The hidden depths of new word knowledge: Using graded measures of orthographic and semantic learning to measure vocabulary acquisition

Jessie Ricketts[1], Nicola Dawson[2] and Robert Davies[3]

[1]Department of Psychology, Royal Holloway, University of London, Egham Hill, Egham, Surrey, TW20 0EX, UK

[2]Department of Experimental Psychology, University of Oxford, Oxford OX2 6GG, UK

[3]Department of Psychology, Lancaster University, Bailrigg, Lancaster, LA1 4YW, UK

Corresponding author: Jessie Ricketts, jessie.ricketts@rhul.ac.uk

Cite as: Ricketts, J., Dawson, N., & Davies, R. (accepted). The hidden depths of new word knowledge: Using graded measures of orthographic and semantic learning to measure vocabulary acquisition. *Learning and Instruction*.

Acknowledgements

We would like to thank Rachel Tomkinson, Fay Bainbridge and Natascha Ahmed for their assistance with data collection and all schools and families for participating. Amrita Bains and Eleanor Gray contributed to manuscript preparation. We would also like to acknowledge Emmanuel Keuleers' assistance in generating R scripts for calculating Levenshtein distance measures earlier in this project.

Funding. This work was supported by the Economic and Social Research Council (grant number ES/K008064/1) and funding from the University of Reading.

Open science. Analysis data and code are shared through an OSF repository accessible at: https://osf.io/e5gzk/?view_only=a43914620dae4cc1b56bf3c15ef8d6c6

Highlights

- Children learned more words that had been taught with, compared to without, visual forms
- Unusually, retention of word knowledge was assessed longitudinally, over a period of eight months
- Word knowledge was well-retained over time
- We introduce new learning measures that capture the incremental nature of vocabulary acquisition
- These measures revealed learning effects that would be masked by traditional measures

Abstract

We investigated whether the presence of orthography promotes new word learning (orthographic facilitation). In Study 1 (N = 41) and Study 2 (N = 74), children were taught 16 unknown polysyllabic words. Half of the words appeared with orthography present and half without orthography. Learning assessments captured the degree of semantic and orthographic learning; they were administered one week after teaching (Studies 1 and 2), and, unusually, eight months later (Study 1 only). Bayesian analyses indicated that the presence of orthography was associated with more word learning, though this effect was estimated with more certainty for orthographic than semantic learning. Newly learned word knowledge was well retained over time, indicating that our paradigm was sufficient to support long-term learning. Our paradigm provides an example of how word learning studies can look beyond simple accuracy measures to reveal the cumulative nature of lexical learning.

Key words: orthographic facilitation; vocabulary; word learning; oral vocabulary acquisition; reading

The hidden depths of new word knowledge: Using graded measures of orthographic and semantic learning to measure vocabulary acquisition

1. General Introduction

Vocabulary knowledge is essential for processing language in everyday life and it is vital that we know how to optimise vocabulary teaching. One strategy with growing empirical support is *orthographic facilitation*: children and adults are more likely to learn new spoken words that are taught with their orthography (visual word forms; for reviews, see Colenbrander, Miles & Ricketts, 2019; Ehri, 2014; 2020). Across two studies, we used an experimental word-learning paradigm to investigate theoretical accounts of orthographic facilitation and to evaluate how orthographic forms can be used to maximise oral vocabulary learning. We used fine-grained measures to assess the outcomes of learning with the aim of capturing partial word learning and the incremental nature of word knowledge. A better understanding of the role of orthography in vocabulary development will inform theory, and practical approaches to teaching.

Spoken and written communication requires knowledge of many words. Before learning to read, children learn the spoken forms (phonology) and meanings (semantics) of words from their spoken language experiences. As reading commences, representations of known words can expand to include visual forms (orthography), and new word learning can involve learning orthography as well as phonology and semantics. The lexical quality hypothesis (Perfetti & Hart, 2002) emphasises the importance of knowing many words and having ‘high quality’ representations for these lexical entries. High quality lexical representations include detailed information about phonology, semantics and orthography that is well integrated such that one component of the representation (phonology, orthography, semantics) will readily activate the other components. For example, when a

child reads a word for which they have a high quality representation, the orthographic form activates an accurate phonological form and rich semantic information. In alphabetic languages, speech sounds (phonemes) are represented by a finite set of visual symbols such as letters. In such languages, when words are unknown (i.e. the lowest quality), the orthographic form can still activate phonology because there are systematic relationships between orthography and phonology (Valentini, Ricketts, Pye, & Houston-Price, 2018). Similarly, on hearing a novel word, a child may use what they know about the way that the sounds in that word are usually spelled to set up expectations about the word's visual form, or an 'orthographic skeleton' (Wegener et al., 2017).

In emphasising the importance of orthography as well as phonology and semantics in lexical representations, the lexical quality hypothesis (Perfetti & Hart, 2002) is consistent with the prediction that orthographic facilitation will occur in word learning. In other words, when we teach new words, exposing learners to orthography as well as to phonology and semantics should result in greater learning and knowledge about words. A recent systematic review by Colenbrander et al. (2019) showed strong evidence that the presence of orthography supports the learning of phonological forms. There was also evidence that the presence of orthography aids semantic learning (e.g., Li et al., 2016; Ricketts et al., 2009; Rosenthal & Ehri, 2008). For example, in Rosenthal and Ehri (2008), children aged 7-8 years who saw and heard 'gam' whilst hearing its definition (orthography present condition) showed greater recall for its phonological form and meaning than children who just heard 'gam' and its definition (orthography absent condition). However, in some studies this effect was marginal (Ricketts, Dockrell, Patel, Charman, & Lindsay, 2015) or nonsignificant (e.g., Chambre, Ehri, & Ness, 2017).

Why might orthographic facilitation occur? Compared to the continuous speech stream, orthography clearly marks where one word, letter or sentence ends, and the next begins. The speech that we hear comes and goes, whereas the written word stays put on the page, allowing more time for processing. Moreover, whilst spoken and written representations of language vary across contexts as a result of changes in voice, accent, handwriting, font and so on, arguably, this is more pronounced for speech. Therefore, orthographic forms may be more readily learned than phonological forms, providing a more effective anchoring device, or hook, on which to hang semantic information (for similar ideas, see Rosenthal & Ehri, 2008; Krepel, de Bree, & de Jong, 2020).

In alphabetic languages, orthographic facilitation could also be driven by the systematic relationships that exist between letters and sounds. In some languages (e.g., Spanish, Finnish), orthography-phonology mappings are highly consistent (Seymour, Aro, & Erskine, 2003; Share, 2008) and orthography provides a reliable cue to phonology and vice versa. In less transparent languages, like English, words contain many consistent mappings, but they can also include spelling patterns that are not pronounced in the usual way (e.g., the letter *s* in *sugar* versus *sit*) and sounds that are not spelled in the usual way (e.g., /ɒ/ in *yacht* versus *pot*). For English, orthography will be a more reliable cue to phonology for more consistent words (i.e. words with a greater number of consistent spelling-sound mappings). For example, *yacht* contains two consistent mappings represented by the letters *y* and *t* whereas *pot* contains three consistent mappings represented by *p*, *o* and *t*. Therefore, it seems plausible that the degree of spelling-sound consistency will moderate the orthographic facilitation effect, with the presence of orthography more useful for learning the phonological forms, and therefore the meanings, of more consistent words.

