Journal of Experimental Psychology: Learning, Memory, and Cognition

Letter–Sound Inconsistency Impacts Word Learning and Forgetting

Yani Qiu and J. S. H. Taylor

Online First Publication, October 2, 2025. https://dx.doi.org/10.1037/xlm0001522

CITATION

Qiu, Y., & Taylor, J. S. H. (2025). Letter–sound inconsistency impacts word learning and forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. https://dx.doi.org/10.1037/xlm0001522

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: Learning, Memory, and Cognition

ISSN: 0278-7393

https://doi.org/10.1037/xlm0001522

Letter–Sound Inconsistency Impacts Word Learning and Forgetting

Yani Qiu and J. S. H. Taylor
Department of Language and Cognition, Division of Psychology and Language Sciences, University College London

In alphabetic writing systems, letters and sounds have systematic mapping relations. Words that display common letter–sound relations are high in consistency (e.g., "speak," "weak"; consistent words), whereas those that use less common relations are low in consistency ("break"; inconsistent words). This study tested how letter–sound consistency affects word learning (Experiment 1) and forgetting (Experiment 2), considering various aspects of lexical knowledge, including orthography (O), phonology (P), semantics (S), and bindings between them (P–O, S–O, S–P). Eighty-six native English-speaking adults learned novel meanings for eight spoken pseudowords and then read sentences containing the written forms of these pseudowords. Half the pseudowords were consistent, whereas the other half were inconsistent. Knowledge of the pseudowords was tested immediately after learning (Experiment 1; $N = 86$) and with a delay ($M = 77$ days; Experiment 2; $N = 58$). Results showed that inconsistency impaired learning of most aspects of lexical knowledge (P, P–O, S–O, and S–P). After the delay, participants also showed more forgetting of P, but interestingly less forgetting of S–P, for inconsistent relative to consistent items. Together, these findings revealed that lexical development is a complex, interactive, and dynamic process.

Keywords: spelling–sound regularity, reading development, lexical quality, self-teaching, memory retention

Supplemental materials: https://doi.org/10.1037/xlm0001522.supp

According to the lexical quality hypothesis (Perfetti, 2007; Perfetti & Hart, 2002), a word is a triplet of orthography (e.g., "peak"), phonology (e.g., /piːk/), and semantics (e.g., the top of a mountain). One characteristic of high lexical quality is precision, that is, fully specified orthographic, phonological, and semantic representations of the word, as well as tight bindings between these representations (orthography–phonology, orthography–semantics, semantics–phonology bindings), as represented in Figure 1. This allows words to be accurately and rapidly retrieved. This in turn lays a foundation for efficient comprehension (Perfetti, 2007; Perfetti & Hart, 2002) and communication (Koizumi, 2013), which has implications for educational attainment, occupational outcomes, and well-being (Johnson et al., 2010; Olsson et al., 2013).

In alphabetic writing systems, relations between orthography and phonology have statistical regularities. The dual-route cascaded model (Coltheart et al., 1993, 2001) assumes a rule system that governs these regularities at the level of grapheme–phoneme correspondences. However, research shows that human readers (except those in the very beginning stages of learning) consider the context in which graphemes and phonemes occur and are sensitive to

L. Robert Slevc served as action editor.
Yani Qiu https://orcid.org/0000-0002-3997-2275
J. S. H. Taylor https://orcid.org/0000-0002-1109-8539
This project was supported by the University College London Studentship in the Division of Psychology and Language Sciences awarded to Yani Qiu. The authors thank Jennifer Rodd who suggested conducting a follow-up study as in Experiment 2.

Yani Qiu played a lead role in conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, visualization, and writing–original draft, a supporting role in writing–review and editing, and an equal role in resources. J. S. H. Taylor played a lead role in supervision and writing–review and editing, a supporting role in conceptualization, formal analysis, investigation, methodology, project administration, visualization, and writing–original draft and an equal role in resources.

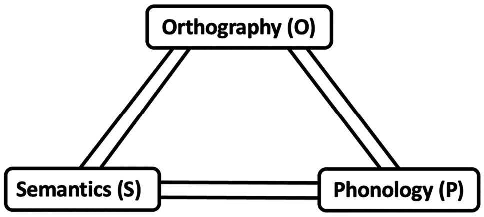
The data are available on Experiment 1 at https://osf.io/dqcbw/; Experiment 2 at https://osf.io/6ty32/.

The experimental materials are available on Experiment 1 at https://osf.io/dqcbw/; Experiment 2 at https://osf.io/6ty32/.

The preregistered design is available on Experiment 1 at https://osf.io/zbgja; Experiment 2 at https://osf.io/gknd2.

Correspondence concerning this article should be addressed to Yani Qiu, Department of Language and Cognition, Division of Psychology and Language Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom. Email: ucjtyqi@ucl.ac.uk

Figure 1
Illustration of a High-Quality Lexical Entry



regularities at the body-rime level (i.e., vowel and the following consonant) that are commonly observed in words (Chee et al., 2020; Jared, 2002; Treiman et al., 1995). For example, letters "ea," when followed by "k," are commonly pronounced as /iː/ and the sound /iː/, when followed by /k/, is commonly spelled as "ea," as in "speak" and "weak." We can describe words that display common letter–sound relations as high in consistency (e.g., "speak," "weak"). In contrast, words that display less common letter–sound relations are low in consistency. For example, in the word "break," letters "ea," followed by "k," are pronounced as /eɪ/ instead. Moreover, the sound /eɪ/, when followed by /k/, is commonly spelled as "ake" as in "cake" and "fake," but is spelled as "eak" in "break." Henceforth, we will use the term "consistency" to refer to a general estimate of letter–sound statistical regularities, encompassing both the grapheme–phoneme and body-rime measures used in previous research.

It is well-documented that consistency affects lexical processing. In reading aloud (measuring orthography-to-phonology bindings), readers made more errors and took longer to read inconsistent than consistent words (Jared, 1997; Jared et al., 1990; Treiman et al., 1995; Weekes et al., 2006). In spelling (measuring phonology-to-orthography bindings), higher error rates were observed for inconsistent words (Brown et al., 1991; Nation, 1997; Weekes et al., 2006). In auditory lexical decision (measuring phonological representations), participants showed longer response times and more errors when trying to recognize inconsistent words compared with consistent words (Ziegler et al., 2004, 2008; Ziegler & Ferrand, 1998). However, the consistency effect has not been reliably observed in visual lexical decision (measuring orthographic representations; Jared, 1997; Jared et al., 1990; Stone et al., 1997; Ziegler et al., 2008). The effect of consistency on semantic knowledge (semantics, semantics–orthography bindings, semantics–phonology bindings) is understudied.

The studies just reviewed measured participants' lexical knowledge at the time of testing, which is the result of accumulated prior experiences and processing of the words. They did not consider processes that are involved in the development of lexical knowledge. How does consistency affect the development of lexical knowledge? Training studies showed that consistency impacts learning of orthography–phonology bindings (Burt & Blackwell, 2008; McKay et al., 2008; Murray et al., 2022; Wang et al., 2012, 2013). In these studies, participants were instructed to learn pseudowords, some of which were consistent, whereas others were inconsistent. After learning, participants were asked to read aloud written words (measuring orthography-to-phonology bindings) or to spell auditorily presented words (measuring phonology-to-orthography bindings). Results showed that participants developed more accurate orthography-to-phonology and phonology-to-orthography bindings for consistent words than for inconsistent words.

There is some evidence that consistency affects learning of orthographic representations. In Wang et al. (2011, 2012), participants were instructed to indicate whether visually presented items were trained words or not. Consistent items were recognized more accurately than inconsistent words, suggesting that participants developed more precise orthographic representations for consistent than inconsistent words. However, this effect is not always observed (McKay et al., 2008; Taylor et al., 2011; Wang et al., 2013).

No study has investigated how consistency would affect learning of other aspects of lexical knowledge, that is, phonology, semantics, semantic–orthography, and semantics–phonology bindings. A relevant study is Hulme et al. (2022), which manipulated letter-to-sound decoding ease in pseudowords. Decoding ease was defined on the basis of reading aloud variability among 41 adults (Mousikou et al., 2017), such that an easy-to-decode item (e.g., "bamper") received a single pronunciation from readers, whereas a hard-to-decode item (e.g., "uzide") was read aloud in various ways. In Hulme et al. (2022), participants were instructed to silently read paragraphs that presented written pseudowords and described their meanings. Half the pseudowords were easy to decode and the other half were hard to decode. Posttests showed that easy-to-decode items were learned with more accurate orthographic representations, orthography-to-phonology, and semantics-to-phonology bindings, compared with hard-to-decode items. Learning of orthography-to-semantics bindings was not affected by decoding ease. Phonology and semantics were not tested. Although both decoding ease and consistency concern letter–sound relations, they are different in important ways. Consistency as defined here depends on both letter-to-sound and sound-to-letter directions. In contrast, decoding ease was shown to depend on letter-to-sound consistency as well as neighborhood size (the number of words that differ from the pseudoword by one letter; Mousikou et al., 2017). Therefore, it remains unclear how consistency would affect learning of phonology, semantics, semantics–orthography, and semantics–phonology bindings.

The development of lexical knowledge is a dynamic process. Learners acquire lexical knowledge, but the knowledge is subject to forgetting, as evidenced by a decrease in recall or recognition accuracy after a delay (Nation et al., 2007; Tamminen et al., 2017; Vlach & Sandhofer, 2014; Wang et al., 2011). For example, Vlach and Sandhofer (2014) instructed participants to learn novel objects and their spoken names (i.e., semantics, phonology, and their bindings). Participants were asked to choose objects for spoken names (measuring phonology-to-semantics bindings) immediately after learning and 1 week later. Results showed that participants were highly accurate immediately after learning but were significantly less accurate after a week. Regarding orthographic learning, Nation et al. (2007) exposed children to written pseudowords embedded in stories and then tested their memory using a four-alternative orthographic choice task. Results indicated that children were more likely to recognize correct forms over distractors when tested 1 day after learning, compared with 7 days after learning. Similar findings following a 10-day delay were reported by Wang et al. (2011). These studies therefore demonstrate that learners acquire lexical knowledge but forget the lexical knowledge over time. It is, however, unknown whether consistency would modulate forgetting of lexical knowledge. The present study aimed to fill in the gaps, by systematically investigating the effect of consistency on learning and forgetting of lexical knowledge (i.e., orthographic, phonological, and semantic representations, orthography–phonology, orthography–semantics, semantics–phonology bindings) via self-teaching.

The present study instructed participants to self-teach pseudowords. Self-teaching (Share, 1995) assumes that, before encountering a novel written word (e.g., "peak"), children have acquired it as a spoken word (or an oral vocabulary item). They know its pronunciation (/piːk/; phonological representation) and meaning (the top of a mountain; semantic representation). Later, when children encounter the novel written word ("peak"), they can decode it into sounds (/piːk/), which provides phonological cues. Sentences (e.g., "They climbed to the peak") can provide semantic cues (the top of a mountain) for the word. Using both phonological and semantic cues, children can identify the known spoken word. They can then learn its orthographic form and associate existing phonological and semantic representations to the new orthographic form (learning phonology–orthography and semantics–orthography, respectively).

Empirical studies provide evidence for the self-teaching hypothesis. In Share (1999) and Cunningham et al. (2002), children read stories including pseudowords (e.g., invented objects, city names, food). In an orthographic choice posttest (measuring orthographic representations), children had to select the correct written form from four choices. In a spelling test, they spelled the pseudowords when given their meanings (e.g., invented objects, city names, food; measuring semantic-to-orthography bindings). They also read aloud the written pseudowords (measuring orthography-to-phonology bindings). Results showed that the target pseudowords were recognized correctly more often, spelled more accurately, and read aloud more quickly than untrained control pseudowords. This indicates that children taught themselves orthographic representations, semantics–orthography, and orthography–phonology bindings through reading stories.

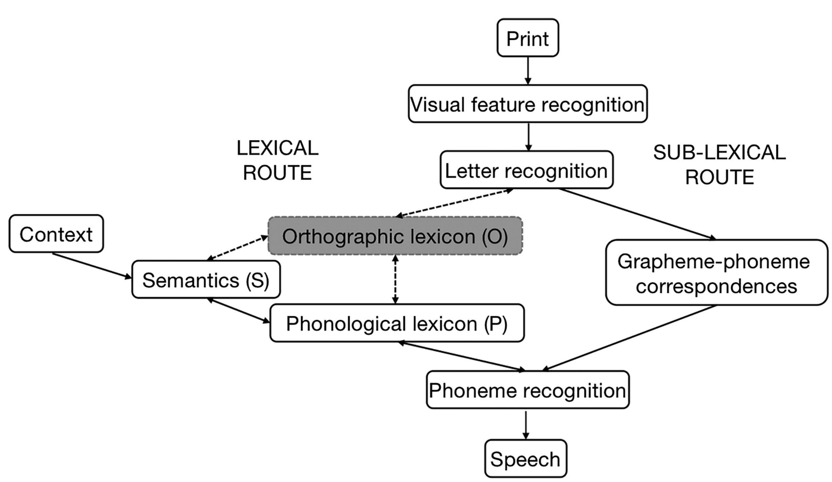
Adults can also self-teach new words. Brysbaert et al. (2016) reported that adults learn 6,000 new words between the ages of 20 and 60. Consider, for example, adults acquire new terms related to new technology or their new hobbies through podcasts or conversations. Later, they encounter written forms of the new terms during reading. Smejkalova and Chetail (2023) instructed adults to read books in which pseudowords were embedded. Posttests showed that adults recognized the correct written forms above chance level and spelled the target pseudowords more accurately than the untrained pseudowords. This provides evidence that the adults could self-teach novel written words. We now consider how consistency might be expected to affect learning and forgetting of lexical knowledge in the context of self-teaching.

