically matching participants across group, while the potential nuisance variable of upper- versus lower-case BACCS font was assigned randomly.

The average accuracy on the matching orthography was 5.7 items out of 12 (SD = 4.4) for the complex+complex condition, and 5.8 out of 12 (SD = 4.7) for the complex+unpredictable condition. The averages ages for the two conditions were 23.7 years (SD = 4.3) and 22.6 (SD = 2.8), respectively. Out of the 22 participants in the complex+complex condition, 17 had a pair-wise match. One participant in the complex+complex condition did not pass the training phase for the experimental orthography while their matched participant did; and the reverse was true for 2 participants in the complex+unpredictable condition. Thus, we had a total of 37 participants who passed the training task, but only 14 pairwise-matched pairs. As the average performance on the matching task and the participants' ages were very similar across the two groups, we retained the data from all 37 participants in the group comparisons.

Other tasks

For explorative purposes, we also tested the participants' reading ability, with a standardised one-minute reading test, the SLRT-II (Moll & Landerl, 2010). The SLRT-II includes two lists, one with real words and one with pseudowords. Participants are given 1 minute for each list, and are instructed to read aloud as many words as possible. The reading score is the number of items read correctly within one minute. We also collected information about foreign language skills (which languages they know in addition to German, self-assessed proficiency).

Results

The script and data for the pre-registered hypotheses as well as the exploratory analyses can be found at https://osf.io/c3wkv/. Consistent with the pre-registration, we analyse the following measures as dependent variables:

A. Number of participants who were unable to reach the accuracy threshold of 70% after 10 repetitions of the training block.

B. Number of repetitions of the training block needed for the remaining participants to reach the accuracy threshold of 70%.

C. Transcriptions of the final repetition of the training block for each participant who successfully completed the training: Number of correct responses across conditions and for different types of items.

D. Transcriptions of the generalisation items: Number of responses which are correct (for consistent items) or plausible (one of the two possible pronunciations for the inconsistent items),

E. From the generalisation task, for the complex-predictable items: The percentage of plausible responses which were correct under the context-sensitive rule, and

F. For the unpredictable condition in the generalisation set: For each participant, the entropy, calculated as described on p. 14, using the proportion of Vowel Response A over the number of total plausible responses (Vowel Response A + Vowel Response B).

Hypothesis 1: Does adding an unpredictable component make the orthography more difficult to learn?

We proposed to test this hypothesis using two measures: (A) The number of participants who did not pass the training for each of the two conditions (using a χ^2-test), and (B) the number of repetitions needed for each of the remaining participants to reach the 70% accuracy threshold.

The pre-registered analysis on Measure (A), a χ^2-test, showed no significant difference between the proportion of participants who passed the training in the complex+complex compared to the complex+unpredictable condition, $\chi^2(1) < 0.001$, $p < 0.999$. However, it should be noted that, given the small number of participants who did not pass the training phase (<5), the data does not meet the assumptions for a χ^2-test (McHugh, 2013).

For Measure (B), participants in the complex+complex condition needed, on average, 5.1 repetitions (SD = 1.8) to reach the training criterion. For the participants in the complex+unpredictable condition, the average number of repetitions was 4.5

(SD = 1.6). We computed a Bayes Factor for the *t*-test in R (R Core Team, 2013), using the package BayesFactor (Morey & Rouder, 2014). We used the `ttestBF` command, with the default parameters (`rscale = "medium"`). The Bayes Factor in favour of the alternative hypothesis ($\mu_1 \neq \mu_2$) was 0.5. Numerically, the Bayes Factor therefore supports the hypothesis of no difference between the groups, however, it falls substantially above our pre-registered threshold at which we would have concluded that there is evidence for H_0, BF < 0.17.

Hypothesis 2: Are consistent items easier to learn than inconsistent items?

We transcribed the participants' responses for their last training block offline, and scored each response as correct or incorrect. We then calculated the overall accuracy percentage for each participant and each condition (Measure C). At 59%, the overall accuracy rate is lower than the expected 70%: this is because the recording was cut off for some of the participants' responses, when they took more than 5 seconds to pronounce a given pseudoword. In these cases, their response was scored as incorrect. Splitting up the items by their consistency (consistent versus either complex or unpredictable), the accuracy rate for consistent items is 80.3% (SD = 26.7), and for the inconsistent items 48.6% (SD = 22.7). A Bayes Factor analysis, with identical parameters to the previous section, showed overwhelming evidence for the alternative hypothesis of a consistency effect, BF > 600,000.

For the generalisation data, we calculated, for each participant and condition, the percentage of responses (out of 24) which were either correct or plausible. The overall percentage of accurate or plausible responses was 79.3%. For the consistent versus inconsistent (unpredictable, complex) items, the averages were 83.1% (SD = 18.9) and 78.1% (SD = 18.8), respectively. We performed a Bayes Factor analysis, with the same parameters as above. The Bayes Factor for the presence of a group differences is BF = 0.5. Thus, the Bayes Factor is numerically in favour of the absence of a group effect, but does not reach the critical threshold to allow us to draw inferences about the absence of an effect.

The accuracy for the training task and the percentage of accurate or plausible responses for the training and generalisation task, respectively, for each of the cells shown in Table 4, are summarised in Table 5.

Table 5: Percentage of responses which were accurate (training) and accurate or plausible (generalisation) across condition (see Table 4). By-participant SD in brackets.