Ricketts et al. (2009) investigated orthographic facilitation in a paired-associate learning paradigm in which children learned 12 nonword-referent mappings. In this study, consistency varied across nonwords. A third of the stimuli contained only consistently spelled phonemes (e.g., the vowel sound /ʌ/ as *bus*), while the remaining stimulus words contained either an inconsistent vowel (e.g., /iː/ can be spelled as in feet and feat) or an inconsistent consonant (e.g., /tʃ/ can be spelled as in much or clutch). There was no significant interaction between the presence of orthography and consistency on measures of oral vocabulary learning. However, it may be premature to conclude that orthographic facilitation is not moderated by consistency. Close inspection of Ricketts et al.'s (2009) data by participants and by items indicates there may have been subtle effects that could not be detected due to the small number of items included in the study (Jubenville, Sénéchal, & Malette, 2014; Ricketts, et al., 2015).

Jubenville et al. (2014) increased the number of items by manipulating orthography and consistency between subjects, rather than using the within-subjects design employed by Ricketts et al. (2009). In their first study, with French-speaking monolingual children, Jubenville et al. observed orthographic facilitation: oral vocabulary learning was greater for the orthography present group, compared to the orthography absent group. Further, children who saw consistent orthographic forms showed greater orthographic facilitation than those seeing inconsistent orthographic forms. In Study 2, bilingual (French-English) children showed the opposite effect. Thus, Jubenville et al. provided evidence that consistency moderates the orthographic facilitation effect, though in a different way for different groups of children, while Ricketts et al. (2009) did not. These mixed findings motivate further work that specifies whether there is an interaction between orthography and consistency. If orthographic facilitation is greater for consistent words, this would indicate that the impact of orthography on the acquisition of semantic information is driven by the relationship between

orthography and phonology. However, there may be a more direct impact of orthography on semantics, as suggested by models of word reading that specify direct links between orthography and semantics (e.g., Harm & Seidenberg, 2004; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001).

Orthographic facilitation for vocabulary acquisition has practical implications, suggesting that practitioners, such as teachers and speech and language therapists, should emphasise orthography when teaching new words. Indeed, teachers do write words on the board whilst explaining their meanings, though this strategy could be adopted more universally (Ricketts et al., 2015). In research, the presence of orthography has usually been incidental, with a few exceptions (Chambre et al. 2017; 2020; Mengoni, Nash, & Hulme, 2013). In Mengoni et al.'s study, the presence of orthography was explicit, with all children alerted to the spelling patterns for all items. Chambre et al. (2017; 2020) have investigated whether directing attention to print moderates orthographic facilitation in beginner readers. In their 2017 study, they compared a group of children for whom the presence of orthography was incidental with a group whose attention was directed to print, finding no difference between groups. In the 2020 study, an incidental group was compared to a group who were asked to read the word aloud, thus also directing attention to it. In this study there was some evidence that reading the word aloud enhanced the orthographic facilitation effect.

We conducted two studies in which children learned phonological forms and meanings of 16 polysyllabic words. To test whether orthographic facilitation would occur, half of the words were taught with access to the orthographic form (orthography present condition) and the other half were taught without orthographic forms (orthography absent condition). In addition, we manipulated the instructions that children received: approximately half of the children were told that some words would appear with their written form (explicit

group); the remaining children did not receive these instructions (incidental group). Finally, we investigated the impact of spelling-sound consistency by including words that varied continuously on a measure of pronunciation variability (after Mousikou, Sadat, Lucas, & Rastle, 2017; see Method for more details). The quality of lexical representations was measured in two ways. A cuing hierarchy was used to elicit semantic knowledge from the phonological forms, providing a fine-grained measure of semantic learning. First, participants were asked to provide a definition. If their response was incorrect, they were given part of the definition (a cue) and asked to provide the rest. If their response was still incorrect, they were asked to select the definition from a choice of four. A spelling task indexed the extent of orthographic learning for each word. We sought to make the experimental paradigm as naturalistic as possible. Therefore, real words were taught, using an instruction and assessment approach adapted from standard educational and speech and language therapy practice. In Study 1, we measured knowledge of newly learned words at two intervals: one week and eight months after teaching. Longitudinal studies of word learning are rare and this is the first longitudinal investigation of orthographic facilitation. Study 2 extended the same experimental paradigm to a larger and more varied sample.

2. Study 1

Forty-one children aged 9-10 years completed the word learning task, followed by semantic and orthographic assessments one week after learning (Time 1), and eight months later (Time 2). Given the paucity of longitudinal data in word learning and orthographic facilitation research, we did not make predictions about the influence of time. We addressed three research questions:

1. Does the presence of orthography promote greater word learning? We predicted that children would demonstrate greater orthographic learning for words that they

had seen (orthography present condition) versus not seen (orthography absent condition). We anticipated that orthographic facilitation might also be observed for semantic learning (Colenbrander et al., 2019).

2. Will orthographic facilitation be greater when the presence of orthography is emphasised explicitly during teaching? We expected to observe an interaction between instructions and orthography, with the highest levels of learning when the orthography present condition was combined with explicit instructions. However, this prediction was tempered by one study showing that this was not the case in younger children (Chambre et al., 2017).

3. Does word consistency moderate the orthographic facilitation effect? For orthographic learning, we expected that the presence of orthography might be particularly beneficial for words with higher spelling-sound consistency, with learning highest when children saw and heard the word, and these codes provided overlapping information. For semantic learning, we reasoned that if the presence of orthography on semantic learning is driven by a beneficial effect of orthography on the learning of phonology (Ricketts et al., 2009; Rosenthal & Ehri, 2008), then orthographic facilitation will be greatest for word forms with more consistent spelling-sound mappings.

2.1. Method

2.1.1. Participants were 41 children from one socially-mixed school in the South-East of England (M_{age} = 9.95, *SD* = .53, 24 female). All spoke English as a first language, and none had any recognised special educational need. Table 1 summarises participant characteristics. The [omitted for blind review] Ethics Committee provided ethical approval for the study, as part of the [omitted for blind review] project. Follow-up data after eight

months were not available for three children; one child had left the school and two were absent on the day of testing.

2.1.2. Materials and procedure.

2.1.2.1. Background measures. Participants completed background measures in one or two sessions, each lasting approximately 45 minutes. Tasks were administered in a fixed order and according to manual instructions. Nonverbal reasoning was measured using the Matrix Reasoning subtest of the Wechsler Abbreviated Scale of Intelligence – Second Edition (WASI-II; Wechsler, 2011), a pattern completion task (split-half reliability: .87; test-retest reliability: .79). Word and nonword reading were assessed using the Sight Word Efficiency (SWE) and Phonemic Decoding Efficiency (PDE) subtests of the Test of Word Reading Efficiency – Second Edition (TOWRE-2; Wagner, Torgesen, & Rashotte, 2011) and the Castles and Coltheart Test 2 (CC2; Castles et al., 2009). For the TOWRE-2 subtests, reading efficiency is indexed by the number of words (SWE) or nonwords (PDE) read correctly in 45 seconds (test-retest reliability SWE: .91; PDE: .90). For the CC2, children were presented with a series of interleaved regular words, irregular words and nonwords (40 of each type) printed on individual cards, which they were asked to read aloud. Oral vocabulary knowledge was indexed by the Vocabulary subtest of the WASI-II (Wechsler, 2011) and the British Picture Vocabulary Scale – Third Edition (BPVS-3; Dunn, Dunn, & NFER, 2009). The WASI-II indexes expressive vocabulary by asking children to verbally define words (split-half reliability: .91; test-retest reliability: .90). The BPVS-3 is a receptive vocabulary measure for which children are asked to indicate which of four pictures represents the meaning of each word.