Inconsistency Should Hinder Self-Teaching of Orthographic Representations, Phonology–Orthography and Semantics–Orthography Bindings

The self-teaching theoretical framework (Pritchard et al., 2018; Share, 1995; Ziegler et al., 2014) predicts that, compared with consistent words, inconsistent words will be learned with less precise orthographic representations, phonology–orthography and semantics–orthography bindings. This will be illustrated with an example of word learning in the self-teaching dual-route cascaded model (ST-DRC; Pritchard et al., 2018; Figure 2). Before encountering novel written words in print, the model has acquired spoken words and has oral vocabulary knowledge along the lexical route. That is, the model has semantic representations at the semantics node (e.g., an invention that is used for cleaning out fish tanks; it has a sponge and is shaped like an arm), phonological representations at the phonological lexicon node (/ziːk/), and bindings between the two nodes. However, orthographic representations (at the orthographic lexicon node), phonology–orthography bindings, and semantics–orthography bindings are to be

Figure 2

The Self-Teaching Dual-Route Cascaded Model



Note. Adapted from "A Computational Model of the Self-Teaching Hypothesis Based on the Dual-Route Cascaded Model of Reading," by S. C. Pritchard, M. Coltheart, E. Marinus, & A. Castles, 2018, *Cognitive Science, 42*(3), pp. 731, 735 (https://doi.org/10.1111/cogs.12571). Copyright 2018 by Cognitive Science Society, Inc. Adapted with permission. To simulate that the model has not yet self-taught the new written word, the orthographic lexicon is grayed out. Dotted lines between the orthographic lexicon, phonological lexicon, semantics, and letter recognition indicate that these are disconnected for the new written word. Once self-taught, the orthographic lexicon is ungrayed and lines become solid, that is, connected, such that the lexical route is fully established.

self-taught. Along the sublexical route, the model is programed with grapheme–phoneme correspondences, so it can use common letter–sound relations to decode letters (from the letter recognition node) into sounds (the phoneme recognition node).

When the model encounters a novel written word (e.g., "zeak") in print, it decodes letters ("zeak") into sounds /ziːk/ (decoded output at the phoneme recognition node) via the sublexical route. At the same time, context (e.g., a semantically rich and unambiguous sentence "when the fish tank was clean, he turned off the zeak and put it away") provides semantic cues that activate semantic representations at the semantics node (the invention that cleans out fish tanks; it has a sponge and is shaped like an arm). This in turn activates phonological representations at the phonological lexicon node (e.g., /ziːk/), generating lexical output. Since the lexical output matches the decoded output (/ziːk/), the *SpokenWordRecognisedThreshold* parameter at the phonological lexicon node is quickly reached and the novel written word "zeak" is recognized as a known spoken word (or oral vocabulary item). Self-teaching is then initiated. Based on the letters encountered in print, the whole-word orthographic representation ("zeak") is stored in the orthographic lexicon. Existing oral vocabulary knowledge, including lexical phonology (/ziːk/) and semantics (the arm-shaped invention that cleans out fish tanks with a sponge), is then mapped onto the new orthography. Phonology–orthography and semantics–orthography bindings are therefore formed and self-taught.

In the illustration above, "zeak" was a consistent word. However, if "zeak" is an inconsistent word with a pronunciation of /zeɪk/ stored in the phonological lexicon instead, self-teaching is more difficult. The model still decodes "zeak" into /ziːk/. However, the decoded /ziːk/ does not match any pronunciation in the phonological lexicon. It is therefore more difficult to recognize the novel written word "zeak" as the known oral vocabulary /zeɪk/ and to initiate self-teaching. Therefore, the ST-DRC model predicts that orthography, phonology–orthography, and semantics–orthography bindings should be learned less well for inconsistent than consistent words. Self-teaching studies showed that inconsistency impaired learning of orthography and phonology–orthography bindings, providing support for this prediction (e.g., Murray et al., 2022; Wang et al., 2012, 2013).

Inconsistency Should Undermine Phonological Representations

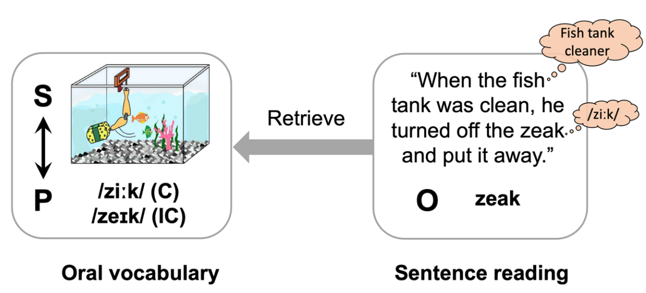
The spelling pronunciation account (Elbro & de Jong, 2017) predicts that inconsistent words would be learned with less precise phonological representations, compared with consistent words. This account proposed that readers store the decoded output (/ziːk/ decoded from "zeak") as a representation in the mental lexicon, in addition to the correct representation (/zeɪk/). Supporting evidence comes from Ranbom and Connine (2011). In a visual lexical decision task, participants made quicker decisions if they heard the correct pronunciation (/kɑːsəl/) prior to seeing the word "castle," compared with hearing an irrelevant control word. Importantly, this priming effect was also observed when participants heard the decoded pronunciation /kɑːstəl/ instead, suggesting that this decoded pronunciation was also stored as a phonological lexical representation for "castle." In Bürki et al. (2012), participants first learned spoken words (e.g., /zeɪk/) and then saw their written forms ("zeak"). In an auditory recognition posttest, participants indicated whether auditorily presented items were trained words. Participants tended to accept the decoded pronunciation (/ziːk/) as the trained word, despite that they learned /zeɪk/ as the correct spoken word. This indicates that participants stored decoded pronunciations as correct pronunciations instead, which undermined phonological representations. By contrast, decoded pronunciations of consistent words (e.g., /ziːk/ decoded from "zeak") match the correct pronunciations (/ziːk/) and should not undermine phonological representations. Therefore, we predict that inconsistent words would be learned with less precise phonological representations, compared with consistent words.

Inconsistency Should Enhance Retention and Mitigate Forgetting of Semantic Representations and Semantics–Phonology Bindings

How might consistency affect forgetting of lexical knowledge? Existing theories and models of reading development make no prediction, as they do not consider forgetting of lexical knowledge (triangle models, Harm & Seidenberg, 2004; Plaut et al., 1996; the self-teaching hypothesis, Share, 1995). However, the episodic context account of retrieval-based learning (Karpicke et al., 2014) predicts that inconsistency should elicit more reinstatement of oral vocabulary learning episodes, which enhances memory retention of semantic representations and semantics–phonology bindings. That is, inconsistency should protect them from forgetting.

According to this account, retrieval involves using cues that are available in the present to reconstruct previous events whose features match the cues. Retrieval is inherent in self-teaching of new written words. As illustrated in Figure 3, sentences such as "when the fish tank was clean, he turned off the zeak and put it away" provide two cues about the novel written word "zeak." One is the phonological cue /ziːk/, decoded from the written form. The other is a semantic cue, that is, "zeak" can be used to clean fish tanks. Readers can use the cues to retrieve a previous event in which they learned the spoken word (or oral vocabulary item) whose features match the cues. Successful retrieval allows readers to recognize the

Figure 3
Retrieval in Self-Teaching



Note. The picture of the fish tank cleaner was taken from "Context Effects on Orthographic Learning of Regular and Irregular Words," by H.-C. Wang, A. Castles, L. Nickels, and K. Nation, 2011, *Journal of Experimental Child Psychology, 109*(1), p. 56 (https://doi.org/10.1016/j.jecp.2010.11.005). Copyright 2011 by Elsevier Inc. Adapted with permission. S = semantic representation; P = phonological representation; O = orthographic representation; C = consistent condition; IC = inconsistent condition. See the online article for the color version of this figure.

novel written word "zeak" as a known spoken word and to map oral vocabulary knowledge onto the novel written form.

The episodic context account of retrieval-based learning also proposed that the effectiveness of cues determines the degree to which a previous event must be reinstated and that reinstatement enhances retention of memories that was retrieved. For a consistent word (e.g., "zeak" pronounced as /ziːk/), both phonological and semantic cues point to a known spoken word (the fish tank cleaner /ziːk/). Learners can therefore easily recognize the item without the need to reinstate previous events in which they learned the spoken word. The benefits of reinstatement are hence minimal.

However, for an inconsistent word (e.g., "zeak" pronounced as /zeɪk/), the phonological cue /ziːk/ cannot effectively elicit the target known spoken word because it does not match any known spoken word. To successfully retrieve the item, learners will have to use the semantic cue to reinstate previous events in which they encountered the fish tank cleaner, and actively recall information about the item, including its semantic features (it has a sponge and is shaped like an arm) and its spoken name (the fish tank cleaner is called /zeɪk/; semantics–phonology bindings). This should benefit memory retention and hence mitigate forgetting of semantic representations and semantics–phonology bindings, for items that are successfully retrieved and recognized.

The Current Experiments

This study investigated the effect of consistency on learning and forgetting of lexical knowledge (orthographic, phonological, and semantic representations, orthography–phonology, orthography–semantics, and semantics–phonology bindings) in adults. Experiment 1 focused on learning of lexical knowledge. It instructed participants to self-teach pseudowords and tested the effect of consistency on learning outcomes. Experiment 2 was a follow-up study. It measured forgetting of the pseudowords after a long delay ($M = 77$ days) and tested the effect of consistency on forgetting.

Experiment 1

Experiment 1 instructed participants to learn four consistent and four inconsistent pseudowords using a self-teaching paradigm. The pseudowords were introduced as Professor Parsnip's inventions. Participants first learned about the inventions and their spoken names, thereby acquiring oral vocabulary knowledge for the pseudowords. They then read sentences presenting the written forms of the pseudowords. Immediately after, they completed posttests measuring various aspects of lexical knowledge. Table 1 summarizes theoretical accounts, predictions, posttests, measures, and hypotheses, as preregistered at https://osf.io/zbgja.

Method

Participants

This experiment was approved by the Language and Cognition Department Ethics Chair at University College London. One hundred and nine adults were recruited through Prolific (https://www.prolific.com). However, 23 of them were excluded, as they failed our preregistered oral vocabulary learning check (their

Table 1

Theoretical Accounts, Predictions, Posttests, Measures, and Hypotheses in Experiment 1

Theoretical account	Prediction	Posttest	Measure	Hypothesis
Self-teaching theoretical framework (Pritchard et al., 2018; Share, 1995; Ziegler et al., 2014)	Inconsistency hinders self-teaching of O	Orthographic decision	A-prime	Lower A-prime scores for inconsistent than consistent items
	Inconsistency hinders self-teaching of P–O	Spelling (P2O)	Accuracy	Lower accuracy for inconsistent than consistent items
		Reading aloud (O2P)	Accuracy, RT	Lower accuracy and/or longer RTs for inconsistent than consistent items
	Inconsistency hinders self-teaching of S–O	Labeling	Accuracy	Lower accuracy for inconsistent than consistent items
Spelling pronunciation account (Elbro & de Jong, 2017)	Inconsistency undermines P	Phonological decision	A-prime	Lower A-prime scores for inconsistent than consistent items
Episodic context account of retrieval-based learning (Karpicke et al., 2014)[a]	Inconsistency enhances retention of S, for items that are successfully retrieved during reading	Semantic feature recall	Number of correct features	More features recalled for inconsistent than consistent items that are recognized in the orthographic decision test (an approximate of successful retrieval during self-teaching)
	Inconsistency enhances retention of S–P, for items that are successfully retrieved during reading	Definition naming	Accuracy, RT	Higher accuracy and/or shorter RTs for inconsistent than consistent items that are recognized in the orthographic decision test

Note. RT = response time; O = orthographic representations; P = phonological representations; S = semantic representations; O–P = bindings between orthographic and phonological representations; P2O = phonology-to-orthography bindings; O2P = orthography-to-phonology bindings; O–S = bindings between orthographic and semantic representations; S–P = bindings between semantic and phonological representations.

[a] Hypotheses regarding retention are tested because there is a short interval between oral vocabulary learning and posttests.

accuracy rate was below 50% in the final picture naming test in the oral vocabulary training phase; for details, see the Procedure section).

The final sample consisted of 86 adults (51 females, 35 males), aged between 18 and 40 years ($M = 29.78$, $SD = 6.03$). They were native standard southern British English speakers with normal or corrected-to-normal vision. They had no hearing loss or difficulties, language or reading related disorders, or attention deficits. A priori power analysis could not be conducted, as some effect sizes of interest were not available in previous research (e.g., the effect of consistency on learning of phonological, semantic representations, semantics–phonology, semantics–orthography bindings). Nevertheless, our sample size is larger than previous self-teaching studies that tested the effect of consistency (Murray et al., 2022; Wang et al., 2011, 2012, 2013).

Materials

Pseudowords. Eight pseudowords were used for training, as previous research (Hulme et al., 2022; Hulme & Rodd, 2021) suggested that eight new meanings are a reasonable number for participants to learn in a single session. Five of the pseudowords were taken from Trudeau (2006), one from Murray et al. (2022), and two were created. As listed in Table 2, they have four or five letters and a consonant(s)–vowel–consonant phonological structure. Each written pseudoword (e.g., "zeak") had two phonological forms, one for the consistent condition (/ziːk/) and one for the inconsistent condition (/zeɪk/). Consistency was manipulated at the vowel with the constraint of the following consonant and was manipulated in both letter-to-sound (or orthography-to-phonology; O2P; e.g., "ea," when followed by "k," is commonly pronounced as /iː/) and sound-to-letter (or phonology-to-orthography; P2O; /iː/, when followed by /k/, is commonly spelled as "ea") directions.

Consistency was quantified using coda-conditional surprisal values (Siegelman et al., 2020). Surprisal values are defined as $-\log_2 p(i)$, in which $p(i)$ is the probability of an event. For example, "ea," when followed by "k," is pronounced as /iː/ in 88% of the words in the Unisyn Lexicon Database (Fitt, 2000), so the surprisal value is $-\log_2 p(0.88) = 0.19$. It is pronounced as /eɪ/ in 12% of the words, so the surprisal value is $-\log_2 p(0.12) = 3$. Lower surprisal values indicate higher probabilities of letter–sound relations, that is, more common and consistent letter–sound relations. The surprisal values are conditioned on the coda, that is, the consonants following the vowels were considered. As listed in Table 2, items in the consistent condition have lower O2P and P2O surprisal values than in the inconsistent condition (consistent, $M_{O2P} = 0.29$, $M_{P2O} = 1.79$; inconsistent, $M_{O2P} = 2.80$; $M_{P2O} = 4.15$). This demonstrates that consistent items follow more common letter–sound relations than inconsistent items. All pronunciations were recorded by a female speaker of standard southern British English.

This surprisal value measure is akin to the standard measure of consistency, which calculates the proportion of words in which the body is pronounced in the same way. The main difference is that surprisal values are log-transformed and negated. Siegelman et al. (2020) showed that the surprisal measure and standard consistency measure are highly correlated ($r = -0.9$). Although the surprisal values are based on words in American English (Siegelman et al., 2020), correlations between these surprisal values and reaction times in the lexical decision task from the British Lexicon Project (Keuleers et al., 2012) are similar to those from the United States-based English Lexicon Project (Balota et al., 2007; ρ, O2P direction: −0.087, −0.082; P2O direction: −0.056, −0.054; for British Lexicon Project and English Lexicon Project, respectively). This suggests that this surprisal measure should be valid for British English speakers.