Condition	**Consistent + Complex + Complex**	**Consistent + Complex + Unpredictable**
Training		
Consistent	73.8% (28.4)	88.2% (22.7)
Complex	48.5% (22.4)	57.7% (23.8)
Complex/unpredictable	45.5 (23.6)	43.4% (19.7)
Generalisation		
Consistent	80.6% (20.7)	86.4% (16.2)
Complex	79.6% (15.8)	88.3% (9.7)
Complex/unpredictable	65.2% (25.1)	81.4% (16.2)
Unpredictable (new body)	74.8% (19.8)	84.4% (11.2)

Hypothesis 3: Does unpredictability affect the learning of context-sensitive rules?

To address the third hypothesis, we calculated, for the generalisation data of the complex conditions, the percentage of context-appropriate responses out of the total number of plausible responses (Measure E). For the participant group that learned the complex+complex orthography, there are two such conditions; for the complex+unpredictable condition, there is only one. The average percentages of context-appropriate responses were 63.7% (SD = 12.5) and 55.8% (SD = 22.0) for the two complex conditions of the former group, and 55.0% (SD = 10.1) for the latter group. The "complex" condition of the latter group contains the same learned context-sensitive rule and items as the first "complex" condition of the former group: as per the pre-registration, we compare these equivalent conditions to each other. The Bayes Factor analysis (*t*-test, same parameters as above) provided numerical support for the presence of a group difference: BF = 2.0, though falling below the cut-off threshold of BF > 6. Numerically, the percentage of appropriate responses is higher for the complex+complex compared to the complex+unpredictable condition (63.7% vs. 55.0%).

Exploratory analyses. To further examine this data, we focus on individuals' data in a set of exploratory analyses. We calculated the percentage of participants who performed above chance in the two equivalent complexity conditions across groups for the generalisation data. This data is shown in Figure 2. In the complex+complex condition, 9 participants (45%) performed above chance. In the complex+unpredictable condition, 2 participants (13%) performed above chance. These results are in line with the numerical trend of better rule learning in the complex+complex condition.

To assess how reliable the percentage of appropriate responses is as a variable, we focussed on the participants from the complex+complex condition. We calculated the percentage of context-appropriate responses for both of the complex conditions. The correlation between them was $r(19) = 0.16$, $p = 0.5$, suggesting that this percentage is likely to be influenced by factors other than the context in which the context-sensitive rules are learned.

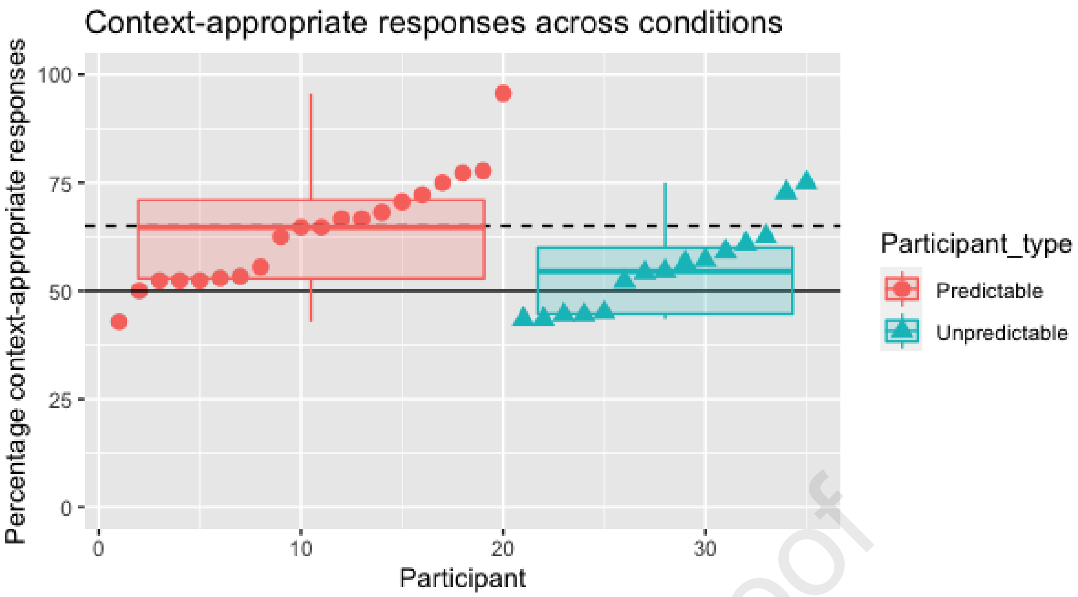


Figure 2: Individual participants' percentage of context-appropriate responses, for the complex+complex (predictable) and complex+unpredictable (unpredictable) conditions in the generalisation task. The line at 50% denotes chance level; the dashed line denotes the approximate cut-off above which each individual participant's performance is significantly above chance (it varies with the total number of trials on which each participant gave either a correct or a plausible vowel pronunciation).

Hypothesis 4: Do participants learn rules for unpredictable items?

For the final pre-registered hypothesis, we aim to assess how participants act in the generalisation phase when the pronunciation is unpredictable from the context. The first possibility is that they create a rule, and pronounce every instance of a given symbol identically. The second possibility is that they learn the distribution of the symbol's pronunciations; in this case, as the experiment was designed such that the two plausible pronunciations were equiprobable, participants should give each of the plausible pronunciations approximately 50% of the time. We examined the generalisation data using Entropy (Measure F), calculated from the plausible responses only. To calculate entropy, we split up the unpredictable conditions by the different vowels which they contained, because across vowel symbols, the plausible pronunciations were, by definition, different, which would inflate entropy even if participants consistently gave the same response to each of the symbols. We used the formula above to calculate the entropy for each participant, each vowel, and each condition, and then calculated the average across vowels for each participant and each condition. In the cases that, for a given vowel, a participant provided only incorrect (implausible) responses, the average was calculated from the entropy values for the remaining vowels.