2.1.2.2. Experiment: design. Children were taught 16 novel words in a 2x2 factorial design. The presence of orthography (orthography absent vs. orthography present) was

manipulated within participants: for all children, eight of the words were taught with orthography present and eight with orthography absent. Instructions (incidental vs. explicit) were manipulated between participants such that children in the explicit condition were alerted to the presence of orthography whereas children in the incidental condition were not. Participants completing explicit and incidental conditions (n = 20 in explicit condition; 21 in incidental condition) were matched in pairs for vocabulary knowledge and word reading ability, and then matched as closely as possible for gender, age and nonverbal reasoning (Fs < 1 for vocabulary, word reading, age and nonverbal reasoning). Items were counterbalanced across instruction and orthography conditions, with all words appearing in both orthography conditions for approximately the same number of children within the explicit and incidental groups.

2.1.2.3. Experiment: word stimuli. Stimuli comprised 16 polysyllabic words, all of which were nouns (see Appendix). Fifty curriculum-relevant words were identified that were unlikely to be known by 12-13 year olds, and could be described as 'tier 2' words: words that are used by mature language users across a variety of domains, and that frequently occur in written texts (Beck, McKeown, & Kucan, 2002). An adult survey (N = 117) and subsequent adolescent survey (N = 42, 15 – 18 years, M_{age} = 16.79, SD = 0.78) were used to identify a set of words that participants were unlikely to know. The surveys were administered online using Bristol Online Survey (now Jisc Online Surveys), and participants were asked to select one of four options in response to each word (following Dale, 1965): a) I've never seen or heard this word before; b) I'm familiar with this word, but don't know what it means; c) I have an idea of what this word means; or d) I definitely know what this word means. Participants were additionally asked to provide the meaning of a word if they knew it. Adult participants were recruited via social media sites, and adolescent participants via colleagues and acquaintances with adolescent children.

Based on these responses, the original list of 50 words was ranked in order of words least well known by respondents. Two lists of eight words were then selected that could be matched for counterbalancing purposes. Words were matched exactly in pairs for number of morphemes (range = 1-2 morphemes) and syllables (range = 2-4 syllables) and the items in each pair were allocated to separate lists. Item lists were also matched closely (all *F*s < 1) for adolescent survey ratings, number of letters (range = 6-11 letters), number of phonemes (range = 4-10 phonemes) and our measure of spelling-sound consistency (see below). Only one word in each list started with a vowel and initial consonants appeared a maximum of once in each list to avoid confusion amongst words. Care was taken to make sure that word meanings were not overlapping.

Spelling-sound consistency relates to the frequency with which letters correspond to sounds and vice versa. Spelling-sound consistency has been conceptualised carefully for monosyllabic words (Kessler & Treiman, 2001) but there is no consensus on how to capture consistency for polysyllabic words. We indexed consistency at the whole word level using *H* (after Mousikou et al., 2017; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995). The stimuli were read aloud by 33 children (17 girls, M_{age}= 13.81 years, SD = .28) recruited from a single school in the South-East of England, none of whom participated in the experiment. The frequency of each alternative pronunciation was recorded, and consistency was then calculated using the formula Σ[-pi x log2(pi)], where pi is the proportion of participants giving a certain pronunciation (see Appendix for values). An H value of 0 would indicate a consistent item (all participants producing the same pronunciation), with values >0 indicating greater inconsistency (pronunciation variability) with increasing magnitude.

2.1.2.4. Experiment: procedure. The experimental procedure is summarised in Figure 1. A pre-test was conducted to establish participants' knowledge of the stimuli. Then, each

child was seen for three 45-minute sessions to complete training (Sessions 1 and 2) and post-tests (Session 3). Sessions were spaced one week apart to emulate the pace of topic-related vocabulary learning in school, and to allow for spaced teaching (Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012). This also enabled newly learned words to be consolidated during sleep (Henderson, Weighall, & Gaskell, 2013). The intended gap of seven days between sessions was achieved for most participants between Sessions 1 and 2 (56%; *M* = 7.37 days, *SD* = 1.09, range = 6-12 days) and Sessions 2 and 3 (71%; *M* = 7.00, *SD* = 0.55, range = 6-8). Post-tests were then re-administered approximately eight months later at Time 2 (*M* = 241.58 days from Session 3, *SD* = 6.10).

All instructions, stimuli and feedback were pre-recorded by a native speaker of English and presented to participants via the E-prime 2.0 programme (Schneider, Eschman, & Zuccolotto, 2012a, 2012b). Instructions, feedback and orthography (where relevant) were also presented in written form on the screen. E-prime was used to randomise order of presentation and record the accuracy of responses. Presenting information in this way also allowed us to ensure that the experiment was presented as intended, pronunciations were standard across children and exposures and all children had the same opportunity to learn. The second author conducted all experimental sessions with all children. Three children were excluded due to an administration error. They are not referred to in the participants section, nor are they included in any tables, figures or analyses.

---Figure 1 about here---

2.1.2.4.1. Pre-test. For each word, children were asked, *'do you know what [word] means?'* If they responded 'yes', they were prompted to give a definition. Participants were excluded if they demonstrated any knowledge of a word's meaning in their definition. Seven

participants were excluded, with individuals knowing *syncopation* (*n* = 6) and *cataclysm* (*n* = 1).

2.1.2.4.2. Training. Each training session comprised two blocks of trials: one phonological block, then one phonology-semantic block. For children in the explicit condition only, the prompt 'for some of the activities, you will see the word written on the screen. You might find this helpful' was given once prior to each semantic-phonological training block. The pre-test provided one exposure to each phonological form; training provided a further 24 exposures. Children were exposed to word definitions 10 times and, for words in the orthography present condition, to orthography four times.

The phonological training block familiarised children with the new phonological forms. In an initial set of trials participants heard and repeated each word once (e.g., *'repeat epigram'*). In the second set of trials they heard each word and then repeated it whilst simultaneously tapping out its syllables to draw attention to the phonological structure of the word (e.g., Lundberg, Frost, & Petersen, 1988; Yopp & Yopp, 2000). This allowed for four exposures to the phonological form per session (eight over training).

In the phonology-semantic block (see Figure 1), children completed five activities with each word, taking one word at a time: 1. repeat it (e.g., *'repeat epigram'*); 2. listen to the word with its three-word definition (e.g., *'listen carefully / you don't need to do anything / epigram is a witty remark'*); 3. listen to the word in sentence context (e.g., *'listen carefully / you don't need to do anything / Ed knew how to use a good epigram to keep his friends entertained over dinner'*); 4. repeat the word with its definition (e.g., *'repeat after me / epigram is a witty remark'*); and 5. repeat the word and definition again, substituting the middle word of the definition (an adjective) for a synonym (e.g., *'repeat it, but this time change the middle word to a different word that means the same thing / epigram is a witty*

remark'). All definitions followed a determiner-adjective-noun structure and included simple vocabulary to ensure understanding. Sentence contexts (15 – 16 words) included the target word and provided additional cues to meaning. All definitions and sentence contexts appear in the Appendix. Repetition trials were included to engage children in the task and the synonym substitution was included to encourage them to actively process the meaning of the word. For words trained in the orthography present condition, the orthographic form appeared during passive activities: 2) listen to the definition; 3) listen to the word in sentence context. The two phonology-semantic blocks allowed for 16 exposures to each phonological form, 10 exposures to the definition and, for orthography present items, four exposures to the orthographic form.