Pseudowords were presented as "Professor Parsnip's inventions" (e.g., a fish tank cleaner), which were taken from Wang et al. (2011) and Wegener, Wang, Beyersmann, Nation, et al. (2023). Each invention has a definition, consisting of a function (e.g., it is used for cleaning out fish tanks) and two features (it has a sponge and is shaped like an arm), as well as a picture depicting this invention (e.g., the fish tank cleaner in Figure 3).

Sentences That Present the Written Pseudowords. For each pseudoword, three sentences were adapted from Wegener et al. (2020). Each sentence presented one pseudoword (e.g., "Ben picked up the fish tank and the arm-shaped zeak to clean the dirty glass"). The sentences provided unambiguous semantic information for the pseudowords. This is particularly important for

Table 2
Pseudowords and Their Surprisal Values

Orthography	Consistency	Phonology	O2P surprisal value	P2O surprisal value
Zeak	Consistent	/ziːk/	0.19	1.36
	Inconsistent	/zeɪk/	3.00	3.64
Vose	Consistent	/vəʊz/	0.29	2.61
	Inconsistent	/vuːz/	2.46	5.07
Kaid	Consistent	/keɪd/	0.36	2.43
	Inconsistent	/ked/	3.17	4.75
Flome	Consistent	/fləʊm/	0.42	1.00
	Inconsistent	/flʌm/	2.00	3.95
Vord	Consistent	/vɔːd/	0.22	1.95
	Inconsistent	/vɜːd/	2.81	4.46
Doad	Consistent	/dəʊd/	0.26	2.63
	Inconsistent	/dɔːd/	2.58	4.09
Breat	Consistent	/briːt/	0.26	1.14
	Inconsistent	/bret/	3.17	3.81
Chone	Consistent	/tʃəʊn/	0.34	1.22
	Inconsistent	/tʃʌn/	3.25	3.46

Note. O2P = orthography-to-phonology; P2O = phonology-to-orthography.

inconsistent words because semantic information serves as the only effective cue for readers to retrieve and recognize the target known spoken words.

Comprehension Questions. To ensure that participants read the sentences carefully, comprehension questions were created. For each pseudoword, one out of three sentences was used to create a yes-or-no question. The questions were easy and did not refer directly to the pseudowords (e.g., "Was Ben too lazy to clean the fish tank?").

Foils for the Orthographic Decision and Phonological Decision Posttests. For each pseudoword, two foils were created for use in the orthographic decision test, one for the consistent condition and the other for the inconsistent condition. For the consistent condition (e.g., when "zeak" was pronounced as /ziːk/), a homophonic foil ("zeek") was created by encoding the pronunciation (/ziːk/) using alternative sound-to-letter mappings that are common in English words (/iː/, when followed by /k/, is commonly spelled as "ee," besides "ea"). For the inconsistent condition (e.g., when "zeak" was pronounced as /zeɪk/), a foil ("zake") was created by encoding the pronunciation (/zeɪk/) using common sound-to-letter mappings (/eɪ/, when followed by /k/, is commonly spelled as "ake"). This followed previous studies (Wang et al., 2011, 2012, 2013).

In the inconsistent condition, foils have much lower surprisal values than the target pseudowords in both letter-to-sound and sound-to-letter directions (foils, $M_{O2P} = 0.04$, $M_{P2O} = 0.79$; targets, $M_{O2P} = 2.80$, $M_{P2O} = 4.15$). In the consistent condition, foils have similar surprisal values to the target pseudowords (foils, $M_{O2P} = 0.01$, $M_{P2O} = 1.65$; consistent targets, $M_{O2P} = 0.29$, $M_{P2O} = 1.79$). Thus, in both consistency conditions, foils are plausible spellings of the spoken forms. Foils in the two consistency conditions are also closely matched in neighborhood (mean Levenshtein distance to 20 closest orthographic neighbors; Yarkoni et al., 2008; consistent, $M = 1.71$; inconsistent, $M = 1.61$) and distance to corresponding targets (measure by Levenshtein distance; Levenshtein, 1966; consistent, $M = 1.63$; inconsistent, $M = 1.63$).

Foils in the phonological decision test constituted the consistent or inconsistent pronunciations that were not learned by the participants. For example, if a participant learned "zeak" in the inconsistent condition (pronounced as /zeɪk/), the foil was /ziːk/ (pronunciation in the consistent condition). This was motivated by the spelling pronunciation account (Elbro & de Jong, 2017), which proposed that readers decode novel written words (e.g., "zeak") using common letter-to-sound mappings and store the decoded output (/ziːk/) as a phonological representation. For the consistent condition, ideally, foils would be pronunciations that are decoded using alternative common letter-to-sound mappings. However, for most items, the only alternatives were those used in the inconsistent condition. For example, "zeak" was pronounced as /ziːk/ in the consistent condition, and its only alternative pronunciation is /zeɪk/ as in the inconsistent condition. Therefore, foils for pseudowords in the consistent condition were their pronunciations in the inconsistent condition.

All the materials (including pseudowords, recordings, definitions, pictures, sentences, comprehension questions, and foils) can be found at https://osf.io/dqcbw/ (Qiu & Taylor, 2025a).

Design

This experiment had a within-subjects design. Participants learned eight pseudowords, four from the consistent condition and four from the inconsistent condition. Four versions of the experiment were created to counterbalance the assignment of phonology and semantics to orthography (e.g., in Versions 1 and 2, Items 1–4 were consistent, whereas Items 5–8 were inconsistent. In Versions 3 and 4, the reverse was true. In Versions 1 and 3, Items 1–4 were assigned Inventions A–D and Items 5–8 were assigned Inventions E–H. In Versions 2 and 4, the reverse was true). Participants were randomly assigned to one of the four versions using the Gorilla (https://gorilla.sc).

Procedure

We adapted the paradigm from Wang et al. (2011). There were three phases. First, in the oral vocabulary training phase, participants learned spoken pseudowords as Professor Parsnip's novel inventions. Second, in the self-teaching phase, participants silently read sentences about the inventions. Written forms of the pseudowords were presented for the first time. Last, in the posttest phase, participants completed tests that assessed knowledge of the pseudowords. We used this paradigm because it met the key assumption of the self-teaching hypothesis: oral vocabulary knowledge is available prior to self-teaching the written form (Share, 1995). In the current paradigm, the oral vocabulary training phase equipped participants with oral vocabulary knowledge and prepared them for self-teaching. Other self-teaching paradigms were not used (e.g., Chaves et al., 2020; Cunningham et al., 2002; Lien, 2017; Share, 1999; Smejkalova & Chetail, 2023), due to a lack of the oral vocabulary training phase. The current experiment was conducted online using Gorilla (https://gorilla.sc; Anwyl-Irvine et al., 2020) and was self-paced by participants. It took an average of 28 min to complete.

Oral Vocabulary Training Phase. During this phase, participants were instructed to learn Professor Parsnip's novel inventions. This included learning of semantic representations, phonological representations, and their bindings of the pseudowords. Items were trained in four blocks. In Block 1, four items were trained one at a time. Participants first saw a picture of an invention (e.g., fish tank cleaner) and heard its name (e.g., /ziːk/). They then repeated the invention name. Next, they saw its definition on the screen (e.g., "It is used to clean fish tanks. It has a sponge and is shaped like an arm"), with the invention name auditorily presented at the same time. Participants then repeated the invention name again. Finally, they were asked to say aloud what the invention is used for. In Block 2, another four items were trained using the same procedure as in Block 1.

In Block 3, all eight items were practiced one at a time. For each item, participants were presented with a picture and definition on the screen and were asked to say aloud the invention name. The correct invention name was then provided as feedback. Participants repeated it. They then made up a sentence and said it aloud, starting with "If I had a (invention name), I would use it for _____." No feedback was given. In Block 4, all items were practiced again following the same procedure as in Block 3.

At the end of each block, participants completed a picture naming task. Pictures were presented one at a time and participants named each of them. Correct pronunciations and definitions were presented as feedback. Participants' responses were recorded and then analyzed offline for accuracy. The picture naming task in Block 4 was used as the learning check. Following Wegener, Wang, Beyersmann, Nation, et al. (2023), participants whose accuracy was below 50% in the learning check were excluded.

Self-Teaching Phase. After learning oral vocabulary, participants were instructed to silently read sentences. They were told that the sentences would be about the inventions that they had learned, they should try to understand the sentences and that they would be asked questions following some of the sentences. Participants saw each written pseudoword three times across three blocks. Each block had eight sentences, each presenting one of the pseudowords. Sentences were randomized within blocks.

Posttest Phase. Immediately after the self-teaching phase, participants completed a series of posttests, measuring various aspects of lexical knowledge (summarized in Figure 4). No feedback was given in any of the tests.

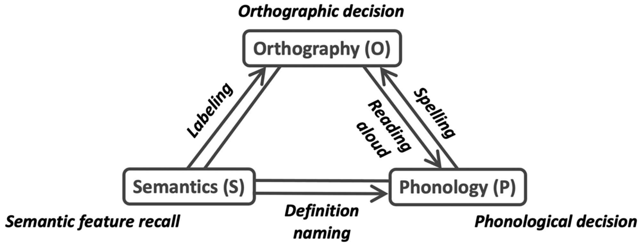
Orthographic Decision. This test was adapted from Wang et al. (2011) and assessed participants' orthographic representations (O). Pseudowords and their foils were presented on the screen, one at a time in a random order. Participants were instructed to indicate whether the presented item was an invention name or not as accurately as possible, by pressing the "A" key (it is an invention name) or "L" key (it is not). Their responses were recorded and then categorized into hits (correct recognition of the invention names), misses (failure to recognize the invention names), correct rejections (of the foils as invention names), or false alarms (incorrectly accepting the foils as invention names). Based on these scores, A-prime scores were generated to reflect precision of orthographic representations, considering not only correct recognition of the invention names but also correct rejections of the foils (Stanislaw & Todorov, 1999), using the psycho package (Makowski, 2018) in R (R Core Team, 2023). Although D-prime is more commonly used, its assumptions about distributions of probabilities underlying decisions on targets and foils "cannot be actually tested" (Stanislaw & Todorov, 1999, p. 140). Therefore, we used the nonparametric A-prime measure.

Phonological Decision. This test assessed participants' phonological representations (P). Pseudowords and their foils were auditorily presented one at a time in a random order. Participants indicated whether the presented item was an invention name or not as accurately as possible. Their responses were recorded and then processed in the same way as in the orthographic decision test.

Semantic Feature Recall. This test assessed participants' semantic representations (S). According to the lexical quality hypothesis (Perfetti, 2007), high relative to low quality semantic knowledge has "a fuller range of meaning dimensions to discriminate among words in same semantic field" (p. 360). In our study, each invention had three meaning dimensions, that is, a function (e.g., the invention is used for cleaning out fish tanks) and two features (it has a sponge and is shaped like an arm). To assess participants' semantic knowledge, we presented inventions' functions one at a time on the screen. Participants were instructed to use the provided function as a cue to recall as many features as possible for the invention and type them into blank text boxes. The number of features that were correctly recalled was the outcome measure.

Figure 4

Posttests Measuring Various Aspects of Lexical Knowledge



Note. Posttests are in italics.

Pilot data showed that participants recalled not only semantic features that were trained during the oral vocabulary training phase (e.g., the fish tank cleaner has a sponge and is shaped like an arm) but also other features that they inferred from pictures and sentences (e.g., colors, ways to use the inventions). Therefore, we used two measures. One measure gave points to trained features only. The other measure gave points to trained features as well as inferred features. The latter measure was scored by two researchers independently (90.39% interrater agreement). Any disagreements were later resolved on a case-by-case basis through discussion.

Spelling. This test assessed knowledge of phonology-to-orthography bindings (P2O). In each trial, participants heard an invention name and typed its spelling into a blank text box as accurately as possible. The spelling was scored as being accurate if it fully matched the correct orthographic form.

Labeling. This test assessed knowledge of semantics-to-orthography bindings (S2O). In each trial, participants saw a picture of an invention and typed its name into a blank text box as accurately as possible. The response was scored as being accurate if it fully matched the correct orthographic form.

Definition Naming. This test assessed knowledge of semantics-to-phonology bindings (S2P). In each trial, participants saw the definition of an invention on the screen and said its name aloud as accurately and quickly as possible. Spoken responses were recorded and phonetically transcribed following the phonemic vocabulary of the dual-route cascaded model (Rastle & Coltheart, 1999). The transcribed pronunciation was scored as being accurate if it fully matched the correct phonological form. Response times (RTs) for correct trials were derived by hand-marking acoustic onsets of responses using CheckVocal (Protopapas, 2007).

Reading Aloud. This test assessed knowledge of orthography-to-phonology bindings (O2P). In each trial, participants saw an invention name on the screen and read it aloud as accurately and quickly as possible. Spoken responses were recorded and processed for accuracy and RT measures as in the definition naming test.

Participants completed the posttests in a fixed order as above. The three tests that measured individual lexical components (O, P, S) were completed before those that measured bindings between these components (P2O, S2O, S2P, O2P). This ensured that tests of bindings did not provide participants with the answers to tests of individual lexical components. For example, the definition naming test could have provided participants with answers to the semantic feature recall test. Furthermore, we ensured that tests of the outcome of self-teaching came before other tests. Self-teaching involves learning of new orthographic representations (O) and associating existing oral vocabulary knowledge (P and S) to this orthographic information (hence, P2O and S2O). Thus, for tests of individual lexical components, the orthographic decision test (measuring O) was completed first and for tests of bindings, spelling (measuring P2O) and labeling (measuring S2O) tests were completed before other tests.

Due to time constraints, orthography-to-semantics (O2S) and phonology-to-semantics (P2S) bindings were not tested. A survey by Revilla and Höhne (2020) indicated that participants accepted 28 min as the maximum suitable duration for online studies. The present experiment took an average of 28 min to complete, thereby limiting the feasibility of including additional tests.

Analysis

Operationalization of Successful Retrieval During Self-Teaching. The episodic context account of retrieval-based learning (Karpicke et al., 2014) predicts a consistency effect only on items that are successfully retrieved during self-teaching. A direct measure of retrieval success would require participants to self-report, during reading, what they retrieved about previous oral vocabulary learning episodes. We did not implement this as it would interrupt the reading process and reduce the ecological validity of the self-teaching paradigm. Instead, successful recognition of the targets as inventions in the orthographic decision task was used as an indirect measure of retrieval success. In this task, which occurred immediately after the self-teaching phase, participants indicated whether the orthographic form was an invention name. Successful recognition thus indicates that they successfully retrieved associated oral vocabulary knowledge and mapped this onto the orthographic form.