The equivalent condition for the complex+complex and complex+unpredictable group consists of the "no body" items, where an inconsistent item was followed by a consonant with which it never occurred in the training set. Here, the complex+complex participants had an average entropy of 0.3 (SD = 0.3), and the complex+unpredictable participants had an average entropy of 0.5 (SD = 0.2). A Bayesian *t*-test (same parameters as above) provided equivocal evidence for a group difference, BF = 1.0. Figure 3 shows each individuals' entropy values for the no-body items.

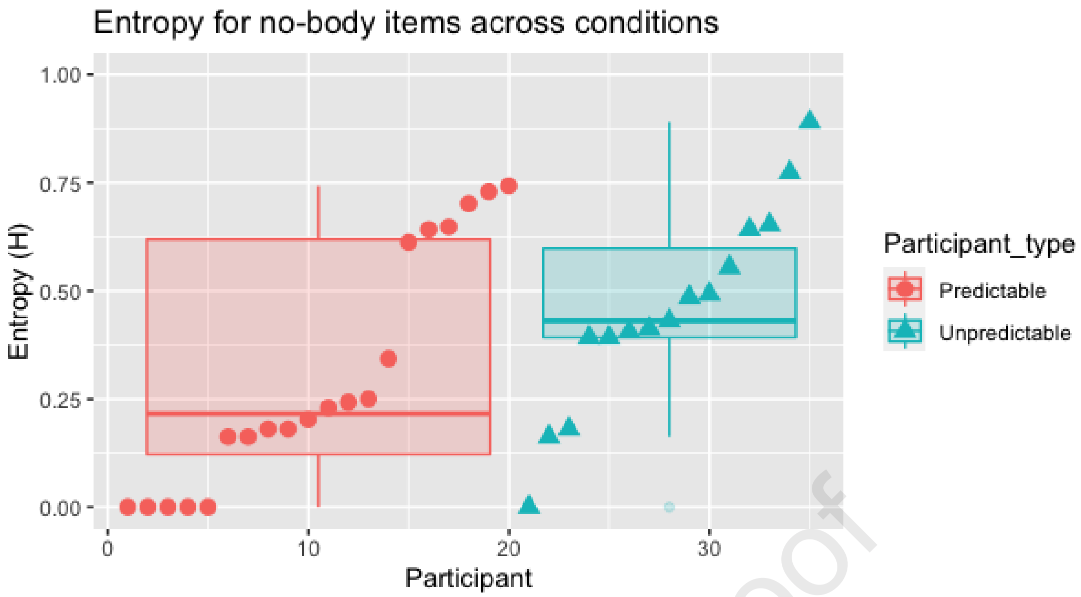


Figure 3: Entropy values for individual participants in the complex+complex (predictable) and complex+unpredictable (unpredictable) conditions.

Exploratory analyses

In a set of exploratory analyses, we aimed to assess if any participant-level variables explain variance in learning or generalisation performance. To this end, we performed Bayesian regression analyses. Each model was fitted using the default parameters of the `lmBF` function; then, we extracted the mean and 95% credibility interval (95% CI) of the posterior distribution using the `posterior` function (1000 iterations). In this exploratory context, we flag predictor variables as potentially worthy of future investigation if the 95% CI does not include the value of 0.

For all models, the independent variables were participant age (in years), their accuracy on the orthographic learning task that served as a matching task, whether they had been randomly assigned to the upper-case or lower-case version of the experimental artificial orthography learning task (coded as 1 or 0), their SLRT-II word reading score, SLRT-II pseudoword reading score, and condition (predictable coded as -1, unpredictable as +1). The dependent variables were, in turn, the number of repetitions needed (Measure B), number of plausible responses in the generalisation task (Measure D), percentage of plausible responses for the matched complex items (Measure E), and entropy for the new-body items (Measure F). The only model containing a 95% CI which did not overlap with zero was the one using the number of repetitions (Measure B) as the dependent variables, where the accuracy on the matching orthography emerged as a predictor, with participants who had higher accuracy needing fewer repetitions of the training block in the experimental task to reach threshold; mean = -0.14, 95% CI = [-0.25, -0.03].

An unexpected finding is the numerically higher performance in the generalisation task for participants who learned the unpredictable correspondences compared to those who learned only predictable correspondences (see Table 5). This is in line with the findings from Pilot Experiment 2, where the between-group comparison was significant in the same direction. To follow up on this finding, we compared a series of Bayes Factor linear mixed effect models. Here, accuracy was predicted by item type (consistent, complex, unpredictable) and group (complex+complex versus complex+unpredictable). Participant was included as a random factor. First, we compared an interactive model (item type * participant group) to an additive model (item type + participant group): here, we found evidence against the interactive model, BF = 0.26 (± 2.88%). Then, we compared the additive model to one excluding the main effect of item type. Here, we found evidence against the additive model and thus against an effect of item type, BF = 0.16 (±1.54%). Critically, we compared the additive model against a model excluding the effect of participant group. Here, we found evidence for the additive model and thus for the presence of a group effect, BF = 15.89 (± 1.54).

General discussion

The pre-registered study set out to test four hypotheses, which were partially derived from the results of the pilot experiments. In the first hypothesis, we tested whether adding an unpredictable component makes an orthography more difficult to learn. We did not find any evidence for this hypothesis, as participants whose orthography contained unpredictable items did not take longer to learn the orthography compared to participants learning an orthography where the pronunciation of all inconsistent items followed a predictable rule. In the second hypothesis, we aimed to assess if inconsistent items (i.e., items where a symbol has two possible pronunciations) are more difficult to learn than consistent items. Here, we found overwhelming evidence that this is indeed the case, which also replicating the findings of the two pilot experiments.