2.1.2.4.3. Semantic post-test. The semantic post-test assessed knowledge for the meanings of newly trained words. We took a dynamic assessment or cuing hierarchy approach (Hasson & Joffe, 2007), providing children with increasing support to capture partial knowledge and the incremental nature of acquiring such knowledge (Dale, 1965). Each word was taken one at a time and children were given the opportunity to demonstrate knowledge in three steps: definition, cued definition, recognition. In the definitions step, each child was asked, '*what does [word] mean?*' If they were able to provide the target definition or a close approximation, the next word was presented. If not, they were given a semantic cue, using a set format: '*it is a type of [noun]*. *Can you tell me what type?*' If the child provided the target adjective or a close synonym the next word was presented. Otherwise, the child was asked to select the correct definition from an array of four, comprising the target definition and three distractors.

For the recognition step, the distractors were identical to the target definition with the exception of the adjective, which was substituted with a plausible alternative (e.g., for

epigram, target definition: ‘a witty remark’; distractors: ‘an unfunny remark’, ‘a kind remark’, ‘an indignant remark’). Adjectives were not used more than once across target and distracter definitions, and distractor adjectives that were similar in meaning to the target were avoided. Where possible, one distractor adjective was opposite in meaning to the target adjective (i.e. ‘unfunny’ for ‘witty’). The four multiple-choice options for each word were presented on the screen in a grid format until a response was made. Position was randomised and participants heard each option once in order: top left, top right, bottom left, bottom right.

The semantic post-test score captured depth of semantic knowledge for the newly learned words. A score of three was allocated for a correct response in the definition task, two for a correct response in the cued definition task, one for a correct response in the recognition task, and zero if the item was not correctly defined or recognised. For this measure, the maximum score was 48 (24 per orthography condition). Reliability (Cronbach’s α) was calculated for a binary score (1 = definition or cued definition, otherwise 0), and was acceptable (α = .71).

2.1.2.4.4. Orthographic post-test. This post-test was included to ascertain the extent of orthographic knowledge after training. Children were asked to spell each word and their spelling productions were transcribed so that they could be scored. Responses were scored using a Levenshtein distance measure, using the *stringdist* library (van der Loo, 2019) in R (R Core Team, 2018). This score indexes the number of letter deletions, insertions and substitutions that distinguish between the target and the child’s response. For example, the response ‘epegram’ for target ‘epigram’ attracts a Levenshtein score of 1 (one substitution). Thus, this score gives credit for partially correct responses, as well as entirely correct responses. The maximum score is 0, with higher scores indicating less accurate responses. Reliability was calculated with accuracy scores; this was acceptable (Cronbach’s α = .81).

For the interested reader, accuracy data are also available online alongside Levenshtein distance scores (see OSF: https://osf.io/e5gzk/?view_only=a43914620dae4cc1b56bf3c15ef8d6c6).

2.2. Results

Analysis data and code are shared through an OSF repository accessible at: https://osf.io/e5gzk/?view_only=a43914620dae4cc1b56bf3c15ef8d6c6.

2.2.1. Participant characteristics. Table 1 summarises performance on background measures. Norm-referenced scores are reported for all measures except the Castles and Coltheart Test 2 (CC2) where raw scores are reported instead. Norm-referenced scores indicate age-appropriate performance in relation to nonverbal reasoning, word reading, nonword reading and oral vocabulary knowledge.

---Table 1 about here---

2.2.2. Approach to analysis. We used mixed-effects models to analyse data because this approach permits modelling of both participant- and item-level variability simultaneously, unlike more traditional approaches such as ANOVA. In this study, multiple participants responded to multiple items, meaning that both participants and items were sources of nonindependence in our data (i.e. responses from the same participant are likely to be correlated, as are responses to the same item). Compared to ANOVA, mixed-effects models offer a more flexible approach, and are better able to handle missing data without significant loss of statistical power (Baayen, Davidson, & Bates, 2008; Brauer & Curtin, 2018).

We fitted Bayesian mixed-effects models using the brms (Bayesian regression models using 'Stan') library (Bürkner, 2017, 2018) in R (R Core Team, 2018). We adopted Bayesian rather than frequentist methods for three reasons. First, Bayesian approaches are highly flexible, enabling us to model the sequential and categorical nature of the semantic post-test responses. Second, while it is recommended that mixed-effects models fully take into account random effects (i.e. a maximal effects structure; Barr, Levy, Scheepers & Tily, 2013), convergence issues are common (Meteyard & Davies, 2020). Bayesian models will typically converge to accurate values of effects estimates for any sample (Liddell & Kruschke, 2018). Third, as we discuss below, Bayesian analyses allowed us to combine data sets in Study 2 without risk of elevating Type 1 error (Kruschke, 2013).

More generally, Bayesian models are scientifically advantageous because they yield accurate representations of the posterior distribution. For each parameter (including fixed and random effects), Bayesian models generated a probability distribution representing the differing probabilities of each potential value of the coefficient for an effect. This means that we were able to report the most probable value of the estimate for an effect, given the posterior distribution, data and model assumptions. In tables summarising our models (Tables 2-3), we report each estimate, along with its 95% credibility intervals (lower and upper bound). The credibility interval indicates the range within which we can suppose that the "true value" of a parameter lies (see OSF: word-learning-supplementary_2020-09-30.pdf for a graphical illustration of this). In tables we also report the proportion of the distribution that sits either above or below 0, depending on the direction of the effect. That proportion indicates the probability of an effect in that direction. Where lower and upper bounds of the credibility interval cross zero, the direction and the magnitude of effects are estimated with less certainty. To allow for comparison, equivalent frequentist models with p values are

included in Supplementary Materials, though were subject to convergence issues (see OSF: word-learning-supplementary_2020-09-30.pdf for details).

In the semantic post-test, participants worked their way through three steps, only progressing from one step to the next step if they provided an incorrect response or no response. Given the sequential nature of this task, we analysed data using sequential ratio ordinal models (Bürkner & Vuorre, 2019). In sequential models, we account for variation in the probability that a response falls into one response category (out of *k* ordered categories), equal to the probability that it did not fall into one of the foregoing categories, given the linear sum of predictors. We estimate the k-1 thresholds and the coefficients of the predictors. Orthographic post-test performance was scored using a Levenshtein distance measure where 0 corresponds to an accurate response and higher scores indicate less accurate responses. Because, for any response, the distance corresponds to the frequency of edits made, and because there is no upper limit to the potential number of edits, this outcome variable can be treated as count data and analysed under the assumption that values stem from a Poisson probability distribution (Gelman & Hill, 2007). This approach allowed us to estimate the effects of potential influences on scores, whilst allowing that many responses may be partially correct to varying degrees.

For the semantic and orthographic models, we took a hypothesis-driven approach, estimating the fixed effects of time (Time 1 vs. Time 2), orthography (absent vs. present), instructions (incidental vs. explicit) and consistency (standardized H), as well as the interaction between orthography and instructions and the interaction between orthography and consistency. Different levels of the three binary fixed effects were sum coded, with orthography as -1 (absent) vs. +1 (present), instructions as -1 (incidental) vs. +1 (explicit), and time as -1 (Time 1) vs. +1 (Time 2). Consistency H, as a numeric predictor variable, was

standardized to z scores before entry to models as predictors. Models were specified to include maximal random effects (after Barr et al., 2013).