Statistical Analysis. Mixed-effects models were used to test the effect of consistency (consistent and inconsistent pseudowords coded as 0.5 and −0.5, respectively) on posttest measures. With the lme4 package (Bates et al., 2015) in R (R Core Team, 2023), linear mixed-effects (LME) models fitted RT (from the definition naming and reading aloud tests) and A-prime data (from the orthographic decision and phonological decision tests). Logistic LME models fitted accuracy data from the spelling, labeling, definition naming, and reading aloud tests (correct and incorrect responses coded as 1 and 0, respectively), and additionally accuracy data from the final picture naming task in the oral vocabulary training phase. Cumulative link mixed models (Christensen, 2024) fitted data from the semantic feature recall test (the number of features that were correctly recalled).

To decide the best fitting random effects structure for each model, the buildmer package (Voeten, 2023) was used. First, a maximal model structure was specified. This included a fixed effect of consistency, by-participant random effects, and by-item random effects. The exceptions were models that fitted A-prime data, in which by-item random effects were not included. This is because A-prime scores are not calculated at the item level. For all models, the buildmer package started with an empty model and kept adding effects until convergence cannot be achieved. It then applied backward elimination on random effects that did not make significant contribution to the model.

When assumptions of homoscedasticity and/or normality of residuals were violated in RT or A-prime measures, we log-transformed and inverse-transformed (1,000/RT) raw data. To decide among the raw, log-, and inverse-transformed data, we compared scatterplots of the residuals versus fitted values, as well as histograms showing distributions of residuals. Data that most closely met the assumptions were used for analysis.

To test the fixed effect of consistency, likelihood ratio tests were conducted to determine whether the model including the fixed effect of consistency provided a significantly better fit to the data than the model without the effect. An α level of .05 was used. One exception should be noted. A Wilcoxon signed-rank test, rather than LME models, was used to analyze A-prime data in the phonological decision test. This is because buildmer showed that the best fitting model did not have any random effects. We therefore could not proceed with LME models as planned.

As suggested by reviewers, we confirmed the robustness of our results by conducting additional analyses using mixed-effects models in which the random effects structure was kept maximal, while ensuring convergence (Barr et al., 2013). These maximal models showed the same effects of consistency as the best fitting models across all posttests that used mixed-effects models. In the Results section, we report results of the best fitting models as planned, unless indicated in the text. The best fitting models, maximal models, and their likelihood ratio test results are summarized in Supplemental Table S1.

Transparency and Openness

This experiment was preregistered at https://osf.io/zbgja. Materials, data, and analysis codes can be found at https://osf.io/dqcbw/ (Qiu & Taylor, 2025a). The online experiment is available at https://app.gorilla.sc/openmaterials/863359 (logging into a Gorilla account is required to access the webpage).

Results

Oral Vocabulary Learning Outcome

The final picture naming task in the oral vocabulary training phase was used to measure oral vocabulary learning outcome. After excluding participants whose accuracy rate was below 50% (N = 23), the overall mean accuracy rate was 82.41% (N = 86, SD = 16.54%). Logistic LME models indicated no effect of consistency. Participants learned consistent (M = 81.10%, SE = 2.01%) and inconsistent (M = 83.72%, SE = 2.01%) items equally well, $\chi^2(1)$ = 0.87, p = .35.

Reading Comprehension During Self-Teaching

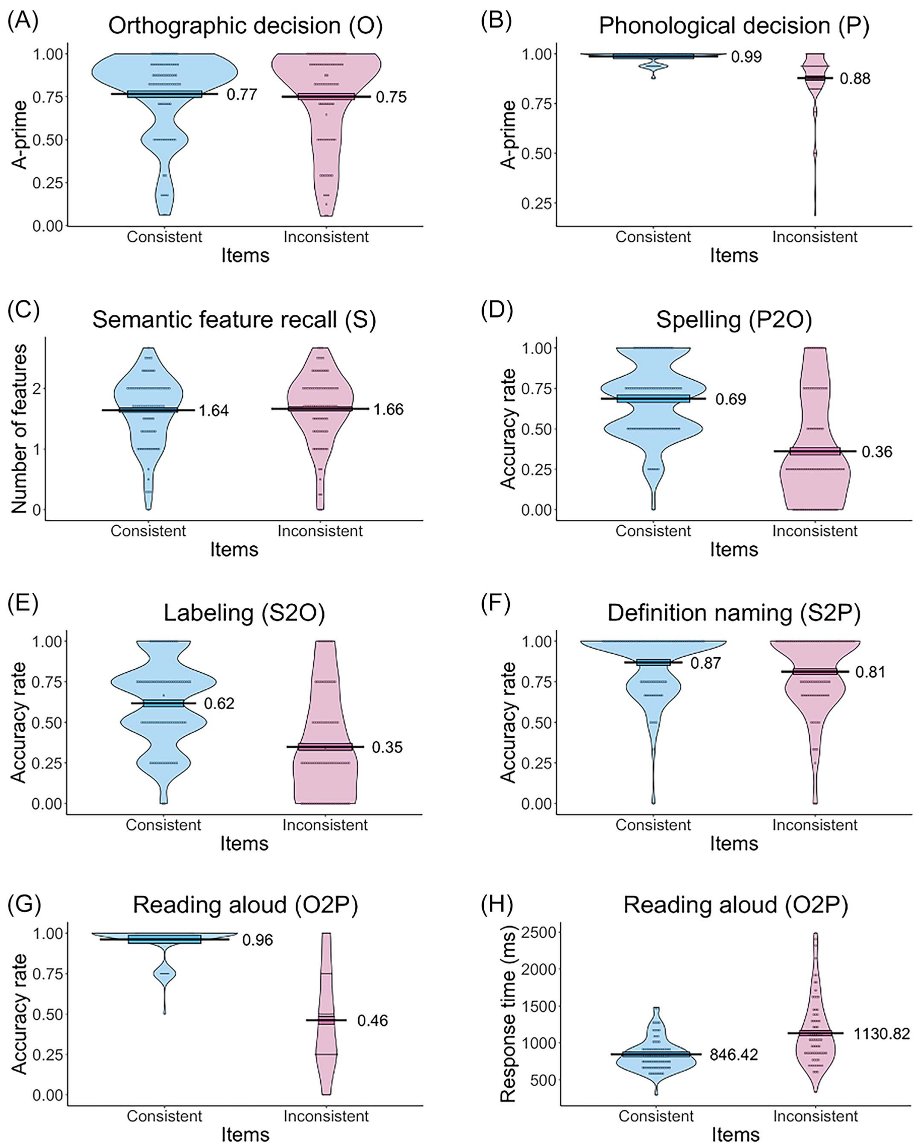
To ensure participants read sentences carefully during the self-teaching phase, we asked them to complete reading comprehension questions. All participants performed above the 50% chance level (M = 95.06%, SD = 7.25%, minimum = 62.5%, maximum = 100).

Posttests

Orthographic Decision (Measuring O). A-prime scores are displayed in Figure 5A. A score of 1 indicates perfect performance in recognizing target pseudowords and rejecting foils; a score of .5 indicates confusion between targets and foils; 0 is the minimum possible score. In both consistency conditions, mean scores were above .5. This suggests that participants learned orthographic forms for consistent and inconsistent items after self-teaching. A-prime scores were log-transformed and fitted with LME models. No effect of consistency was found, $\chi^2(1) = 0.67$, p = .41. Participants self-taught orthographic representations equally well for consistent and inconsistent pseudowords. Analyses of hits, misses, correct rejections, and false alarms (Table 3) showed that, across consistency conditions, participants correctly recognized target pseudowords

Figure 5

Posttest Performance Immediately After Self-Teaching



Note. Performance in the (A) orthographic decision test, measuring orthographic representations; (B) phonological decision test, measuring phonological representations; (C) semantic feature recall test, measuring semantic representations; (D) spelling test, measuring phonology-to-orthography bindings; (E) labeling test, measuring semantics-to-orthography bindings; (F) definition naming test, measuring semantics-to-phonology bindings; (G) and (H) reading aloud test, measuring orthography-to-phonology bindings. Dots represent participants, lines represent means, and boxes represent standard errors corrected for the within-subjects design. See the online article for the color version of this figure.

approximately 80% of the time and rejected foils approximately 60% of the time.

Phonological Decision (Measuring P). As shown in Figure 5B, participants achieved high A-prime scores in both consistency conditions. The scores were fitted with LME models. However, buildmer suggested that the best fitting model did not have any random effects. The maximal model converged but its nested model without the consistency effect showed convergence issues. We therefore could not proceed with LME models. A Shapiro–Wilk test showed that difference scores between consistency conditions were not normally distributed ($W = .71$, $p < .001$), so a Wilcoxon signed-rank test was conducted. Results showed a significant effect of consistency ($V = 2{,}233$, $p < .001$). Participants developed less precise phonological representations for inconsistent than consistent items. Analyses of hits, misses, correct rejections, and false alarms (Table 3) revealed that participants were highly accurate in recognizing target pseudowords in both consistency conditions (above 90% hit rate). They were also highly accurate in rejecting foils in the

Table 3
Hit, Miss, Correct Rejection, and False Alarm Rates in the Orthographic and Phonological Decision Tests

		Target (%)		Foil (%)	
Test	Consistency	Hit	Miss	Correct rejection	False alarm
Orthographic decision	Consistent	83.38	16.62	59.77	40.23
	Inconsistent	77.19	22.81	67.73	32.27
Phonological decision	Consistent	97.07	2.93	97.66	2.34
	Inconsistent	93.59	6.41	68.60	31.40

consistent condition. However, they often failed to reject foils in the inconsistent condition.

Semantic Feature Recall (Measuring S). Based on Karpicke et al. (2014), we first analyzed items that were correctly recognized in the orthographic decision task (83.38% of consistent items and 77.19% of inconsistent items). There were two measures of semantic feature recall for these items. One measure gave points to trained features only. The other measure gave points to trained features as well as inferred features. Cumulative link mixed models found no effect of consistency for either measure, trained features only, $\chi^2(1) = 0.90$, $p = .34$; trained and inferred features, $\chi^2(1) = 0.01$, $p = .91$. Participants recalled a similar number of semantic features for consistent and inconsistent items. Figure 5C shows the total number of trained and inferred features that participants correctly recalled. When all items were analyzed, the same results were found, that is, there was no effect of consistency on the number of trained features, $\chi^2(1) = 1.79$, $p = .18$; consistent, $M = 1.32$, $SE = 0.03$; inconsistent, $M = 1.26$, $SE = 0.03$; or the total number of trained and inferred features that were correctly recalled, $\chi^2(1) = 0.02$, $p = .90$; consistent, $M = 1.65$, $SE = 0.03$; inconsistent, $M = 1.64$, $SE = 0.03$.

Spelling (Measuring P2O). Figure 5D shows that participants were 69% and 36% accurate in spelling consistent and inconsistent items, respectively. This suggests that participants learned phonology-to-orthography bindings in both consistency conditions, although less well in the inconsistent condition. Logistic LME models fitted accuracy data and showed a significant effect of consistency, $\chi^2(1) = 12.03$, $p < .001$. Response patterns were explored by categorizing them into three types (Table 4). Phonologically derived errors refer to spellings that can be derived from phonology using alternative sound-to-letter relations as observed in English words in the CELEX Lexical database (Baayen et al., 1996). Other errors refer to responses that are neither correct responses nor phonologically derived errors. Phonologically derived errors accounted for more than half the responses for inconsistent items and approximately a quarter of responses for consistent items. There were a small number of other errors, for example, responses that were derived from phonology using implausible sound-to-letter relations (/dɔːd/ spelled as "daod").

Labeling (Measuring S2O). Accuracy data show that participants learned semantics-to-orthography bindings in both consistency conditions, although the learning was less successful for inconsistent (35%) than for consistent items (62%; Figure 5E). Accuracy data were fitted with logistic LME models. A significant effect of consistency was found, $\chi^2(1) = 8.69$, $p = .003$. Similar to the spelling test, phonologically derived errors constituted approximately half the responses for inconsistent items and a quarter for consistent items (Table 4). Other errors were also observed, for example, responses that were derived from phonology using implausible sound-to-letter relations, other trained pseudowords (correctly or incorrectly spelled), and "don't know."

Definition Naming (Measuring S2P). Based on Karpicke et al. (2014), items that were correctly recognized in the orthographic decision task were analyzed (83.38% of consistent items and 77.19% of inconsistent items). Accuracy data were fitted with logistic LME models. A significant effect of consistency was found, $\chi^2(1) = 6.98$, $p = .008$, but in the opposite direction to our hypothesis (Figure 5F). Participants were more accurate in naming consistent than inconsistent items. RT data were log-transformed and fitted with LME models. There was no significant difference between consistent ($M = 2015.41$ ms; $SE = 67.85$ ms) and inconsistent ($M = 2167.96$ ms; $SE = 68.67$ ms) conditions, $\chi^2(1) = 3.32$, $p = .07$. When all items were analyzed, inconsistent items had significantly longer RTs ($M = 2183.30$ ms; $SE = 56.70$ ms) than consistent items ($M = 2037.41$ ms; $SE = 56.70$ ms), $\chi^2(1) = 5.65$, $p = .02$. There was no difference in accuracy based on the maximal model, $\chi^2(1) = 1.72$, $p = .19$; consistent, $M = 86.63\%$, $SE = 0.016$; inconsistent, $M = 80.81\%$, $SE = 0.016$. The best fitting model converged but its nested model without the consistency effect showed convergence issues.

Again, response patterns were explored. Responses were categorized into three types as in Table 5. Orthographically derived errors refer to phonological forms that can be derived from orthography using alternative letter-to-sound relations as observed

Table 4
Response Patterns in the Spelling and Labeling Test

Test	Consistency	Correct response (%)	Phonologically derived error (%)	Other error (%)
Spelling	Consistent	68.60	27.33	4.07
	Inconsistent	36.05	52.91	11.05
Labeling	Consistent	61.81	25.95	12.24
	Inconsistent	34.99	45.78	19.24

Table 5
Response Patterns in the Definition Naming and Reading Aloud Test

Test	Consistency	Correct response (%)	Orthographically derived error (%)	Other error (%)
Definition naming (recognized items)	Consistent	87.06	1.05	11.89
	Inconsistent	79.55	9.47	10.98
Definition naming (all items)	Consistent	86.63	1.45	11.92
	Inconsistent	80.81	8.14	11.05
Reading aloud	Consistent	96.13	0.89	2.98
	Inconsistent	46.13	46.73	7.14

Note. Recognized items refer to items that are correctly recognized in the orthographic decision test.

in English words (CELEX Lexical database; Baayen et al., 1996). Other errors refer to responses that are neither correct responses nor orthographically derived errors. Across measures (recognized or all items) and consistency conditions, only a small number of responses (less than 10%) were orthographically derived errors. However, there were more orthographically derived errors when naming inconsistent than consistent items. Other errors accounted for 10%–12% of the responses. They were responses that were derived from orthography using implausible letter-to-sound relations (e.g., /deɪd/ for the invention whose name was "doad"), other trained pseudowords (correctly or incorrectly pronounced), and "don't know."