In the third hypothesis, we assessed whether adding an unpredictable aspect to the orthography prevents participants from learning a context-sensitive rule. Numerically, participants were less likely to give context-appropriate responses when their orthography contained an unpredictable component, but the Bayes Factor did not provide strong evidence for the presence of a group difference. This may be because of the limited number of participants. Power calculations are a frequentist concept and not central to inference using Bayes Factor. Bayes Factors should provide clearer evidence with an increasing number of participants, or with an increase in the effect size, *ceteris paribus*. Unfortunately, collecting more data was unfeasible due to funding constraints.

In the fourth hypothesis, we assessed whether participants create rules for symbols with unpredictable pronunciations or instead learn the distribution of the input. Here, as in the pilot experiments, we found variability between the participants.

We also found a numerical trend with participants who had an unpredictable component in their orthography being less likely to create a rule, but, again, the Bayes Factor did not provide strong evidence for this group difference.

A striking difference between the pilot experiments and pre-registered study was the number of participants who did not pass the training phase (Pilot Experiment 1: 35%, Pilot Experiment 2: 32%, Pre-registered Experiment: 8%). There were numerous differences between the experiments. The most likely explanation for the difference in training performance is the number of items: the final experiment had more items, meaning that the participants had more exposure to the symbol-sound correspondences in a given training block. There may have been differences in the motivation of the participants, as the pilot studies did not receive payment for their participation. Another major difference between the experiments is language, as the first two experiment were conducted with Italian speakers and the final one with German speakers. Both orthographies are considered to be shallow: they contain some complex rules. However, in German, there is additional ambiguity in vowel length pronunciation (e.g., “Mond” [moon] is pronounced with a long vowel, /mo:nt/ and “blond” [blonde] is pronounced with a short vowel, /blɔnt/). Thus, it is possible that the cognitive system of German readers is better equipped to deal with unpredictability in an orthography. *A priori*, we did not expect a difference due to the participants’ languages: In a previous artificial orthography learning experiment with German participants and similar characteristics as the pilot experiments here, we found that 20% of participants were unable to learn the orthography (Schmalz et al., 2021). Thus, the performance of the participants was closer to the Italian experiments reported here, rather than to the final German experiment, making it unlikely that language is the only determining factor.

Overall and across two experiments, participants who learned unpredictable correspondences performed better in the generalisation task than participants who learned only predictable correspondences. This is contrary to what one would expect, intuitively. We have no straight-forward explanation for this finding. One possibility is that participants in the unpredictable conditions need more training before reaching the required accuracy threshold. Though we did not find any significant differences between the participant groups in the length of training, participants in Experiment 3 needed numerically more training if they were in the unpredictable condition (5.1 runs through on average, versus 4.5 for the complex condition). In Experiment 2, this difference was negligible (6.3 for the unpredictable condition versus 6.6 for the predictable condition). Therefore, this is an unlikely explanation for our results. Rather than speculating about other potential reasons, we leave this as an open question for future research.

Item-level, participant-level, and orthography-level effects

The only clear result of the current study is a consistency effect: when a symbol has two possible pronunciations, the item is more difficult to learn. Here, the reason for the inconsistency does not seem to matter: across three experiments, we showed this consistency effect both for predictable-inconsistent and for unpredictable-inconsistent items. Bringing this finding back to the literature on reading, this finding is in line with findings of a consistency effect for reading real words or pseudowords: when a written stimulus contains an orthographic cluster with more than one possible pronunciation, participants take longer to respond or are more likely to give an inaccurate response (e.g., Jared, 2002).

The consistency effect is an item-level variable, as the presence of two or more pronunciations of a given orthographic unit is a characteristic of each individual item. It is noteworthy that this variable had by far the clearest effect on the orthography acquisition process. In contrast, participant-level variables such as reading ability or age had no noticeable effect on learning or generalisation[3]. Also the effects of group assignment, which was an orthography-level variable, were more elusive than the effect of consistency. Thus, item-level variables seem to have stronger effects on learnability than either orthography-level or participant-level variables. Extending this observation to learning to read in the real world, this is relevant for the question of why it is more difficult to learn to read in deep orthographies, which are characterised by inconsistency of grapheme-phoneme correspondences, compared to shallow orthographies (e.g., Seymour et al., 2003). There are two possibilities for the effects of orthographic depth on reading acquisition (which are not mutually exclusive): One is that the presence of inconsistency places a heavier demand on the cognitive system, and qualitatively changes the way in which readers learn or process words, for example by readers adopting a more whole-word based reading strategy, where even the reading of consistent words is influenced only minimally by phonological decoding (Katz & Frost, 1992). The second is that words in deep orthographies are more likely to contain inconsistent correspondences, which makes them more difficult to learn or to read (De Simone, Beyersmann, Mulatti, Mirault, & Schmalz, 2021; Egan, Oppenheim, Saville, Moll, & Jones, 2019). The strong effect of the item-level variable of consistency provides support for the second possibility. The orthography-level effects were less clear; however, taking together the results of the three experiments, some consistent patterns emerge, which leads us

[3] Note that we did not look at differential effects of gender, as we do not expect systematic differences in the way in which orthographies are learned as a function of gender.

to speculate that the first possibility also holds true to some extent, as discussed below.