2.2.3. Semantic post-test. Table 2 summarises the semantic model (see OSF: word-learning-supplementary_2020-09-30.pdf for full model summaries). Figure 2 (top panel) illustrates marginal effects, with four panels showing how each fixed effect influenced the probability that children would produce a response scored 0, 1, 2 or 3. It is clear that there were very few full definitions (coded 3) or incorrect responses (coded 0), with the majority of responses either reflecting recognition of the definition (category 1) or cued definitions (category 2), and the fixed effects primarily influencing the relative contribution of category 1 and 2 responses to the total.

---Table 2 and Figure 2 about here---

Time was estimated to have a negative effect, with children producing lower scored responses at Time 2 than Time 1. At Time 2, there were fewer cued definition (category 2) responses and more recognition (category 1) responses, compared to Time 1. Importantly though, at Time 2, our estimates reveal good retention of knowledge about each word, as reflected in the high probability of recognition responses. There was some evidence that instructions influenced performance, with higher responses in the explicit than incidental condition. The credibility intervals for instructions, consistency and the interactions (see Table 2) show that the evidence was not sufficient to resolve the magnitude or the direction of these effects.

2.2.4. Orthographic post-test. Table 2 summarises the orthographic model, and Figure 2 (bottom panel) illustrates marginal effects, showing how each fixed effect influenced the accuracy of spelling responses. Note that 0 indicates a correct response and higher scores correspond to less accurate responses. Spelling productions were more accurate

for items taught in the orthography present condition, compared to the orthography absent condition. Other effects were estimated with less certainty.

2.3. Discussion

Phonological forms and meanings for sixteen polysyllabic words were taught, with half of the words taught with orthography present, and half without orthography. We measured learning for semantic and orthographic information just after teaching (Time 1), and eight months later (Time 2). We analysed our data using Bayesian mixed-effects models. In relation to our hypotheses, there was evidence for orthographic facilitation, with more accurate spelling responses for words that had been taught with orthographic support than those taught without. In comparison, the orthographic facilitation effect was estimated with less confidence for our semantic learning measure. Stronger effects of orthography on the learning of orthographic rather than semantic information are congruent with previous findings (Colenbrander et al., 2019). We did not observe the hypothesised interactions between orthography and instructions, or between orthography and consistency. An advantage of using Bayesian models was that they allowed us to estimate the magnitude of effects so that we can quantify confidence about our findings, instead of using the significant/nonsignificant dichotomy. There was uncertainty in the estimation of the orthographic facilitation effect for semantic learning, and little confidence in the hypothesised interactions for both orthographic and semantic learning. This uncertainty could reflect limited power or minimal individual differences, and Study 2 set out to explore this possibility. Further discussion of Study 1 findings is included following Study 2, in the General Discussion below.

3. Study 2

Study 1 provided evidence for orthographic facilitation, though the effect was estimated with more certainty for orthographic than semantic learning. Analyses did not support the hypothesised interactions between orthography and consistency, or between orthography and instructions. In Study 2, the Study 1 sample was combined with an older sample of children (total $N = 74$) in order to increase diversity within the sample, and provide more power for analyses. Increasing sample size and then re-running analyses does not increase the Type 1 error rate in Bayesian analyses in the way that it does for more traditional significance testing (Kruschke, 2013). The research questions and hypotheses were the same as for Study 1 except that longitudinal analyses were not possible for Study 2.

3.1. Method

3.1.1. Participants. Thirty-three children from an additional three socially mixed schools in the South-East of England were added to the Study 1 sample (total $N = 74$). These additional children were older ($M_{age} = 12.57$, $SD = .29$, 17 female) and their characteristics are summarised in Table 1. The same exclusionary criteria and ethics procedures were used.

3.1.2. Materials and procedure. These were identical to Study 1. For the background measures, one child from the additional older age group did not complete the TOWRE. For the experiment, there were now 37 participants completing each condition (explicit and incidental) and for most, there was a 7-day time difference between Sessions 1 and 2 (76%; $M = 7.20$ days, $SD = .83$, range = 6-12 days) and Sessions 2 and 3 (76%; $M = 7.43$, $SD = 1.81$, range = 6-17). Four children, including the three described for Study 1 were excluded due to an administration error. After the pre-test, a further 22 participants were excluded, including the seven described for Study 1, because they knew *dormancy* ($n = 11$), *syncopation* ($n = 8$), *accolade* ($n = 5$), *cataclysm* ($n = 4$), *nonentity* ($n = 2$) and *debacle* ($n = 1$). Excluded participants are not referred to in the participants section, nor are they included

in any tables, figures or analyses. Reliability for the semantic and orthographic post-tests were acceptable for this larger sample (semantic: Cronbach's α = .72; orthographic: Cronbach's α = .74).

3.2. Results and Discussion

3.2.1. Participant characteristics. Table 1 summarises performance on background measures for participants included in Studies 1 and 2. Again, norm-referenced scores (where available) show means and standard deviation scores that are in line with the test norms.

3.2.2. Semantic and orthographic post-tests. We analysed post-test data to test our hypotheses and establish whether the magnitude of our effects would increase with a larger and more varied sample. Models were identical to those used for Study 1 but without the effect of time, including fixed effects of orthography, instructions, consistency, orthography x instructions and orthography x consistency and a maximal random effects structure (see Table 3). Compared to Study 1, the effect of orthography on semantic learning was estimated with more certainty (*P* = .93 vs. .86), indicating a trend for higher quality semantic responses when orthography was present, rather than absent (for marginal effects plots, see top panel of Figure 3). The increased probability is also consistent with the notion that the presence of orthography influences semantic learning, but that this effect is small and our study was underpowered to detect it. Other effects were estimated with uncertainty, as for Study 1. Findings for the orthographic post-test also replicated Study 1 (for marginal effects, see bottom panel of Figure 3). There was clear evidence for more accurate spelling patterns when orthography was present rather than absent but other effects were not supported.

---Table 3 and Figure 3 about here---

4. General Discussion

Children were taught phonological forms and meanings for 16 unknown polysyllabic words. Half of the words were taught with orthographic forms available, and the remaining words were taught without orthographic forms. Fine-grained measures of semantic and orthographic learning were used to ascertain lexical quality for the newly learned words. In line with our predictions, we observed orthographic facilitation: children were more likely to learn words that they had seen during training. This effect was robust for orthographic learning but less clear for semantic learning. We did not find evidence for our hypothesised interactions: that orthographic facilitation would be moderated by consistency or the instructions that children received. Particularly novel was the longitudinal aspect of our study. Post-tests were administered one week after the end of teaching (Studies 1 and 2), and eight months later (Study 1 only), and analyses showed that over this time frame knowledge was well retained and orthographic facilitation effects endured.

4.1. Orthographic facilitation for word learning

The presence of orthography resulted in more accurate spelling responses and shifted the weighting of semantic responses towards deeper semantic knowledge. For orthographic learning, this effect was robust. For semantic learning, it was less clear, though it was estimated with high probability, especially in Study 2, where analyses were better powered. In a systematic review, Colenbrander et al. (2019) concluded that effects on orthographic learning are strong and consistent whereas effects on semantic learning can be nonsignificant (e.g., Chambre et al., 2017) or range from small to large (e.g., Rosenthal & Ehri, 2008; Ricketts et al., 2009). Colenbrander et al. concluded that the magnitude of the semantic learning effect could not readily be explained by differences in the teaching or assessment approach used in the studies. They called for further research. Indeed, many factors will determine whether an individual can learn a new word meaning, such as the learning context

(e.g., in the classroom, in conversation, while reading, background noise), word characteristics (e.g., whether the word has multiple meanings, its meaning is more concrete or abstract) and individual differences (e.g., pre-existing knowledge). It might be that in some cases the presence of orthography exerts only a small influence relative to these other forces. However, this effect may still be important. Consistently encountering orthography with phonology and semantics may lead to subtle changes in lexical quality that promote reading comprehension (e.g., Perfetti & Hart, 2002). Furthermore, presenting orthography whilst teaching is a strategy that many teachers already use, and it is low cost in terms of time and resources (Ricketts et al., 2015). Even a small effect on learning words on one or two occasions in the classroom can accumulate over the many encounters with words that occur during each hour, each day, each year, resulting in a large effect across words, learning opportunities and development.