Reading Aloud (Measuring O2P). Accuracy data showed that 96% of consistent items and 46% of inconsistent items were read correctly. This indicates learning of orthography-to-phonology bindings in both consistency conditions, although inconsistent items were learned less well (Figure 5G). A significant effect of consistency was found, $\chi^2(1) = 9.49$, $p = .002$. Most errors in reading inconsistent items were orthographically derived errors. There were a small number of other errors, for example, responses that were derived from orthography using implausible letter-to-sound relations. RT data were log-transformed and fitted with LME models. A significant effect of consistency was also found, $\chi^2(1) = 44.25$, $p < .001$. It took longer to read aloud inconsistent than consistent items (Figure 5H).

Discussion

In Experiment 1, participants learned eight pseudowords, four consistent and four inconsistent. Immediately after learning, participants completed posttests assessing various aspects of lexical knowledge of the pseudowords. We discuss findings with respect to our predictions and hypotheses.

Inconsistency Hindered Self-Teaching of Phonology–Orthography and Semantics–Orthography Bindings, but Not That of Orthographic Representations

The self-teaching theoretical framework (Pritchard et al., 2018; Share, 1995; Ziegler et al., 2014) predicts that inconsistency should hinder self-teaching of orthographic representations, phonology–orthography, and semantics–orthography bindings. Accordingly, we hypothesized that, in posttests measuring these aspects of lexical knowledge, participants should show lower accuracy and/or longer RT for inconsistent than consistent items. Results were mixed.

In the spelling and reading aloud tests (measuring phonology-to-orthography and orthography-to-phonology bindings, respectively), participants were significantly less accurate for inconsistent than consistent pseudowords. This supports our hypothesis and the prediction that inconsistency impairs self-teaching of phonology–orthography bindings. The results are also in line with previous self-teaching studies (e.g., Murray et al., 2022; Wang et al., 2012, 2013). Response pattern analysis showed that participants made more phonologically derived errors when spelling inconsistent items (52.91%) compared with consistent items (27.33%) and made more orthographically derived errors when reading aloud inconsistent items (46.73%) compared with consistent items (7.14%). That is, compared with consistent items, participants were more likely to spell and read inconsistent words using alternative letter–sound relations. In addition, we found that participants' responses were significantly slower when reading aloud inconsistent relative to consistent pseudowords.

In the labeling test (measuring semantics-to-orthography bindings), accuracy rate was significantly lower for inconsistent than consistent items, supporting the prediction that inconsistency should hinder self-teaching of semantics–orthography bindings. Response pattern analysis indicated that most responses were phonologically derived errors when participants labeled inconsistent items (45.78%). Phonologically derived errors were also observed when participants labeled consistent items, although less often (25.95%). This gives clear evidence that phonology was involved when participants tried to generate orthographic forms from semantics. It is possible that participants produced the phonological form using robust semantics–phonology bindings developed during oral vocabulary learning and then used knowledge of phonology–orthography bindings to generate orthography. That is, access from semantics to orthography is mediated by phonology (semantics-to-phonology-to-orthography pathway). This is similar to the phonologically mediated pathway proposed in the triangle model (Harm & Seidenberg, 2004), except that the direction is different (orthography-to-phonology-to-semantics pathway). The triangle model relies on the phonologically mediated pathway early in learning, with the direct orthography-to-semantics pathway developing later.

The results from the labeling test also have implications for the self-teaching theoretical framework (Pritchard et al., 2018; Share, 1995; Ziegler et al., 2014). According to the framework, readers self-teach the novel orthographic form, phonology–orthography, and semantics–orthography bindings, after successfully recognizing the novel written word as a known spoken word. Our results suggest that it was not easy for participants to learn the direct semantics–orthography bindings and that they used the semantics-to-phonology-to-orthography pathway to help generate orthography. Our study provides some of the first evidence that consistency affects self-teaching of semantics–orthography bindings.

In contrast to our prediction, inconsistency did not affect self-teaching of orthographic representations, as A-prime scores were similar in the two consistency conditions in the orthographic decision test. This null effect is in fact in line with previous studies (McKay et al., 2008; Taylor et al., 2011; Wang et al., 2013). Analyses of hits, misses, correct rejections, and false alarms showed that, for consistent and inconsistent items, participants were equally likely to recognize the target pseudowords (approximately 80% of the time) and to reject the foils (approximately 60% of the time). It is interesting that participants tended to accept the foils as invention names (approximately 40% of the time) across consistency conditions. The orthographic skeleton hypothesis (Wegener et al., 2018) might provide an account. According to this hypothesis, people develop orthographic representations for spoken words before seeing them in print, by encoding sounds of spoken words into letters using common sound-to-letter mappings. In our experiment, participants might have developed their own orthographic representations for spoken inventions during the oral vocabulary training phase, before seeing the actual orthographic forms in the self-teaching phase. The self-generated representations might match foils in the orthographic decision test. Thus, participants might accept the foils as correct invention names. In the inconsistent condition, this is highly plausible because foils (e.g., "zake") are more likely spellings of the spoken forms (/zeɪk/) than the targets ("zeak"). This is also plausible in the consistent condition, since foils (e.g., "zeek") are similarly likely spellings of the spoken forms ("ziːk") as the targets ("zeak").

Inconsistency Undermined Phonological Representations

In the phonological decision test, participants showed significantly lower A-prime scores for inconsistent than consistent items. Analyses of hits, misses, correct rejections, and false alarms revealed that participants were highly accurate (above 90%) in recognizing the target pseudowords in both consistency conditions. However, they often accepted foils in the inconsistent condition (31.40%), but rarely did so in the consistent condition (2.34%). These results support the prediction of the spelling pronunciation account (Elbro & de Jong, 2017). That is, when participants encounter novel written words (e.g., "zeak") during self-teaching, they would decode the word into sounds (/ziːk/) using common letter-to-sound relations and then store the decoded output into the mental lexicon. This means that, in the inconsistent condition ("zeak" pronounced as /zeɪk/), the decoded output (/ziːk/) would be stored as a new phonological representation in addition to the existing correct pronunciation (/zeɪk/). When participants hear a foil in the form of the decoded output (/ziːk/), they are likely to accept it as an invention name. In the consistent condition (e.g., "zeak" pronounced as /ziːk/), participants would still decode "zeak" into /ziːk/, which matches the correct pronunciation /ziːk/. Therefore, participants would recognize /ziːk/ as the phonological representation. The foil (i.e., /zeɪk/) would not interfere with this process and therefore would not be recognized as an invention name. Our findings add to the existing literature (Bürki et al., 2012; Ranbom & Connine, 2011) that provides empirical support for the spelling pronunciation account.

Inconsistency Did Not Enhance Retention of Semantic Representations and Semantics–Phonology Bindings

The episodic context account of retrieval-based learning (Karpicke et al., 2014) predicts that inconsistency should enhance memory retention of semantic representations and semantics–phonology bindings for pseudowords that are successfully retrieved during self-teaching. We therefore hypothesized that, for items that are correctly recognized in the orthographic decision test (operationalization of successful retrieval), inconsistent relative to consistent items should have more features recalled in the semantic feature recall test (measuring semantic representations), and higher accuracy rates and/or shorter RTs in the definition naming test (measuring semantics–phonology bindings). However, the results did not support our hypotheses. In the semantic feature recall test, participants recalled a similar number of features for consistent and inconsistent items. In the definition naming test, RTs were similar across conditions. Opposite to the hypothesis, participants were less accurate in naming inconsistent than consistent items when prompted with meanings.

The lack of a beneficial effect of consistency on memory retention might be due to the short interval between learning and testing. Participants learned semantics and semantics–phonology bindings in the oral vocabulary training phase. They then read three blocks of eight sentences in the self-teaching phase, before completing posttests that assessed various aspects of lexical knowledge including semantic representations and semantics–phonology bindings. The interval between learning and testing varied across participants, but was short, with a maximum of 5 min and 10 s. During such a short interval, memory decay and forgetting might be minimal. Indeed, post hoc analyses showed that the mean accuracy rate in the semantics-to-phonology task was 82.41% at the end of oral vocabulary learning (the picture naming task) and remained at a similar level after the interval (83.72%; the definition naming posttest). Due to a lack of forgetting, the effect of consistency on forgetting could not be validly tested. To allow forgetting and hence testing of the consistency effect on forgetting, we conducted a follow-up study with a long delay (see Experiment 2).

The negative effect of inconsistency in the definition naming test was unexpected. We analyzed participants' response patterns to explore reasons for the reduced accuracy in inconsistent words. Noticeably, more orthographically derived errors were observed for inconsistent (recognized items, 9.47%; all items, 8.14%) than consistent items (recognized items, 1.05%; all items, 1.45%). Inconsistent and consistent items did not differ much in other errors (10%–12%). This provides additional evidence for the spelling pronunciation account (Elbro & de Jong, 2017). Participants stored the decoded pronunciations as additional phonological representations. Crucially, they might use these decoded representations to name the inventions.

The results coincide with those of Hulme et al. (2022) who also showed that letter–sound relations affect learning of semantics-to-phonology bindings. In Hulme et al., participants learned pseudowords by reading paragraphs that described their meanings. Half the pseudowords were easy to decode from letters to sounds, whereas the other half were hard to decode. After reading the paragraphs, participants saw definitions and were asked to say aloud the corresponding pseudowords. Participants were significantly less accurate with hard-to-decode items compared with easy-to-decode items, suggesting that letter-to-sound decoding ease affected learning of semantics-to-phonology bindings. This finding as well as ours (inconsistency impacted learning of semantics-to-phonology and semantics-to-orthography bindings) suggests that letter–sound relations have knock-on effects on learning to use forms to represent meaning. More broadly, these findings revealed that there are interactions between orthography, phonology, and semantics during word learning, which is in line with models of reading development such as the ST-DRC (Pritchard et al., 2018) and triangle model (Harm & Seidenberg, 2004).

Experiment 2

Experiment 2 was a follow-up study. Participants from Experiment 1 returned to complete the posttests again, with a mean delay of 77 days. This allowed us to test forgetting of lexical knowledge after a delay and the effect of consistency on forgetting. It was preregistered at https://osf.io/gknd2.

We hypothesized that participants would show forgetting in every aspect of lexical knowledge, such that there would be a significant decline in participants' posttest performance in Experiment 2 (delayed testing condition) compared with Experiment 1 (immediate testing condition).

The episodic context account of retrieval-based learning (Karpicke et al., 2014) predicts that inconsistency should enhance memory retention and thus mitigate forgetting of semantic representations and semantics–phonology bindings, for items that are successfully retrieved during self-teaching. Accordingly, we tested the hypothesis that, for items that were recognized in the immediate orthographic decision test (operationalization of the successful retrieval), inconsistent relative to consistent items should show less forgetting (i.e., smaller decline in performance) in semantic representations and semantics–phonology bindings, as measured by the semantic feature recall test and the definition naming test, respectively. This would be characterized by a significant interaction between testing (immediate vs. delayed testing) and consistency (consistent vs. inconsistent items).

In addition, we explored how consistency would modulate forgetting of other aspects of lexical knowledge (orthographic and phonological representations, phonology–orthography, and semantics–orthography bindings). We also explored the main effect of consistency on overall development of lexical knowledge, by considering lexical knowledge at two time points (i.e., immediately after the learning and after the delay).

Method

Participants

This experiment was approved by the Language and Cognition Department Ethics Chair at University College London. Participants that formed the final sample in Experiment 1 ($N = 86$) were invited to take part in this follow-up study. Fifty-eight participants returned (39 females, 19 males), after a mean delay of 77 days ($SD = 10$, minimum = 53, maximum = 92). They were aged between 19 and 40 years ($M = 30.47$, $SD = 6.02$).

Materials

The same pseudowords and foils were used as in Experiment 1.

Procedure

Participants were informed that this experiment was to test how well they remembered Professor Parsnip's inventions that they learned before. They completed the same posttests in the same order as in Experiment 1. The only exception was that participants were no longer instructed to respond as quickly as possible (in the definition naming and reading aloud tests). This is because we decided not to analyze RT data. We suspected that participants would find the posttests more difficult following the delay and therefore that the number of correct trials would be too small to conduct well-powered RT analyses. No oral vocabulary training or self-teaching was involved in this follow-up study.

Design

The within-subjects design had two independent variables, that is, testing (immediate vs. delayed testing) and consistency (consistent vs. inconsistent pseudowords).

Analysis

Operationalization of Successful Retrieval During Self-Teaching. As reasoned in Experiment 1, retrieval success during self-teaching was operationalized as correct recognition of target pseudowords as invention names in the orthographic decision test in Experiment 1.

Statistical Analysis. Data from participants who completed posttests at both time points (immediate testing in Experiment 1, delayed testing in Experiment 2) were analyzed. Mixed-effects models tested the effect of testing (immediate vs. delayed testing, coded as 0.5 and −0.5, respectively), consistency (consistent vs. inconsistent items, coded as 0.5 and −0.5, respectively), and their interaction on posttest measures. RT data were not analyzed. The maximal structure for the buildmer included fixed effects of testing, consistency, and their interaction. It also included by-participant and by-item random effects of testing, consistency, and the interaction. Likelihood ratio tests were conducted to test the main effects of testing and consistency and their interaction. Significant interactions were followed up by pairwise comparisons using the emmeans package (Lenth et al., 2024). Mixed-effects models fitted data from all posttests, except that a two-way within-subjects analysis of variance test was used to analyze A-prime data from the phonological decision test. This is because buildmer showed that the best fitting model did not have any random effects. We therefore could not proceed with mixed-effects models. Due to violations of assumptions, these A-prime data were additionally rank-transformed (Holbert, 2022). All other aspects of analysis were the same as in Experiment 1.

Additional analyses were also conducted in which the random effects structure of the mixed-effects models was kept maximal (Barr et al., 2013). These maximal models showed the same effects of consistency as in the best fitting models. The best fitting models, maximal models, and their likelihood ratio test results are summarized in Supplemental Table S2. The Results section reports results from the best fitting models.

Transparency and Openness

This experiment was preregistered at https://osf.io/gknd2. Data and analysis codes are available at https://osf.io/6ty32/ (Qiu & Taylor, 2025b). The online experiment can be found at https://app.gorilla.sc/openmaterials/866018 (logging into a Gorilla account is required to access the webpage).