The orthography-level variable of group assignment seems to have much smaller effects on artificial orthography learning compared to item-level consistency. For none of the effects did the Bayes Factor exceed the pre-registered threshold for drawing strong conclusions, therefore our discussion of these is based on numerical trends and serves as a suggestion for future research. In the generalisation data, we found (i) that participants whose orthography had an unpredictable component were numerically less likely to learn the context-sensitive rule, and (ii) participants whose orthography had the unpredictable component were also less likely to assign a rule to the unpredictable correspondence, as shown by greater variability of vowel pronunciations that they assigned to a given symbol when its pronunciation was unpredictable. These two findings have in common the notion that participants whose orthography was less systematic were less likely to create rules. This is in line with our speculations following the two pilot experiments, where (i) participants were more likely to learn the context-sensitive rule in a between-subject design, where all correspondences were predictable, than in a within-subject design, where the item types were mixed, and (ii) participants had predominantly low entropy values when complex rules were present, and entropy values close to 1 when the pronunciation of inconsistent vowels was completely unpredictable.

The findings above suggest that systematicity of an orthography encourages rule formation, be it of simple or of complex rules. Bringing this back to the original broad aim of this study, this allows us to speculate about an effect that unpredictability, but not complexity exerts on the learning of a novel orthography:

namely, it impairs the build-up of rules and leads readers to learn the probabilistic distribution of symbol-sound correspondences instead.

Distributional learning and rule extraction: Modelling of parallel processing

In the literature on reading, rule knowledge is implemented in the Dual-Route Cascaded (DRC) model of reading (Coltheart et al., 2001), whereas probabilistic learning of inconsistent print-to-speech correspondences defines knowledge in connectionist models (Plaut et al., 1996). Our data supports neither model unequivocally: instead, it suggests that both models may reflect the learning of grapheme-phoneme correspondences under different circumstances.

Below, we propose a theoretical model which could explain our data. Specifically, our data could be explained by a model where both rule-extraction and distributional learning processes work in parallel. We model knowledge gain during the training phase. We assume that knowledge is gained about the symbol-sound correspondences. We remain agnostic about how exactly this knowledge can be measured: while the purpose of the generalisation task is to measure such symbol-sound knowledge, it remains unclear how it reflects graded knowledge in the cognitive system.

Assuming a distributional learning model, learning of symbol-sound correspondences is likely to take place on a token basis: with each encounter of the symbol-sound correspondence, the cognitive system updates its knowledge about the distribution of pronunciations. Once all tokens are encountered, the state of knowledge reaches a plateau (s), as the distributions of all correspondences are known. To demonstrate this mathematically, distributional learning (D) as a function of token encounters (t) can be modelled as:

$$D(t) = \begin{cases} t \text{ if } t < s \\ s \text{ if } t \geq s \end{cases}$$

This is likely to be different for rule learning: learning a single rule leads to a jump in knowledge rather than a linear increase associated with continuous upgrading of the distributional knowledge. Thus, initial learning due to rule extraction should lead to faster knowledge gain than for distributional learning, because a rule can be immediately applied to any other instances of the same symbol. When an orthography is complex, the knowledge gain should be slower compared to a consistent orthography, because the learning of a simple rule leaves out knowledge about the complex rules.

Furthermore, rule learning is confined by the usefulness of the rules of the orthography: when the symbol-sound correspondences have some level of unpredictability, rule knowledge cannot result in comprehensive knowledge as distributional knowledge; thus, the plateau cannot be reached. Thus, rule learning (R), as a function of token encounters (*t*) can be modelled by the arctangent function, because it has the characteristic of increasing with positive values of *t* and of the steepness (R*'*) decreasing with an increase in *t*. We can model rule learning with additional two parameters, *c* and *a*, which correspond to the degree of unpredictability and complexity, respectively:

$$R(t) \ = \ c \ \tan^{-1}(at)$$

The parameter *c* defines the height of the plateau. Low values of *c* correspond to high unpredictability; as $c \ \rightarrow \ \frac{s}{\pi/2}$, we approach a completely predictable system. The parameter *a* defines the steepness of the curve for initial values of *t*, and may thus correspond to complexity: when an orthography has low complexity, the rules are easy to extract, and thus the learning process will be fast until it reaches the plateau.

Figure 4 shows the learning performance for a set of hypothetical parameters: *s* = 15; for the simple, predictable system, $c \ = \ \frac{s}{\pi/2}$ and *a* = 2; for the complex, predictable system, $c = \frac{s}{\pi/2}$ and a = 0.5; and for the complex, unpredictable system, c = 6 and a = 0.5. The figure shows that, initially, the rule-based process overtakes the level of knowledge of the distributional learning process. However, when unpredictability is high, the distributional process provides more complete knowledge at comparatively lower numbers of token encounters. If the learning process is interrupted at a point t when the distributional line is above the rule-extraction line (e.g., at t = 10 in the figure below), distributional knowledge is more complete and may thus be preferred to make inferences in a generalisation task. This is more likely to be the case for low values of c, and thus low unpredictability.