We hypothesised that the presence of orthography might be more beneficial to learning if it was explicitly emphasised. However, telling participants that orthography would be present for some items did not influence orthographic facilitation. Therefore, it seems that when orthography was there, children attended to it, even when their attention was not explicitly directed to it (see also Chambre et al., 2017). It is worth noting that our instructions were not very directive and placing more emphasis on processing the orthographic form might influence orthographic processing (see Chambre et al., 2020).

4.2. The role of consistency in word learning and orthographic facilitation

In this study, we deliberately characterised the spelling-sound consistency of words to see if this would moderate the orthographic facilitation effect. In so doing, we aimed to test a key mechanistic account of orthographic facilitation: that the presence of orthography confers an advantage on word learning via its impact on phonology. We reasoned that if this is the

case, orthographic facilitation should be greater for more consistent items where orthography is a more reliable cue to phonology. However, our models did not support an interaction between orthography and consistency. Our findings indicate that the presence of orthography promoted orthographic learning, and to a lesser degree semantic learning, irrespective of item-level consistency. Notably, whilst our findings resonate with some previous studies (Jubenville et al., 2014, Study 1; Ricketts et al., 2009; Ricketts et al., 2015), others have indicated that consistency moderates orthographic facilitation (Jubenville et al., 2014, Study 2; Li et al., 2016; Rastle et al., 2011).

It may be premature to conclude that the impact of orthography on word learning is not moderated by consistency. We hypothesised that orthographic facilitation would be greater when orthography-phonology mappings are more consistent. However, the opposite could also occur. Inconsistency may render items more salient, with inconsistent items attracting more attention than consistent items and therefore driving greater orthographic facilitation. Preliminary evidence for this idea comes from a study showing that less 'wordlike' stimuli can be more readily learned than more 'wordlike' forms (Storkel, Armbruster & Hogan, 2006). Another possibility relates to the orthographic skeleton proposal (Wegener et al., 2017), which suggests that when children hear a novel word, some orthography is activated on the basis of what they know about spelling-sound mappings. With this in mind, orthographic learning for consistent items in the orthography absent condition could already be quite high, with little room for improvement. Therefore, the presence of orthography might be particularly beneficial for more inconsistent words with spelling patterns that would be harder to infer from phonology.

There are other more methodological reasons for remaining tentative about our consistency findings. First, the effect of orthography was limited for semantic learning. If this

reflected insufficient statistical power, this may also have constrained any interactions. Second, there was not much variation in our consistency measure across items (see Appendix), which may have limited its prediction of outcomes and associations with other variables. Third, since we chose multisyllabic words that were aligned with what children would be learning, capturing consistency was a challenge as there is no consensus for how this should be done for multisyllabic words. A fruitful area for future research would be to explore further the conditions under which consistency exerts an influence on word learning and orthographic facilitation (or not). Indeed, consistency is known to impact spelling performance (Caravolas, Kessler, Hulme, & Snowling, 2005) and some studies have shown that it moderates orthographic facilitation (Jubenville et al., 2014, Study 2; Li et al., 2016; Rastle et al., 2011). A study that included a greater number of words and therefore a greater range of consistency would be useful, as would further exploration of the appropriate way to capture consistency in multisyllabic words. In our study, we captured consistency from orthography to phonology (variation in pronunciation), though in English this is not the same as phonology-orthography consistency and the latter will be more important in underpinning spelling generation. Further, as in the consideration of monosyllabic consistency, it would be beneficial to consider more carefully the locus of inconsistency (e.g., vowel vs. consonant) and how consistency can be conditional on the context (Kessler & Treiman, 2001).

4.3. Lexical learning over time

In Study 1, children completed post-tests one week after teaching ended, and eight months later. Tracking learning of specific words over more than a few days or weeks is extremely unusual (for an exception, see Gellert & Elbro, 2013) and our findings are quite striking: our paradigm supported lexical quality that was well maintained over time. Orthographic knowledge did not degrade with time and semantic knowledge was well

retained, despite no intervening teaching. It is possible that children were exposed to these words in the interim. However, our pilot data (see Method) showed that older adolescents knew little about these words, indicating that this is unlikely. As a cautionary measure, teachers were not given the list of words until after data collection was complete. Notably, semantic responses indicated deeper knowledge of meaning one week after learning, compared to eight months later. Nonetheless, at both time points children exhibited semantic knowledge about many words that was durable and at least sufficient to support recognition of the correct definition. This level of knowledge may well support a range of language processing tasks. For example, even minimal semantic knowledge of *debacle*, when combined with other knowledge and skills, could allow for the successful comprehension of a text that includes this word.

4.4. The importance of using fine-grained outcome measures

Our measures of learning were novel in going beyond simple accuracy to capture knowledge in a more fine-grained manner. For orthographic learning, we administered a spelling task, which is widely argued to be a precise measure of orthographic representations (cf. Andrews, Veldre, & Clarke, 2020). Instead of analysing binary accuracy as usual, we gave credit for partially correct responses, indexing the distance between spelling responses and targets. The semantic post-test followed a 'cuing hierarchy' (Hasson & Joffe, 2007) or 'dynamic assessment' approach to provide progressively greater support for performance and to adequately capture depth of the knowledge learned.

These measures allowed us to look below the 'tip of the iceberg' and capture the partial knowledge that may lie beneath a simple correct or incorrect classification. Lexical learning must be incremental, and our measures capture that. By taking this approach, we were able to observe the way that time, and to some extent orthography, changed the

contribution of correct recognitions and cued definitions to responses. If we had measured simple accuracy, this would have obscured these effects. In addition, had we used definition accuracy as our outcome, we would have concluded that our paradigm did not teach semantic information as there were very few correct definition responses. Indeed, our learning task was challenging. Though we provided more teaching than is usual, we taught 16 complex forms with richer meanings than are typically presented in the field (for a review, see Colenbrander et al, 2019). By measuring partial knowledge, it was clear that our paradigm was sufficient to support substantial learning: either cued definition or recognition responses made up 80% of responses. This sensitivity in measurement recommends our approach to future research and brings it closer to practice. It is important to know how close children are to knowing word forms and meanings, not just whether they know them or not.

4.5. Strengths and limitations

In order to maximise the relevance of this study to practice, we drew heavily on educational and speech and language therapy expertise, discussing our methods with school teachers, and speech and language therapists. We adopted an unusually naturalistic approach, teaching real words over multiple sessions and carefully selecting words that were just beyond the reach of our participants. This was balanced with idealised learning conditions, where teaching was one-to-one and distractions were minimised. As discussed above, our outcome measures were sensitive to the incremental nature of learning. Our approach was also evidence informed. We aligned our teaching and assessment approach with memory and learning research that highlights the importance of spacing (Carpenter et al., 2012) and sleep-related consolidation (e.g., Henderson et al., 2013).