Results

Orthographic Decision (Measuring O)

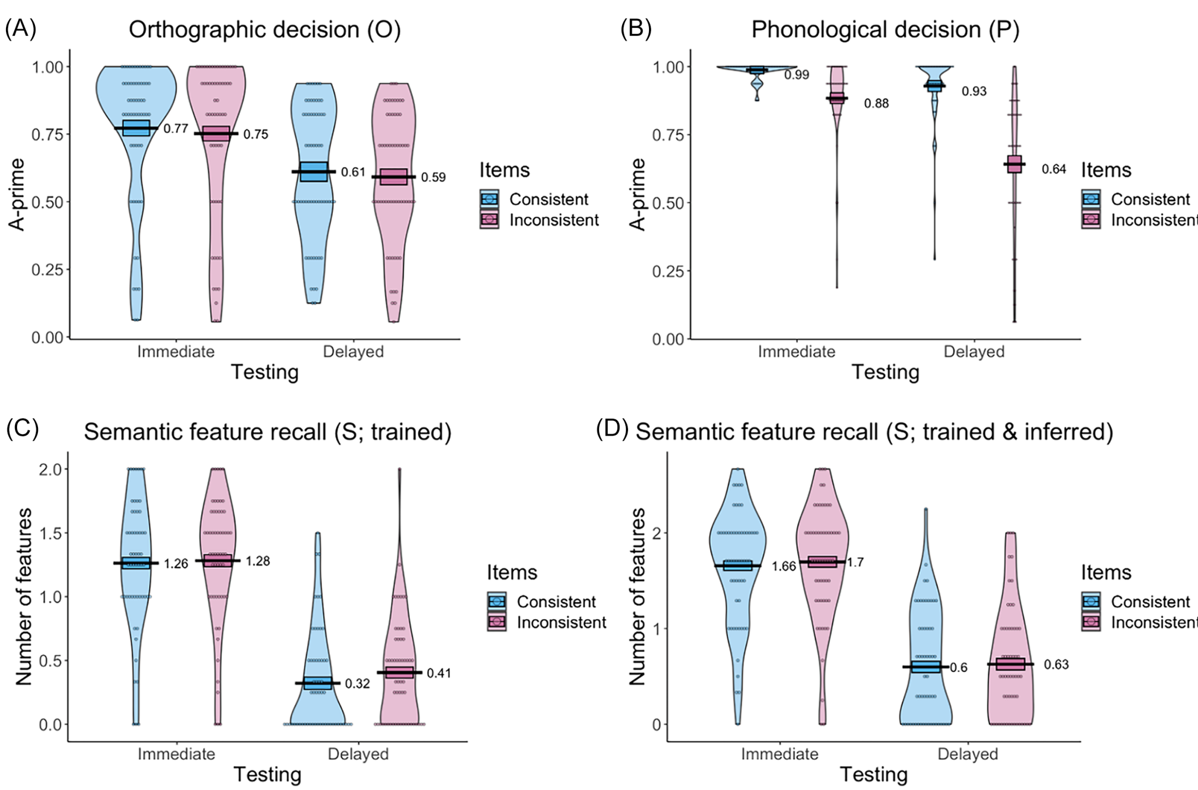
A-prime scores were fitted with LME models. A main effect of testing was found, $\chi^2(1) = 18.26$, $p < .001$. However, there was no significant effect of consistency, $\chi^2(1) = 0.52$, $p = .47$, and no interaction, $\chi^2(1) = 0.0004$, $p = .98$. As shown in Figure 6A, A-prime scores dropped on the delayed compared with the immediate test. This indicates that participants forgot precise orthographic representations after the delay. However, this process did not differ between consistent and inconsistent pseudowords. Analyses of hits, misses, correct rejections, and false alarms (Table 6) revealed that, after the delay, participants were less accurate in recognizing target pseudowords and in rejecting foils. The decline in accuracy was similar across the consistency conditions. The lack of a main effect of consistency also suggested that consistency did not affect overall development of orthographic representations.

Phonological Decision (Measuring P)

Rank-transformed A-prime scores were analyzed using a two-way within-subjects analysis of variance. Results showed significant effects of testing ($F = 63.83$, $p < .001$) and consistency ($F = 156.09$, $p < .001$), as well as an interaction ($F = 4.58$, $p = .03$). Follow-up pairwise comparisons found a simple effect of testing in both consistent ($t = -4.14$, $p < .001$) and inconsistent conditions ($t = -7.16$, $p < .001$). Figure 6B shows that participants forgot precise phonological representations after the delay, but, to a greater extent for inconsistent than consistent items. Table 6 revealed that, following the delay, both hit and correct rejection rates declined by approximately 10% for consistent items, but by about 20% for inconsistent items. The main effect of consistency suggests that, considering test performance at both time points, participants developed less precise phonological representations for inconsistent relative to consistent items.

Figure 6

Performance in Three Posttests Immediately After Self-Teaching and With a Delay



Note. Performance in the (A) orthographic decision test, measuring orthographic representations; (B) phonological decision test, measuring phonological representations; (C) and (D) semantic feature recall test, measuring semantic representations. Dots represent participants. Lines represent mean scores. Boxes represent standard errors corrected for the within-participants design. See the online article for the color version of this figure.

Table 6
Hit, Miss, Correct Rejection, and False Alarm Rates in the Orthographic and Phonological Decision Tests in Immediate and Delayed Testing Conditions

			Target (%)		Foil (%)	
Test	Consistency	Testing	Hit	Miss	Correct rejection	False alarm
Orthographic decision	Consistent	Immediate	83.55	16.45	61.90	38.10
	Consistent	Delayed	65.52	34.48	50.43	49.57
	Inconsistent	Immediate	77.92	22.08	67.67	32.33
	Inconsistent	Delayed	59.91	40.09	51.72	48.28
Phonological decision	Consistent	Immediate	97.40	2.60	97.84	2.16
	Consistent	Delayed	90.95	9.05	87.88	12.12
	Inconsistent	Immediate	94.40	5.60	70.69	29.31
	Inconsistent	Delayed	72.73	27.27	47.41	52.59

Semantic Feature Recall (Measuring S)

Based on Karpicke et al. (2014) and our preregistered hypothesis, we analyzed items that were correctly recognized in the immediate orthographic decision test (83.55% of consistent items and 77.92% of inconsistent items). As in Experiment 1, there were two measures. In one measure, points were given to trained features only. In the other measure, points were given to trained features as well as inferred features (97.32% interrater agreement). Cumulative link mixed models fitted the data. In both measures, a main effect of testing was found, trained features only, $\chi^2(1) = 23.51$, $p < .001$; trained and inferred features, $\chi^2(1) = 96.55$, $p < .001$, but there was no effect of consistency, trained features only, $\chi^2(1) = 0.73$, $p = .39$; trained and inferred features, $\chi^2(1) = 0.20$, $p = .66$; nor an interaction, trained features only, $\chi^2(1) = 1.68$, $p = .19$; trained and inferred features, $\chi^2(1) = 0.28$, $p = .59$. Figure 6C and 6D show that, in both measures, participants forgot semantic features after the delay and to a similar extent across consistency conditions. When all items were analyzed, the same results were found, that is, an effect of testing, trained features only, $\chi^2(1) = 24.57$, $p < .001$; trained and inferred features, $\chi^2(1) = 98.70$, $p < .001$; no effect of consistency, trained features only, $\chi^2(1) = 0.02$, $p = .90$; trained and inferred features, $\chi^2(1) = 0.40$, $p = .53$; and no interaction, trained features only, $\chi^2(1) = 1.08$, $p = .30$; trained and inferred features, $\chi^2(1) = 0.05$, $p = .83$.

Spelling (Measuring P2O)

Data were fitted with logistic LME models. Likelihood ratio tests indicated main effects of testing, $\chi^2(1) = 15.53$, $p < .001$; and consistency, $\chi^2(1) = 15.70$, $p < .001$. However, there was no significant interaction between testing and consistency, $\chi^2(1) = 0.51$, $p = .48$. As shown in Figure 7A, participants were less accurate in spelling inconsistent relative to consistent items both before and after the delay. The decline in accuracy over time was similar across consistency conditions. Response patterns were explored and categorized in the same way as in Experiment 1 (Table 7). Across consistency conditions, the correct response rate decreased as compared with Experiment 1 by approximately 10%, whereas the phonologically derived error rate increased by approximately 10%, and the rate of other errors remained at a similar level. This suggests that, after the delay, participants forgot precise phonology-to-orthography bindings and used alternative sound-to-letter relations, for both consistent and inconsistent items.

Labeling (Measuring S2O)

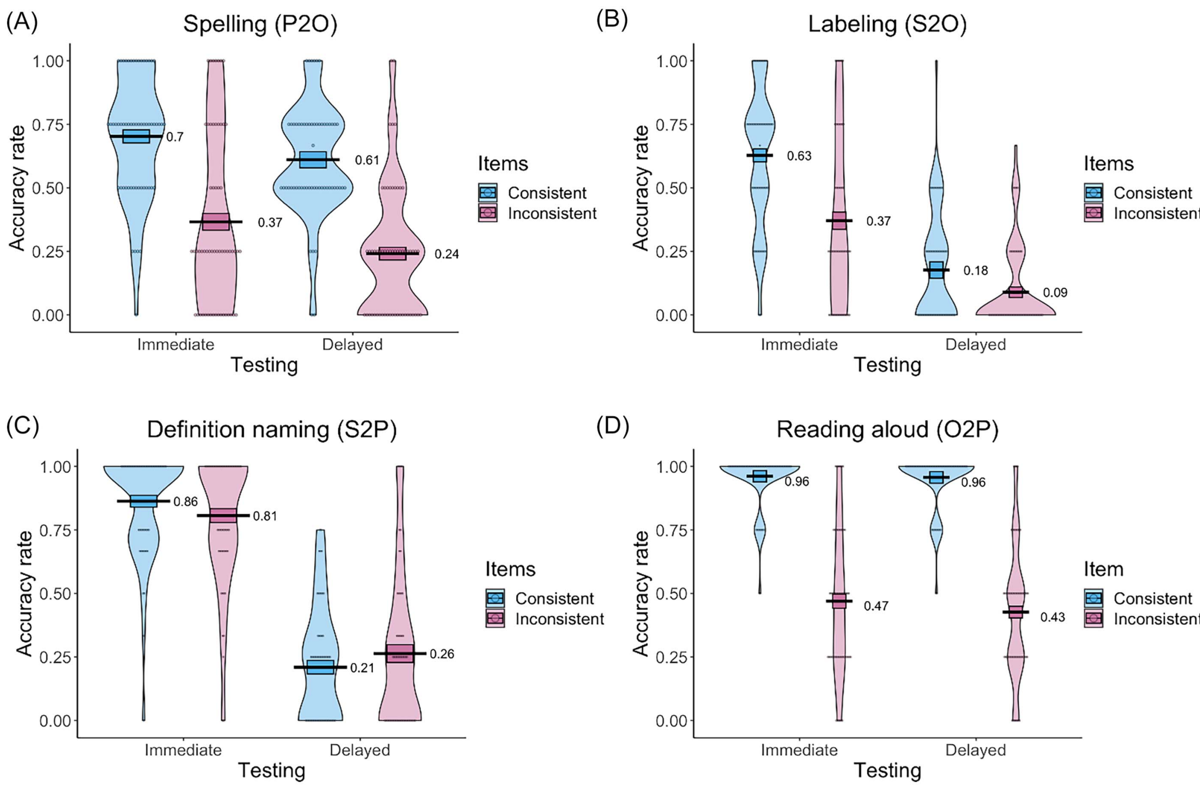
Accuracy data were analyzed with logistic LME models. There were main effects of testing, $\chi^2(1) = 183.93$, $p < .001$; and consistency, $\chi^2(1) = 6.02$, $p = .01$, but no interaction, $\chi^2(1) = 2.05$, $p = .15$. Figure 7B shows that accuracy declined after the delay across consistency conditions. Participants forgot semantics-to-orthography bindings and the amount of forgetting did not differ significantly between consistent and inconsistent items. Analysis of response patterns (Table 7) revealed that, after the delay, there were less correct responses and phonologically derived errors, but substantially more other errors (e.g., responses that were derived from phonology using implausible sound-to-letter relations; other trained pseudowords, correctly or incorrectly spelled; "don't know"). This suggests that bindings between semantic and orthographic representations diminished over time.

Definition Naming (Measuring S2P)

Based on our preregistered hypothesis, we analyzed items that were correctly recognized in the immediate orthographic decision test (83.55% of consistent items and 77.92% of inconsistent items). Accuracy data were fitted with logistic LME models. There was a main effect of testing, $\chi^2(1) = 302.22$, $p < .001$; no main effect of consistency, $\chi^2(1) = 1.2$, $p = .27$; and a significant interaction between testing and consistency, $\chi^2(1) = 5.47$, $p = .02$. Follow-up pairwise comparisons showed a simple effect of testing in both consistent ($z = -11.32$, $p < .001$) and inconsistent conditions ($z = -9.60$, $p < .001$). There was an effect of consistency in the immediate test ($z = -2.25$, $p = .02$), but not in the delayed test ($z = 0.93$, $p = .35$). As illustrated in Figure 7C, participants forgot semantics-to-phonology bindings after the delay in both consistency conditions. However, inconsistent items showed less forgetting compared with consistent items, such that inconsistent and consistent items were no longer significantly different in the delayed test. When all items were analyzed, similar results were found, that is, an effect of testing, $\chi^2(1) = 413.9$, $p < .001$; no effect of consistency, $\chi^2(1) = 0.35$, $p = .55$; and a marginally significant interaction, $\chi^2(1) = 3.27$, $p = .07$.

Figure 7

Performance in the Other Four Posttests Immediately After Self-Teaching and With a Delay



Note. Performance in the (A) spelling test, measuring phonology-to-orthography bindings; (B) labeling test, measuring semantics-to-orthography bindings; (C) definition naming test, measuring semantics-to-phonology bindings; and (D) reading aloud test, measuring orthography-to-phonology bindings. Dots represent participants. Lines represent means. Boxes represent standard errors corrected for the within-participants design. See the online article for the color version of this figure.

Again, response patterns were explored and categorized in the same way as in Experiment 1 (Table 8). Across measures (recognized items and all items) and consistency conditions, there were fewer correct responses but more other errors (e.g., responses that were derived from orthography using implausible letter-to-sound relations; other trained pseudowords, correctly or incorrectly pronounced; "don't know"). It indicates that semantic and phonological representations became less tightly bound over time. Orthographically derived errors remained negligible after the delay for consistent items. There were fewer orthographically derived errors for inconsistent items.