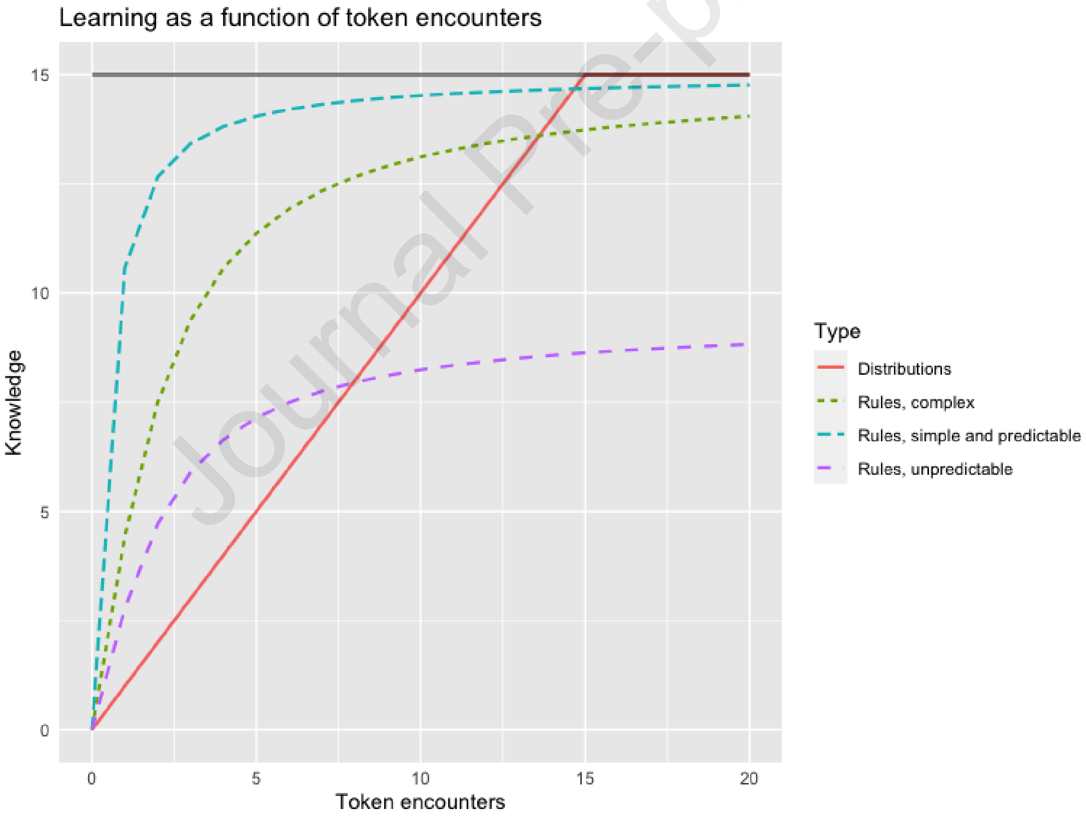


Figure 4: State of knowledge for a distrubutional learning process or rule extraction with different parameters corresponding to different levels of complexity and unpredictability. The black horizontal line corresponds to s = 15. The scales are abitrary. Note that the arctan lines with the same predictability paramater will

converge as $t \rightarrow \infty$. The code to generate the figure can be found at https://osf.io/pv6t9/.

Differential effects of complexity and unpredictability?

The original broad question that we set out to answer is whether complexity and unpredictability of an orthographic system have different effects on learning to read (Schmalz et al., 2015). From a practical perspective, this will help us to better predict in which orthographies children are more likely to struggle to read, and the cognitive processes that are especially taxed by increased processing difficulty associated with the orthography's characteristics.

In our artificial orthography learning experiment, inconsistency affected learnability of an orthography, regardless the reason for the multiple possible pronunciations. This is likely to be true also for the initial stages of reading acquisition: as outlined in the introduction, complexity should increase the amount of time needed to learn the correspondences, because it increases the number of rules. A question for further research is whether complexity continues to exert a negative effect on reading acquisition at later stages. If the learning of grapheme-phoneme correspondences by rule extraction can be, indeed, modelled by an arctangent function, the *a* parameter affects the steepness of the slope at the beginning stages of acquisition, but has a negligible effect once we approach *s*. This leads to several empirical predictions: in a pseudoword reading task, participants should be able to consistently apply complex rules and give context-appropriate responses. This seems to be, indeed the case (e.g., Martensen, Maris, & Dijkstra, 2003; Schmalz, Porshnev, & Marinus, 2017). However, evidence suggests that complexity continues to exert an effect in reading aloud latencies, even in skilled readers (e.g., Rastle & Coltheart, 1998; but see Schmalz, Beyersmann, Cavalli, & Marinus, 2016). Such an effect on latencies does not contradict the possibility that skilled readers approach ceiling knowledge of context-sensitive rules: the explanation for the effect on reading aloud latencies is that there is a conflict between the pronunciation of the single letters and of the multi-letter cluster. The high accuracy of participants reading complex pseudowords shows that participants are able to use their knowledge of context-sensitive rules to resolve such a conflict. Predictions could be made for direct cross-linguistic comparisons: overall accuracy should be lower for children learning to read in a complex orthography, such as French, compared to simple orthographies, such as Italian. The difference in accuracy should decrease with increasing reading experience. Unfortunately, we know of no data set which would allow us to test this prediction.