One clear limitation of our study is sample size, an issue that plagues learning and longitudinal research as such research is costly and resource intensive. Given that the effect

of orthography might be small for semantic learning, or in the real world where learning takes place amongst distractions, larger studies are particularly warranted. As discussed above, our measure of consistency would benefit from further consideration. Finally, for our measure of semantic learning, we provided the phonological form and requested information about meaning. Given the link between orthography and phonology, it may be that orthographic facilitation is greater for tasks that require phonological output. There is some evidence for this (Miles, Ehri, & Lauterbach, 2016; Ricketts et al., 2015, though see Colenbrander et al., 2019) and a large body of evidence supports orthographic facilitation for phonological form learning (e.g., Ehri & Wilce, 1979; Reitsma, 1983). We did not measure phonological learning separately but rather sought to 'pre-train' phonological forms so that we could focus on the learning of semantics, and phonology-semantic mappings. Had we measured semantic learning using tasks that require production of the phonological form, or measured phonological learning separately, we would likely have observed stronger orthographic facilitation effects.

4.6. Conclusion

In conclusion, the presence of orthography promoted higher quality lexical representations, particularly in terms of orthographic learning. We did not find evidence that the presence of orthography was more beneficial when it was made explicit, suggesting that the effect of orthography was somewhat automatic. Consistency did not influence orthographic facilitation either and further empirical work is needed to specify how orthography exerts its influence on vocabulary acquisition. Our study provides novel evidence that relatively short learning paradigms can lead to lexical knowledge that is well retained over an extended time frame. In addition, it highlights the importance of using measures of learning that probe the incremental nature of word knowledge, instead of crude

accuracy measures that might mask learning. Future studies that capture the incremental nature of word learning will not only inform theory, but will also resonate with vocabulary teaching practice, where even small changes in knowledge may be important for boosting spoken and written language processing.

5. References

Andrews, S., Veldre, A., & Clarke, I. (2020). Measuring lexical quality: The role of spelling ability. *Behavior Research Methods*. doi:10.3758/s13428-020-01387-3

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. doi:10.1016/j.jml.2012.11.001

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412. doi:10.1016/j.jml.2007.12.005

Beck, I., McKeown, M., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York: Guildford Press.

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychol Methods, 23*(3), 389-411. doi:10.1037/met0000159

Bürkner, P.C. (2017). An R package for Bayesian Multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1-28. doi:10.18637/jss.v080.i01

Bürkner, P.C. (2018). Advanced Bayesian Multilevel Modeling with the R package brms. *The R Journal*, 10(1), 395-411. doi:10.32614/RJ-2018-017

Bürkner, P.C., & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A tutorial. *Advances in Method and Practices in Psychological Science*, 2(1), 77-101. Doi: 10.1177/2515245918823199

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: review of recent research and implications for instruction. *Educational Psychology Review, 24*(3), 369-378. doi:10.1007/s10648-012-9205-z

Caravolas, M., Kessler, B., Hulme, C., & Snowling, M. (2005). Effects of orthographic consistency, frequency, and letter knowledge on children's vowel spelling development. *Journal of Experimental Child Psychology, 92*(4), 307-321.

Castles, A., Coltheart, M., Larsen, L., Jones, P., Saunders, S., & McArthur, G. (2009). Assessing the basic components of reading: A revision of the Castles and Coltheart test with new norms. *Australian Journal of Learning Difficulties, 14*(1), 67-88.

Chambre, S. J., Ehri, L. C., & Ness, M. (2017). Orthographic facilitation of first graders' vocabulary learning: does directing attention to print enhance the effect? *Reading and Writing*, 1-20. doi:10.1007/s11145-016-9715-z

Chambre, S. J., Ehri, L. C., & Ness, M. (2020). Phonological decoding enhances orthographic facilitation of vocabulary learning in first graders. *Reading and Writing*, *33*(5), 1133-1162. doi:10.1007/s11145-019-09997-w

Colenbrander, D., Miles, K. P., & Ricketts, J. (2019). To See or Not to See: How Does Seeing Spellings Support Vocabulary Learning? *Language, Speech, and Hearing Services in Schools, 50*(4), 609-628. doi:10.1044/2019_LSHSS-VOIA-18-0135

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

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English, 42*(8), 895-948.

Dunn, L. M., Dunn, D. M., & NFER. (2009). *British Picture Vocabulary Scale - Third Edition* (3rd ed.). London: GL Assessment Ltd.

Ehri, L. C. (2020). The science of learning to read words: A case for systematic phonics instruction. *Reading Research Quarterly, 55*(S1), S45-S60. doi:https://doi.org/10.1002/rrq.334

Ehri, L. C. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading, 18*(1), 5-21. doi:10.1080/10888438.2013.819356

Ehri, L. C., & Wilce, L. (1979). The mnemonic value of orthography among beginning readers. *Journal of Educational Psychology, 71*, 26-40. doi:10.1037/0022-0663.71.1.26

Gellert, A. S., & Elbro, C. (2013). Do experimental measures of word learning predict vocabulary development over time? A study of children from grade 3 to 4. *Learning and Individual Differences, 26*(0), 1-8. doi:http://dx.doi.org/10.1016/j.lindif.2013.04.006

Gelman, A., & Hill, J. (2007). Data analysis using Regression d Multilevel/Hierarchal Models. Cambridge University Press.

Harm, M., & 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. doi:10.1037/0033-295X.111.3.662

Hasson, N., & Joffe, V. (2007). The case for Dynamic Assessment in speech and language therapy. *Child Language Teaching and Therapy, 23*, 9-25.

Henderson, L., Weighall, A., & Gaskell, G. (2013). Learning new vocabulary during childhood: Effects of semantic training on lexical consolidation and integration. *Journal of Experimental Child Psychology, 116*(3), 572-592. doi:https://doi.org/10.1016/j.jecp.2013.07.004

Jubenville, K., Sénéchal, M., & Malette, M. (2014). The moderating effect of orthographic consistency on oral vocabulary learning in monolingual and bilingual children. *Journal of Experimental Child Psychology, 126*(0), 245-263. doi:10.1016/j.jecp.2014.05.002

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

Krepel, A., de Bree, E. H., & de Jong, P. F. (2020). Does the availability of orthography support L2 word learning? *Reading and Writing.* doi:10.1007/s11145-020-10078-6

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*(2), 573-603. doi:10.1037/a0029146

Li, H., Zhang, J., Ehri, L., Chen, Y., Ruan, X., & Dong, Q. (2016). The role of orthography in oral vocabulary learning in Chinese children. *Reading and Writing*, 1-19. doi:10.1007/s11145-016-9641-0

Liddell, T.M., & Kruschke, J.K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimetnal Social Psychology,* 79, 328-348. https://doi.org/10.1016/j.jesp.2018.08.009

Lundberg, I., Frost, J., & Petersen, O.-P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly, 23*(3), 263-284.

Mengoni, S. E., Nash, H., & Hulme, C. (2013). The benefit of orthographic support for oral vocabulary learning in children with Down syndrome. *Journal of Child Language, 40*(Special Issue 01), 221-243. doi:10.1017/S0305000912000396

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 104092. doi:https://doi.org/10.1016/j.jml.2020.104092

Miles, K. P., Ehri, L. C., & Lauterbach, M. D. (2016). Mnemonic value of orthography for vocabulary learning in monolinguals and language minority English-speaking college students. *Journal of College Reading and Learning, 46*(2), 99-112. doi:10.1080/10790195.2015.1125818

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. doi:http://dx.doi.org/10.1016/j.jml.2016.09.003

Nelson, J. R., Balass, M., & Perfetti, C. A. (2005). Differences between written and spoken input in learning new words. *Written Language & Literacy, 8*(2), 25-44.