Reading Aloud (Measuring O2P)

Accuracy data were fitted with logistic LME models. Likelihood ratio tests showed a main effect of consistency, $\chi^2(1) = 11.12$, $p < .001$. This suggests that, in general, participants developed less precise orthography-to-phonology bindings for inconsistent items,

Table 7

Response Patterns in the Spelling and Labeling Tests in Immediate and Delayed Testing Conditions

Test	Consistency	Testing	Correct response (%)	Phonologically derived error (%)	Other error (%)
Spelling	Consistent	Immediate	70.26	25.43	4.31
	Consistent	Delayed	61.04	35.93	3.03
	Inconsistent	Immediate	36.64	50.86	12.50
	Inconsistent	Delayed	24.14	64.22	12.63
Labeling	Consistent	Immediate	62.77	25.97	11.26
	Consistent	Delayed	16.74	9.25	74.00
	Inconsistent	Immediate	37.07	43.97	18.97
	Inconsistent	Delayed	8.37	20.26	71.37

Table 8
Response Patterns in the Definition Naming and Reading Aloud Tests in Immediate and Delayed Testing Conditions

Test	Consistency	Testing	Correct response (%)	Orthographically derived error (%)	Other error (%)
Definition naming (recognized items)	Consistent	Immediate	87.05	1.04	11.92
	Consistent	Delayed	22.80	1.55	75.65
	Inconsistent	Immediate	78.89	11.11	10.00
	Inconsistent	Delayed	26.67	2.22	71.11
Definition naming (all items)	Consistent	Immediate	86.21	1.72	12.07
	Consistent	Delayed	21.98	1.72	76.29
	Inconsistent	Immediate	80.17	9.91	9.91
	Inconsistent	Delayed	23.71	2.59	73.71
Reading aloud	Consistent	Immediate	96.12	1.29	2.59
	Consistent	Delayed	95.69	1.29	3.01
	Inconsistent	Immediate	46.98	44.83	8.19
	Inconsistent	Delayed	42.67	53.02	4.31

Note. Recognized items refer to items that are correctly recognized in the immediate orthographic decision test.

compared with consistent items. There was no effect of testing, $\chi^2(1) = 0.68$, $p = .41$; and no interaction, $\chi^2(1) = 0.14$, $p = .71$. As demonstrated in Figure 7D, there was no evidence of substantial forgetting following the delay. Table 8 showed that, for consistent items, response patterns were similar before and after the delay. For inconsistent items, there were fewer correct responses and other errors, but more orthographically derived errors.

Discussion

In this experiment, participants from Experiment 1 returned to retake the posttests, with a mean delay of 77 days. Their posttest performance in Experiment 1 (immediate testing) and the current experiment (delayed testing) were analyzed. We discuss findings with respect to our predictions and hypotheses. Results from exploratory analyses are also discussed.

Forgetting of Lexical Knowledge

Across posttests, participants performed significantly less well at delayed compared with immediate tests. This suggests that participants forgot various aspects of lexical knowledge after the delay, including orthographic, phonological, and semantic representations, phonology-to-orthography, semantics-to-orthography, and semantics-to-phonology bindings. The results support our predictions and are in line with previous literature that documented forgetting of newly acquired lexical knowledge (e.g., Nation et al., 2007; Tamminen et al., 2017; Vlach & Sandhofer, 2014; Wang et al., 2011).

The only exception to forgetting was orthography-to-phonology bindings. In the reading aloud test, participants maintained the same level of accuracy after the delay, across consistency conditions. It seems that, once orthography-to-phonology bindings were acquired, they remained robust and did not decay. It is possible that participants used long-term memory of letter-to-sound mappings and known words to support orthography-to-phonology bindings and the retention. Reading consistent items was highly accurate (96%) in both immediate and delayed testing conditions, as participants could decode pseudowords using knowledge of common letter-to-sound mappings without necessarily remembering the pseudowords. For inconsistent words, the accuracy rate was 47% on the immediate test and remained at a similar level after the delay (43%). It is possible that participants developed a strategy to memorize O2P for these inconsistent pseudowords. They might search for similar known words in the mental lexicon and apply their pronunciations to these pseudowords. For example, to learn that "zeak" was pronounced as /zeɪk/, participants might find a similar sounding word like "break" and then apply its /eɪk/ to "zeak." This might have supported retention of orthography-to-phonology bindings for inconsistent pseudowords.

The Effect of Consistency on Forgetting

As predicted, participants showed less forgetting of semantics-to-phonology bindings for inconsistent than consistent items in the definition naming test. This supports the episodic context account of retrieval-based learning (Karpicke et al., 2014). That is, successful retrieval of inconsistent items is more difficult and requires active recall of semantics-to-phonology bindings, which in turn benefits memory retention and mitigates forgetting of semantics-to-phonology bindings. More specifically, when readers encounter an inconsistent novel written word in semantically rich sentences (e.g., "when the fish tank was clean, he turned off the zeak and put it away"), semantic cues point readers to an oral vocabulary item (the fish tank cleaner). However, according to the common letter-to-sound mappings, "zeak" should sound /ziːk/, which does not match any known oral vocabulary. Readers might use the effective semantic cues to reinstate oral vocabulary learning episodes in which they encountered the fish tank cleaner. Importantly, they might actively recall the spoken name of the fish tank cleaner (semantics-to-phonology bindings) to determine the correct pronunciation for the novel written word "zeak." This retrieval process enhances memory retention of semantics-to-phonology bindings and protects them from forgetting.

The episodic context account of retrieval-based learning also predicts that inconsistency should mitigate forgetting of semantic representations, as readers might actively recall additional semantic features of oral vocabulary items during reinstatement (e.g., the fish tank cleaner has a sponge and is shaped like an arm). However, participants showed a similar amount of forgetting for consistent and inconsistent items in the semantic feature recall test. The lack of a beneficial effect of inconsistency might be because participants did not actively recall these additional semantic features. Recalling additional semantic features might not be necessary to retrieve the inventions.

Consistency also modulated the forgetting of phonology. Participants exhibited a greater decline in both target recognition and foil rejection for inconsistent items compared with consistent items. Consistency did not modulate forgetting of orthographic representations, phonology-to-orthography, or semantics-to-orthography bindings.

The Main Effect of Consistency on the Development of Lexical Knowledge

By analyzing the main effect of consistency, we explored the general effect of consistency on development of lexical knowledge, considering participants' lexical knowledge at two time points (i.e., immediately after learning and with a delay). A main effect of consistency was found in the phonological decision (measuring phonological representations), spelling (phonology-to-orthography bindings), labeling (semantics-to-orthography bindings), and reading aloud (orthography-to-phonology) tests. Across these tests, inconsistency had a negative effect. That is, development of these aspects of lexical knowledge was more difficult in inconsistent than consistent items. No main effect was found in the other tests (orthographic decision test, measuring orthographic representations; semantic feature recall task, measuring semantic representations; definition naming test, measuring semantics-to-phonology bindings). It should be noted, however, that the null effect of consistency on semantics-to-phonology bindings may be a consequence of two opposing effects, that is, a negative effect on learning and a positive effect on memory retention, which together cancel each other out in the delayed posttest and diminish the main effect (i.e., across immediate and delayed tests).

General Discussion

To our knowledge, the present study was the first study that systematically investigates the effect of consistency on learning and forgetting of all aspects of lexical knowledge (i.e., orthographic, phonological, and semantic representations, orthography–phonology, orthography–semantics, semantics–phonology bindings). Experiment 1 focused on the effect of consistency on learning, and Experiment 2 on forgetting. As discussed earlier, we tested predictions from the self-teaching framework (Pritchard et al., 2018; Share, 1995; Ziegler et al., 2014), spelling pronunciation account (Elbro & de Jong, 2017), and episodic context account of retrieval-based learning (Karpicke et al., 2014). We now discuss more general issues.

Both Learning and Forgetting Are Inherent in Lexical Development

The present study tracked learning and forgetting of lexical knowledge. In Experiment 1, participants were instructed to learn pseudowords through oral vocabulary training and self-teaching. Posttests demonstrated that participants learned various aspects of lexical knowledge. When these participants were tested again after a mean delay of 77 days in Experiment 2, results showed clear evidence of forgetting in all aspects of lexical knowledge, except orthography-to-phonology bindings. Together, the two experiments suggest that lexical knowledge development is a dynamic process, consisting of not only gains in knowledge through learning, but also loss of the knowledge due to forgetting. Most previous word learning studies focused on learning, without considering retention of the learned lexical knowledge over time. There have, however, been some efforts to understand the factors that affect retention of lexical knowledge, for example, the number of exposures (e.g., Nation et al., 2007), temporal spacing of learning materials (Wegener, Wang, Beyersmann, Reichle, et al., 2023), and sleep (e.g., Tamminen et al., 2017). We encourage future research to investigate not only word learning but also its long-term retention, to better understand the dynamic and complex development of lexical knowledge.

Inconsistency Has Multifaceted Effects on the Development of Lexical Knowledge

Inconsistency showed complex and multifaceted effects on the development of lexical knowledge, as summarized in Table 9. First, inconsistency impacted both word learning and forgetting. Previous studies focused on the effect of consistency on learning (Burt & Blackwell, 2008; McKay et al., 2008; Murray et al., 2022; Wang et al., 2012, 2013). It was not clear whether and how consistency would affect word forgetting. The present study filled in the gap and found that inconsistency also impacted word forgetting (Table 9). This highlights that consistency is an important factor throughout the development of lexical knowledge.

Second, inconsistency affected most aspects of lexical knowledge, including phonological representations, phonology-to-orthography, semantics-to-orthography, semantics-to-phonology, and orthography-to-phonology bindings. Previous studies tested the effect on development of orthographic representations and orthography–phonology bindings only (e.g., Burt & Blackwell, 2008; McKay et al., 2008; Murray et al., 2022; Wang et al., 2012, 2013). We extended the scope

Table 9
Summary of Effects of Inconsistency on the Development of Lexical Knowledge and Theoretical Accounts That Were Supported

Lexical knowledge	Learning (Experiment 1)	Forgetting (Experiment 2)
Orthographic representation	No	No
Phonological representation	Yes (negative). Spelling pronunciation account	Yes (negative)
Semantic representation	No	No
Phonology-to-orthography binding	Yes (negative). Self-teaching framework	No
Semantics-to-orthography binding	Yes (negative). Self-teaching framework	No
Semantics-to-phonology binding	Yes (negative)	Yes (positive). Episodic context account of retrieval-based learning
Orthography-to-phonology binding	Yes (negative). Self-teaching framework	No

Note. A negative effect on forgetting means more forgetting and a positive effect on forgetting means less forgetting.

of investigation and tested other aspects of lexical knowledge too (semantic and phonological representations, semantics-to-orthography, and semantics-to-phonology bindings). As summarized in Table 9, inconsistency impacted development of all tested aspects of lexical knowledge except orthographic and semantic representations.

The effects on semantics-to-orthography and semantics-to-phonology bindings are especially noteworthy. Consistency is a property of word form (i.e., relations between orthographic and phonological forms), but it also affected how well semantics were bound to forms. In Experiment 1, learning of semantics-to-orthography and semantics-to-phonology bindings was impaired by inconsistency. This suggests that inconsistent relations between orthographic and phonological forms pose challenges for learning to use precise forms to represent meaning. In Experiment 2, we found that inconsistency reduced forgetting of semantics-to-phonology bindings. According to Karpicke et al. (2014), inconsistency might have encouraged readers to use semantic cues from sentences to actively retrieve semantics-to-phonology bindings. This strengthens memory and mitigates forgetting of semantics-to-phonology bindings. Our findings provide clear evidence that word learning involves complex interactions between orthography, phonology, and semantics, which have immediate effects on learning outcomes and lasting effects on long-term retention.

Third, inconsistency had not only negative but also positive effects on the development of lexical knowledge. As reported by previous studies and this study (Table 9), inconsistency caused difficulties in word learning. However, it benefited long-term retention of semantics-to-phonology bindings. This pattern might be counterintuitive but is in line with a large body of literature related to desirable difficulties (Bjork, 1994). Difficulties, deliberately introduced to training (e.g., reduction in feedback, testing rather than restudying materials), impair initial learning. However, faced with difficulties, learners more actively encode and retrieve learning materials, which benefits long-term memory retention of the materials.

Generalizability, Limitations, and Future Research

Our experiments investigated the effect of consistency on lexical knowledge development using a self-teaching paradigm. The findings might not generalize to other learning approaches, for example, sight word instruction (Castles et al., 2018) or learning through silent reading only (Hulme et al., 2022). More research is needed to understand how inconsistency would impact lexical knowledge development under these different learning conditions. We tested adults. Future research could test children and investigate whether the findings generalize to children.

A priori power analysis was not conducted for this study, as some effect sizes of interest were unavailable in previous literature. Future research could use the current data set to inform power analyses and guide study planning. Replication studies are needed to examine findings from the present study. We also welcome researchers to further explore response patterns in the current data set.

Retrieval success during self-teaching was operationalized as successful recognition of written pseudowords as invention names in the immediate orthographic decision test, as justified in the sections "Operationalization of successful retrieval during self-teaching." As pointed out by a reviewer, this operationalization is not ideal. First, recognition of written pseudowords as invention names is not identical to retrieval of oral vocabulary episodes. Second, self-teaching and the orthographic decision test took place at different time points. Retrieval success at one time point does not necessarily indicate retrieval success at the other. However, although our operationalization was imperfect, Experiment 2 still showed clear evidence for the protective effect of inconsistency on memory retention of semantics-to-phonology bindings. When items that were successfully recognized in the immediate orthographic decision test were analyzed, a significant interaction between testing and consistency was found, indicating that inconsistent items were forgotten less compared with consistent items. When all items were analyzed, similar results were found, and the interaction approached significance.

The order of posttests was fixed in the present study, for reasons that were explained in the Procedure section in Experiment 1. It is not clear whether changing the order of posttests would give rise to different results. Future research could consider counterbalancing posttests of the same type (tests on individual lexical components; tests on bindings) across participants. Due to time constraints on online experiments (Revilla & Höhne, 2020), orthography-to-semantics and phonology–semantics bindings were not tested. Future studies could consider including tests that assess these two aspects of lexical knowledge.

As a reviewer commented, the orthographic and phonological decision tests assessed recognition memory, requiring participants to determine whether a stimulus was a trained item. In contrast, other posttests assessed recall, requiring participants to produce certain aspects of lexical knowledge of previously trained items. According to Mandler (1980), recognition memory is supported by both familiarity, where participants recognize items because they feel familiar, and recollection, in which participants retrieve previous study episodes to recall specific information. In contrast, tests of recall involve recollection exclusively. Therefore, findings from recognition memory tests might not generalize if lexical knowledge was assessed using recall tests instead, and vice versa. Future research could investigate how different types of tests modulate the consistency effect.

Experiments 1 and 2 had a long interval (M = 77 days), which allowed us to test long-term retention of acquired lexical knowledge and the effect of consistency. Radvansky et al. (2022) proposed that retention can be studied at various phases, including working memory (first few minutes following encoding), early long-term memory (the following 12 hr), transitional long-term memory (the following week), and long-lasting long-term memory (beyond the week). Future research could investigate lexical knowledge development and the effect of consistency at different retention phases.