During the initial symbol-sound correspondence learning process which we modelled in our experimental task, we found numerical trends suggesting a differential effect of an orthography's predictability when it comes to using rule-based knowledge as opposed to distributional knowledge: when the orthography contained unpredictability, participants were numerically less likely to create rules either for simple or for complex correspondences. This finding should be followed up either in a larger experiment or with a stronger manipulation. Such an experiment could be performed in an applied setting, with children at the beginning of reading acquisition. Here, knowledge of grapheme-phoneme correspondences could be approximated by performance on a generalisation task: in the case of reading a real orthography, this would be a pseudoword reading aloud task. Distributional versus rule learning could be assessed by pseudoword reading aloud entropy (De Simone et al., 2021): if participants create rules, they should have low entropy by pronouncing a given orthograph cluster in the same way across pseudowords. In a cross-linguistic study, the model in Figure 4 would predict that during the initial stages of reading acquisition, all children should show low entropy as they rely on extracted rules; at later stages, cross-linguistic differences associated with predictability should emerge, with children reading in unpredictable orthographies relying on distributional knowledge and therefore showing greater variability in their pronunciations of a given unit, which should be reflected as higher entropy. Children reading in less unpredictable orthographies should rely on extracted rules and provide consistent responses, even to orthographic units with inconsistent and unpredictable pronunciations. While we know of no study that has tested this prediction in children, we have previously reported that, in adults, pseudoword reading aloud entropy is, indeed, affected by unpredictability, but not complexity (De Simone et al., 2021). A future study aiming to test this prediction should take into account reading instruction: explicit teaching of simple and complex rules may affect the *a* (complexity) parameter of the arctan function.

The focus of our experimental task was on the initial stages of reading acquisition. As such, we also have no information about higher-order effect, such as semantic processing or compensatory strategies. This is an avenue for future research. Recent findings already suggest a differential effect of predictability, but not complexity, in a sentence reading task (Schroeder et al., 2021): children reading in English (complex, unpredictable) seemed to rely more on sentence context compared to children reading in German (complex, predictable) and Finnish (simple, predictable). However, it is unclear whether this finding is due to a general delay in reading acquisition associated with increased difficulty of learning to read in English, or if unpredictability across all age groups encourages semantic (OSP) processing due to the limited information that can be obtained using the OP-route. These two possibilities could be dissociated in an experiment examining orthographic-semantic marker effects during reading acquisition in predictable versus unpredictable orthographies: if these effects are stronger in unpredictable orthographies, we can conclude that unpredictability encourages greater reliance on semantics. Conversely, if they are weaker in unpredictable orthographies, this would suggest a general delay of both the OSP and OS routes.

References

Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics, 24*, 621-635.

Asfaha, Y. M., Kurvers, J., & Kroon, S. (2009). Grain size in script and teaching: Literacy acquisition in Ge'ez and Latin. *Applied Psycholinguistics, 30*(04), 709. doi:10.1017/s0142716409990087

Bitan, T., & Karni, A. (2003). Alphabetical knowledge from whole words training: effects of explicit instruction and implicit experience on learning script segmentation. *Cognitive Brain Research*, *16*(3), 323-337.

Blomert, L. (2011). The neural signature of orthographic–phonological binding in successful and failing reading development. *Neuroimage*, *57*(3), 695-703.

Caravolas, M., Leråg, A., Mousikou, P., Efrim, C., Litavsky, M., Onochie-Quintanilla, E., . . . Hulme, C. (2012). Common Patterns of Prediction of Literacy Development in Different Alphabetic Orthographies. *Psychological Science, 23*(6), 678-686. doi:10.1177/0956797611434536

Caravolas, M., Lervag, A., Defior, S., Malkova, G. S., & Hulme, C. (2013). Different Patterns, but Equivalent Predictors, of Growth in Reading in Consistent and Inconsistent Orthographies. *Psychological Science, 24*(8), 1398-1407. doi:10.1177/0956797612473122

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. doi:10.1037//0033-295x.108.1.204

De Simone, E., Beyersmann, E., Mulatti, C., Mirault, J., & Schmalz, X. (2021). Order among chaos: Cross-linguistic differences and developmental trajectories in pseudoword reading aloud using pronunciation Entropy. *PloS one*, *16*(5), e0251629.

Egan, C., Oppenheim, G. M., Saville, C., Moll, K., & Jones, M. W. (2019). Bilinguals apply language-specific grain sizes during sentence reading. *Cognition*, *193*, 104018.

Forster, K. I., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instuments & Computers, 35*, 116-124.

Frith, U., Wimmer, H., & Landerl, K. (1998). Differences in Phonological Recoding in German- and English-Speaking Children. *Scientific Studies of Reading, 2*(1), 31-54.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for Visual Word Recognition and Orthographic Depth: A Multilingual Comparison. *Journal of Experimental Psychology: Human Perception & Performance*, *13*(1), 104-115.

Hutzler, F., Ziegler, J., Perry, C., Wimmer, H., & Zorzi, M. (2004). Do current connectionist learning models account for reading development in different languages? *Cognition*, *91*, 273-296.

Jackson, N., & Coltheart, M. (2001). *Routes to reading success and failure: Toward an integrated cognitive psychology of atypical reading.* New York, NY: Psychology Press.

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. *Journal of Memory and Language, 46*(4), 723-750.

Katz, L., & Frost, R. (1992). The Reading Process is Different for Different Orthographies: The Orthographic Depth Hypothesis. In R. Frost & L. Katz (Eds.), *Orthography, Phonology, Morphology, and Meaning* (pp. 67-84). Amsterdam: Elsevier Science Publishers.

Kempe, V., & MacWhinney, B. (1998). The acquisition of case marking by adult learners of Russian and German. *Studies in Second Language Acquisition, 20*(04), 543-587.

Landerl, K. (2000). Influences of orthographic consistency and reading instruction on the development of nonword reading skills. *European Journal of Psychology of Education, 15*(3), 239-257. Retrieved from <Go to ISI>://000169003200001

Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppanen, P. H. T., Lohvansuu, K., . . . Schulte-Korne, G. (2013). Predictors of developmental dyslexia in European orthographies with varying complexity. *Journal of Child Psychology and Psychiatry, 54*(6), 686-694. doi:10.1111/Jcpp.12029

Marinus, E., Nation, K., & de Jong, P. (2015). Density and length in the neighbourhood: Explaining cross-linguistic differences in learning to read in English and Dutch. *Journal of Experimental Child Psychology, 139*, 127-147. doi:10.1016/j.jecp.2015.05.006

Martensen, H., Maris, E., & Dijkstra, T. (2003). Phonological ambiguity and context sensitivity: On sublexical clustering in visual word recognition. *Journal of Memory and Language, 49*(3), 375-395.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica, 23*(2), 143-149.