Conclusion

The present study systematically investigated the effect of consistency on learning and forgetting of various aspects of lexical knowledge (i.e., orthographic, phonological, and semantic representations, phonology-to-orthography, semantics-to-orthography, semantics-to-phonology, and orthography-to-phonology bindings). We found that inconsistency impaired learning of phonological representations, phonology–orthography, and semantics-to-orthography bindings. The findings support the self-teaching framework (Pritchard et al., 2018; Share, 1995; Ziegler et al., 2014) and spelling pronunciation account (Elbro & de Jong, 2017). Several months later (M = 77 days), participants were tested again and showed a significant amount of forgetting in all aspects of lexical knowledge except orthography-to-phonology bindings. Inconsistent items showed more forgetting of phonology compared with consistent items. However, inconsistency benefited memory retention of semantics-to-phonology bindings such that less forgetting was observed for inconsistent relative to consistent items. This supports the episodic context account of retrieval-based learning (Karpicke et al., 2014). Our study revealed complex and multifaceted effects of consistency on lexical knowledge development.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *The CELEX Lexical Database (CD-ROM)* [Computer software]. https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_2339741

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–206). MIT Press. https://doi.org/10.7551/mitpress/4561.003.0011

Brown, G. D. A., Romney, J. L., Loosemore, R. P. W., & Watson, F. L. (1991). A neural net model of spelling development. In T. Kohonen, K. Mäkisara, O. Simula, & J. Kangas (Eds.), *Artificial neural networks* (pp. 1727–1730). North-Holland. https://doi.org/10.1016/B978-0-444-89178-5.50174-3

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*, Article 1116. https://doi.org/10.3389/fpsyg.2016.01116

Bürki, A., Spinelli, E., & Gareth Gaskell, M. (2012). A written word is worth a thousand spoken words: The influence of spelling on spoken-word production. *Journal of Memory and Language*, *67*(4), 449–467. https://doi.org/10.1016/j.jml.2012.08.001

Burt, J. S., & Blackwell, P. (2008). Sound–spelling consistency in adults' orthographic learning. *Journal of Research in Reading*, *31*(1), 77–96. https://doi.org/10.1111/j.1467-9817.2007.00362.x

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Chaves, N., Ginestet, E., & Bosse, M.-L. (2020). Lexical orthographic knowledge acquisition in adults: The whole-word visual processing impact. *European Review of Applied Psychology*, *70*(1), Article 100520. https://doi.org/10.1016/j.erap.2019.100520

Chee, Q. W., Chow, K. J., Yap, M. J., & Goh, W. D. (2020). Consistency norms for 37,677 English words. *Behavior Research Methods*, *52*(6), 2535–2555. https://doi.org/10.3758/s13428-020-01391-7

Christensen, R. H. B. (2024). *ordinal: Regression models for ordinal data* (Version 2023.12-4.1) [Computer software]. https://cran.r-project.org/web/packages/ordinal/index.html

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*(4), 589–608. https://doi.org/10.1037/0033-295X.100.4.589

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

Cunningham, A. E., Perry, K. E., Stanovich, K. E., & Share, D. L. (2002). Orthographic learning during reading: Examining the role of self-teaching. *Journal of Experimental Child Psychology*, *82*(3), 185–199. https://doi.org/10.1016/S0022-0965(02)00008-5

Elbro, C., & de Jong, P. F. (2017). The role of spelling pronunciations: Orthographic learning is verbal learning. In K. Cain, D. L. Compton, & R. K. Parrila (Eds.), *Theories of reading development* (pp. 169–190). John Benjamins Publishing. https://doi.org/10.1075/swll.15.10elb

Fitt, S. (2000). *Documentation and user guide to UNISYN lexicon and post-lexical rules* [Data set]. https://www.cstr.ed.ac.uk/projects/unisyn/

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*(3), 662–720. https://doi.org/10.1037/0033-295X.111.3.662

Holbert, C. (2022, April 16). *Nonparametric two-way ANOVA*. Charles Holbert. https://www.cfholbert.com/blog/nonparametric_two_way_anova/

Hulme, R. C., & Rodd, J. M. (2021). Learning new word meanings from story reading: The benefit of immediate testing. *PeerJ*, *9*, Article e11693. https://doi.org/10.7717/peerj.11693

Hulme, R. C., Shapiro, L. R., & Taylor, J. S. H. (2022). Learning new words through reading: Do robust spelling–sound mappings boost learning of word forms and meanings? *Royal Society Open Science*, *9*(12), Article 210555. https://doi.org/10.1098/rsos.210555

Jared, D. (1997). Spelling–sound consistency affects the naming of high-frequency words. *Journal of Memory and Language*, *36*(4), 505–529. https://doi.org/10.1006/jmla.1997.2496

Jared, D. (2002). Spelling–sound consistency and regularity effects in word naming. *Journal of Memory and Language*, *46*(4), 723–750. https://doi.org/10.1006/jmla.2001.2827

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

Johnson, C. J., Beitchman, J. H., & Brownlie, E. B. (2010). Twenty-year follow-up of children with and without speech-language impairments: Family, educational, occupational, and quality of life outcomes. *American Journal of Speech-Language Pathology*, *19*(1), 51–65. https://doi.org/10.1044/1058-0360(2009/08-0083)

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 61, pp. 237–284). Elsevier. https://doi.org/10.1016/B978-0-12-800283-4.00007-1

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. https://doi.org/10.3758/s13428-011-0118-4

Koizumi, R. (2013). Vocabulary and speaking. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (1st ed., pp. 1–7). Wiley. https://doi.org/10.1002/9781405198431.wbeal1431

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Piaskowski, J., Riebl, H., & Singmann, H. (2024). *emmeans: Estimated marginal means, aka least-squares means*

(Version 1.10.4) [Computer software]. https://cran.r-project.org/web/packages/emmeans/index.html

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Lien, B. (2017). *Applying the self-teaching hypothesis to adults: The effects of reading condition and syllable length on orthographic learning* [Undergraduate Thesis, University of Pittsburgh]. Electronic Theses and Dissertations. https://d-scholarship.pitt.edu/30869/1/Bphil%20ETD_Lien.pdf

Makowski, D. (2018). The psycho package: An efficient and publishing-oriented workflow for psychological science. *Journal of Open Source Software*, *3*(22), Article 470. https://doi.org/10.21105/joss.00470

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*(3), 252–271. https://doi.org/10.1037/0033-295X.87.3.252

McKay, A., Davis, C., Savage, G., & Castles, A. (2008). Semantic involvement in reading aloud: Evidence from a nonword training study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1495–1517. https://doi.org/10.1037/a0013357

Mousikou, P., Sadat, J., Lucas, R., & Rastle, K. (2017). Moving beyond the monosyllable in models of skilled reading: Mega-study of disyllabic nonword reading. *Journal of Memory and Language*, *93*, 169–192. https://doi.org/10.1016/j.jml.2016.09.003

Murray, L., Wegener, S., Wang, H.-C., Parrila, R., & Castles, A. (2022). Children processing novel irregular and regular words during reading: An eye tracking study. *Scientific Studies of Reading*, *26*(5), 417–431. https://doi.org/10.1080/10888438.2022.2030744

Nation, K. (1997). Children's sensitivity to rime unit frequency when spelling words and nonwords. *Reading and Writing*, *9*(5), 321–338. https://doi.org/10.1023/A:1007938810898

Nation, K., Angell, P., & Castles, A. (2007). Orthographic learning via self-teaching in children learning to read English: Effects of exposure, durability, and context. *Journal of Experimental Child Psychology*, *96*(1), 71–84. https://doi.org/10.1016/j.jecp.2006.06.004

Olsson, C. A., McGee, R., Nada-Raja, S., & Williams, S. M. (2013). A 32-year longitudinal study of child and adolescent pathways to well-being in adulthood. *Journal of Happiness Studies*, *14*(3), 1069–1083. https://doi.org/10.1007/s10902-012-9369-8

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Perfetti, C., & Hart, L. (2002). Lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189–213). John Benjamins Publishing. https://www.lrdc.pitt.edu/perfettilab/pubpdfs/Lexical%20quality%20hypothesis-%20Hart.pdf

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

Pritchard, S. C., Coltheart, M., Marinus, E., & Castles, A. (2018). A computational model of the self-teaching hypothesis based on the dual-route cascaded model of reading. *Cognitive Science*, *42*(3), 722–770. https://doi.org/10.1111/cogs.12571

Protopapas, A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods*, *39*(4), 859–862. https://doi.org/10.3758/BF03192979

Qiu, Y., & Taylor, J. S. H. (2025a, July 4). *Letter-sound inconsistency impacts word learning*. Open Science Framework. https://osf.io/dqcbw/

Qiu, Y., & Taylor, J. S. H. (2025b, July 4). *Word forgetting and the effect of letter-sound inconsistency*. Open Science Framework. https://osf.io/6ty32/

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Radvansky, G. A., Doolen, A. C., Pettijohn, K. A., & Ritchey, M. (2022). A new look at memory retention and forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(11), 1698–1723. https://doi.org/10.1037/xlm0001110

Ranbom, L. J., & Connine, C. M. (2011). Silent letters are activated in spoken word recognition. *Language and Cognitive Processes*, *26*(2), 236–261. https://doi.org/10.1080/01690965.2010.486578

Rastle, K., & Coltheart, M. (1999). Lexical and nonlexical phonological priming in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(2), 461–481. https://doi.org/10.1037/0096-1523.25.2.461

Revilla, M., & Höhne, J. K. (2020). How long do respondents think online surveys should be? New evidence from two online panels in Germany. *International Journal of Market Research*, *62*(5), 538–545. https://doi.org/10.1177/1470785320943049

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*(2), 151–218. https://doi.org/10.1016/0010-0277(94)00645-2

Share, D. L. (1999). Phonological recoding and orthographic learning: A direct test of the self-teaching hypothesis. *Journal of Experimental Child Psychology*, *72*(2), 95–129. https://doi.org/10.1006/jecp.1998.2481

Siegelman, N., Kearns, D. M., & Rueckl, J. G. (2020). Using information-theoretic measures to characterize the structure of the writing system: The case of orthographic-phonological regularities in English. *Behavior Research Methods*, *52*(3), 1292–1312. https://doi.org/10.3758/s13428-019-01317-y

Smejkalova, A., & Chetail, F. (2023). Learning new words by reading books: Does semantic information help? *Quarterly Journal of Experimental Psychology*, *76*(3), 568–582. https://doi.org/10.1177/17470218221095735

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

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

Tamminen, J., Lambon Ralph, M. A., & Lewis, P. A. (2017). Targeted memory reactivation of newly learned words during sleep triggers REM-mediated integration of new memories and existing knowledge. *Neurobiology of Learning and Memory*, *137*, 77–82. https://doi.org/10.1016/j.nlm.2016.11.012

Taylor, J. S. H., Plunkett, K., & Nation, K. (2011). The influence of consistency, frequency, and semantics on learning to read: An artificial orthography paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 60–76. https://doi.org/10.1037/a0020126

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

Trudeau, J. J. (2006). *Semantic contributions to word naming with artificial lexicons* [Doctoral dissertation, University of Connecticut]. ProQuest Dissertations and Theses Global. https://www.proquest.com/docview/305321622?fromopenview=true&pq-origsite=gscholar&sourcetype=Dissertations%20&%20Theses

Vlach, H. A., & Sandhofer, C. M. (2014). Retrieval dynamics and retention in cross-situational statistical word learning. *Cognitive Science*, *38*(4), 757–774. https://doi.org/10.1111/cogs.12092

Voeten, C. C. (2023). *buildmer: Stepwise elimination and term reordering for mixed-effects regression* (Version 2.11) [Computer software]. https://cran.r-project.org/web/packages/buildmer/index.html

Wang, H.-C., Castles, A., & Nickels, L. (2012). Word regularity affects orthographic learning. *Quarterly Journal of Experimental Psychology*, *65*(5), 856–864. https://doi.org/10.1080/17470218.2012.672996

Wang, H.-C., Castles, A., Nickels, L., & Nation, K. (2011). Context effects on orthographic learning of regular and irregular words. *Journal of Experimental Child Psychology*, *109*(1), 39–57. https://doi.org/10.1016/j.jecp.2010.11.005

Wang, H.-C., Nickels, L., Nation, K., & Castles, A. (2013). Predictors of orthographic learning of regular and irregular words. *Scientific Studies of Reading*, *17*(5), 369–384. https://doi.org/10.1080/10888438.2012.749879

Weekes, B. S., Castles, A. E., & Davies, R. A. (2006). Effects of consistency and age of acquisition on reading and spelling among developing readers. *Reading and Writing*, *19*(2), 133–169. https://doi.org/10.1007/s11145-005-2032-6

Wegener, S., Wang, H.-C., Beyersmann, E., Nation, K., Colenbrander, D., & Castles, A. (2023). Orthographic expectancies in the absence of contextual support. *Scientific Studies of Reading*, *27*(2), 187–197. https://doi.org/10.1080/10888438.2022.2127356

Wegener, S., Wang, H.-C., Beyersmann, E., Reichle, E. D., Nation, K., & Castles, A. (2023). The effect of spacing versus massing on orthographic learning. *Reading Research Quarterly*, *58*(3), 361–372. https://doi.org/10.1002/rrq.492

Wegener, S., Wang, H.-C., de Lissa, P., Robidoux, S., Nation, K., & Castles, A. (2018). Children reading spoken words: Interactions between vocabulary and orthographic expectancy. *Developmental Science*, *21*(3), Article e12577. https://doi.org/10.1111/desc.12577

Wegener, S., Wang, H.-C., Nation, K., & Castles, A. (2020). Tracking the evolution of orthographic expectancies over building visual experience. *Journal of Experimental Child Psychology*, *199*, Article 104912. https://doi.org/10.1016/j.jecp.2020.104912

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Ziegler, J. C., & Ferrand, L. (1998). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin & Review*, *5*(4), 683–689. https://doi.org/10.3758/BF03208845

Ziegler, J. C., Ferrand, L., & Montant, M. (2004). Visual phonology: The effects of orthographic consistency on different auditory word recognition tasks. *Memory & Cognition*, *32*(5), 732–741. https://doi.org/10.3758/BF03195863

Ziegler, J. C., Perry, C., & Zorzi, M. (2014). Modelling reading development through phonological decoding and self-teaching: Implications for dyslexia. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), Article 20120397. https://doi.org/10.1098/rstb.2012.0397

Ziegler, J. C., Petrova, A., & Ferrand, L. (2008). Feedback consistency effects in visual and auditory word recognition: Where do we stand after more than a decade? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(3), 643–661. https://doi.org/10.1037/0278-7393.34.3.643

Received October 30, 2024
Revision received May 9, 2025
Accepted June 20, 2025 ■