Moll, K., & Landerl, K. (2010). SLRT-II: Lese- und Rechtschreibtest; Weiterentwicklyng des Salzburger Lese- und Rechtschreibtests (SLRT). In: Huber.

Moll, K., Ramus, F., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., . . . Landerl, K. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learning and Instruction, 29*, 65-77. doi:10.1016/J.Learninstruc.2013.09.003

Morey, R. D., & Rouder, J. N. (2014). Package "BayesFactor". Retrieved from http://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf

Perry, C., Ziegler, J., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. *Psychological Review, 114*(2), 273-315. doi:10.1037/0033-295X.114.2.273

Perry, C., Ziegler, J., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology, 61*(2), 106-151.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56-115. doi:10.1037/0033-295x.103.1.56

R Core Team, R. (2013). R: A language environment for statistical computing [Computer software manual]. Vienna. Retrieved from http://www.R-project.org/

Rastle, K., & Coltheart, M. (1998). Whammies and double whammies: The effect of length on nonword reading. *Psychonomic bulletin & review, 5*(2), 277-282.

Rey, A., Jacobs, A. M., Schmidt-Weigand, F., & Ziegler, J. C. (1998). A phoneme effect in visual word recognition. *Cognition, 68*(3), B71-B80. doi:10.1016/S0010-0277(98)00051-1

Rey, A., & Schiller, N. O. (2005). Graphemic complexity and multiple print-to-sound associations in visual word recognition. *Memory & Cognition, 33*(1), 76-85. doi:10.3758/Bf03195298

Rouder, J. N., Speckman, P. L., Sun, D. C., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review, 16*(2), 225-237. doi:10.3758/Pbr.16.2.225

Schmalz, X., Beyersmann, E., Cavalli, E., & Marinus, E. (2016). Unpredictability and complexity of print-to-speech correspondences increase reliance on lexical processes: More evidence for the orthographic depth hypothesis. *Journal of Cognitive Psychology, 28*(6), 658-672.

Schmalz, X., Marinus, E., Coltheart, M., & Castles, A. (2015). Getting to the bottom of orthographic depth. *Psychonomic Bulletin and Review, 22*(6), 1614-1629. doi:10.3758/s13423-015-0835-2

Schmalz, X., Porshnev, A., & Marinus, E. (2017). Two distinct parsing stages in nonword reading aloud: Evidence from Russian. *Quarterly Journal of Experimental Psychology, 70*(12), 2548-2559.

Schmalz, X., Schulte-Körne, G., de Simone, E., & Moll, K. (2021). What Do Artificial Orthography Learning Tasks Actually Measure? Correlations Within and Across Tasks. *Journal of cognition, 4*(1).

Schroeder, S., Häikiö, T., Pagan, A., Dickins, J. H., Hyona, J., & Liversedge, S. P. (2021). Eye Movements of Children and Adults Reading in Three Different Orthographies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Seidenberg, M., & Plaut, D. C. (2014). Quasiregularity and its discontents: the legacy of the past tense debate. *Cognitive Science, 38*(6), 1190-1228.

Seymour, P., Aro, M., & Erskine, J. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143-174.

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition, 177*, 198-213.

Tabossi, P., & Laghi, L. (1992). Semantic priming in the pronunciation of words in two writing systems: Italian and English. *Memory & Cognition, 20*(3), 303-313.

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology, 79,* 1-39.

Taylor, J. S. H., Plunkett, K., & Nation, K. (2011). The influence of consistency, frequency, and semantics on learning to read: An artificial orthography paradigm. *J Exp Psychol Learn, 37*(1), 60-76. doi:10.1037/A0020126

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. (1995). The Special Role of Rimes in the Description, Use, and Acquisition of English Orthography. *Journal of Experimental Psychology: General, 124*(2), 107-136.

van den Bosch, A., Content, A., Daelemans, W., & de Gelder, B. (1994). Measuring the complexity of writing systems. *Journal of Quantitative Linguistics, 1*(3), 178-188.

Venezky, R. L. (1970). *The structure of English orthography* (Vol. 82): Walter de Gruyter.

Vidal, C., Content, A., & Chetail, F. (2017). BACS: The Brussels Artificial Character Sets for studies in cognitive psychology and neuroscience. *Behavior research methods,* 1-20.

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language, 65*(1), 1-14.

Yoncheva, Y. N., Wise, J., & McCandliss, B. (2015). Hemispheric specialization for visual words is shaped by attention to sublexical units during initial learning. *Brain and language, 145,* 23-33.

Acknowledgements

We would like to thank Erik Visser and Giovanni Carli for data collection for the pilot experiments, Veronika Jäger for recording the stimuli for the proposed experiment, and Marion Olbricht for data collection of Experiment 3. We thank Ondřej Kudláček for his advice on appropriate functions. X.S. was supported by a grant from the Deutsche Forschungsgemeinschaft (SCHM3450/1-1) and by the Programme “Fellows Freies Wissen” from the Stifterverband, Wikimedia, and Volkswagenstiftung (H190 5909 5096 32135). Declarations of interest: None.