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (Vol. 11). Amsterdam/Philadelphia: John Benjamins Publishing Company.

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rastle, K., McCormick, S. F., Bayliss, L., & Davis, C. J. (2011). Orthography influences the perception and production of speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1588-1594. doi:10.1037/a0024833

Reitsma, P. (1983). Printed word learning in beginning readers. *Journal of Experimental Child Psychology, 36*(2), 321-339. doi:10.1016/0022-0965(83)90036-X

Ricketts, J., Bishop, D. V. M., & Nation, K. (2009). Orthographic facilitation in oral vocabulary acquisition. *Quarterly Journal of Experimental Psychology, 62*(10), 1948-1966. doi:10.1080/17470210802696104

Ricketts, J., Dockrell, J. E., Patel, N., Charman, T., & Lindsay, G. (2015). Do children with specific language impairment and autism spectrum disorders benefit from the presence of orthography when learning new spoken words? *Journal of Experimental Child Psychology, 134*(0), 43-61. doi:http://dx.doi.org/10.1016/j.jecp.2015.01.015

Rosenthal, J., & Ehri, L. C. (2008). The mnemonic value of orthography for vocabulary learning. *Journal of Educational Psychology, 100*(1), 175-191. doi:10.1037/0022-0663.100.1.175

Schneider, W., Eschman, A., & Zuccolotto, A. (2012a). *E-Prime 2 reference guide*. Pittsburgh: Psychology Software Tools Inc.

Schneider, W., Eschman, A., & Zuccolotto, A. (2012b). *E-Prime 2 user's guide*. Pittsburgh: Psychology Software Tools Inc.

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, *94*(2), 143-174. doi:10.1348/000712603321661859

Share, D. L. (2008). On the Anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin, 134*(4), 584-615. doi:10.1037/0033-2909.134.4.584

Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research, 49*(6), 1175-1192. doi:10.1044/1092-4388(2006/085)

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. doi:10.1037/0096-3445.124.2.107

Valentini, A., Ricketts, J., Pye, R. E., & Houston-Price, C. (2018). Listening while reading promotes word learning from stories. *Journal of Experimental Child Psychology, 167*, 10-31. doi:https://doi.org/10.1016/j.jecp.2017.09.022

Van der Loo, M.J. (2014). The stringdist package for approximate string matching. *The R Journal*, 6(1), 111-122.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2011). *Test of Word Reading Efficiency - Second Edition (TOWRE-2).* Austin, TX: Pro-Ed.

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence - Second Edition (WASI-II).* San Antonio, TX Pearson.

Wegener, S., Wang, H.-C., de Lissa, P., Robidoux, S., Nation, K., & Castles, A. (2017). Children reading spoken words: Interactions between vocabulary and orthographic expectancy. *Developmental Science*.

Yopp, H. K., & Yopp, R. H. (2000). Supporting phonemic awareness development in the classroom. *The Reading Teacher, 54*(2), 130-143.

Tables and Figures

Table 1. Performance on background measures

Measure	Study 1 (*N* = *41*)			Study 2 additional group (*N* = 33)		
	M	*SD*	range	*M*	*SD*	range
WASI-II nonverbal reasoning[1]	47.88	9.48	27-73	48.09	9.23	30-71
TOWRE-2 word reading[2]	103.83	12.51	79-129	104.00	11.75	83-132
TOWRE-2 nonword reading[2]	106.29	12.58	83-131	103.50	12.85	73-129
CC2 regular word reading[3]	36.49	2.59	28-40	37.48	2.69	29-40
CC2 irregular word reading[3]	24.76	4.00	18-35	26.03	3.40	17-32
CC2 nonword reading[3]	31.93	5.68	14-39	32.33	6.88	13-40
WASI-II expressive vocabulary[1]	52.95	7.12	35-69	51.42	8.84	31-66
BPVS-3 receptive vocabulary[2]	93.59	11.84	72-120	93.88	12.99	69-120

Notes. WASI-II = Wechsler Abbreviated Scale of Intelligence – Second Edition; TOWRE-2 = Test of Word Reading Efficiency – Second Edition; CC2 = Castles and Coltheart Test 2; BPVS-3 = British Picture Vocabulary Scale – Third Edition; [1]T-score (*M* = 50, *SD* = 10); [2]Standard score (*M* = 100, *SD* = 15); [3]Maximum score = 40

Table 2. Model summaries for Study 1 (semantic and orthographic post-tests)

Post-test	Effect	Estimate	Est.Error	95% credibility interval		Proportion of distribution above or below 0
				Lower bound	Upper bound	
Semantic	Intercept[1]	-1.89	0.25	-2.39	-1.40	
	Intercept[2]	1.65	0.25	1.16	2.14	
	Intercept[3]	2.73	0.30	2.15	3.32	
	Time	-0.92	0.08	-1.08	-0.76	1.00
	Orthography	0.08	0.08	-0.07	0.23	0.86
	Instructions	0.30	0.14	0.03	0.57	0.98
	Consistency	0.31	0.21	-0.10	0.72	0.94
	Orthography:Instructions	0.07	0.07	-0.06	0.20	0.86
	Orthography:Consistency	-0.04	0.06	-0.16	0.09	0.73
Orthographic	Intercept	0.19	0.15	-0.11	0.47	
	Time	-0.03	0.03	-0.09	0.04	0.80
	Orthography	-0.14	0.03	-0.20	-0.08	1.00
	Instructions	0.00	0.07	-0.15	0.15	0.50
	Consistency	0.13	0.12	-0.11	0.37	0.86
	Orthography:Instructions	-0.04	0.03	-0.09	0.02	0.91
	Orthography:Consistency	0.04	0.03	-0.01	0.10	0.94

Note. Model for both semantic and orthographic post-test outcomes:
score ~ time + orthography + instructions + consistency + orthography:instructions + orthography:consistency + [random effects associated with]
(time + orthography + consistency + 1 | participant) + (time + orthography + instructions + 1 | word)

Table 3. Confirmatory model summaries for Study 2 (semantic and orthographic post-tests)

Semantics

Post-test	Effect	Estimate	Est.Error	95% credibility interval		Proportion of distribution above or below 0
				Lower bound	Upper bound	
Semantic	Intercept[1]	-3.66	0.34	-4.34	-3.02	
	Intercept[2]	0.67	0.30	0.08	1.27	
	Intercept[3]	2.53	0.33	1.89	3.20	
	Orthography	0.11	0.08	-0.03	0.26	0.93
	Instructions	0.30	0.13	0.05	0.56	0.99
	Consistency	0.39	0.27	-0.15	0.92	0.93
	Orthography:Instructions	0.02	0.06	-0.10	0.14	0.64
	Orthography:Consistency	0.00	0.07	-0.14	0.14	0.51
Orthographic	Intercept[3]	0.11	0.16	-0.21	0.42	
	Orthography	-0.17	0.04	-0.24	-0.10	1.00
	Instructions	0.00	0.06	-0.11	0.12	0.54
	Consistency	0.16	0.15	-0.13	0.46	0.86
	Orthography:Instructions	-0.01	0.03	-0.07	0.05	0.68
	Orthography:Consistency	-0.01	0.03	-0.07	0.05	0.64

Note. Model for both semantic and orthographic post-test outcomes:
score ~ orthography + instructions + consistency + orthography:instructions + orthography:consistency + [random effects associated with]
(orthography + consistency + 1 | participant) + (orthography + instructions + 1 | word)

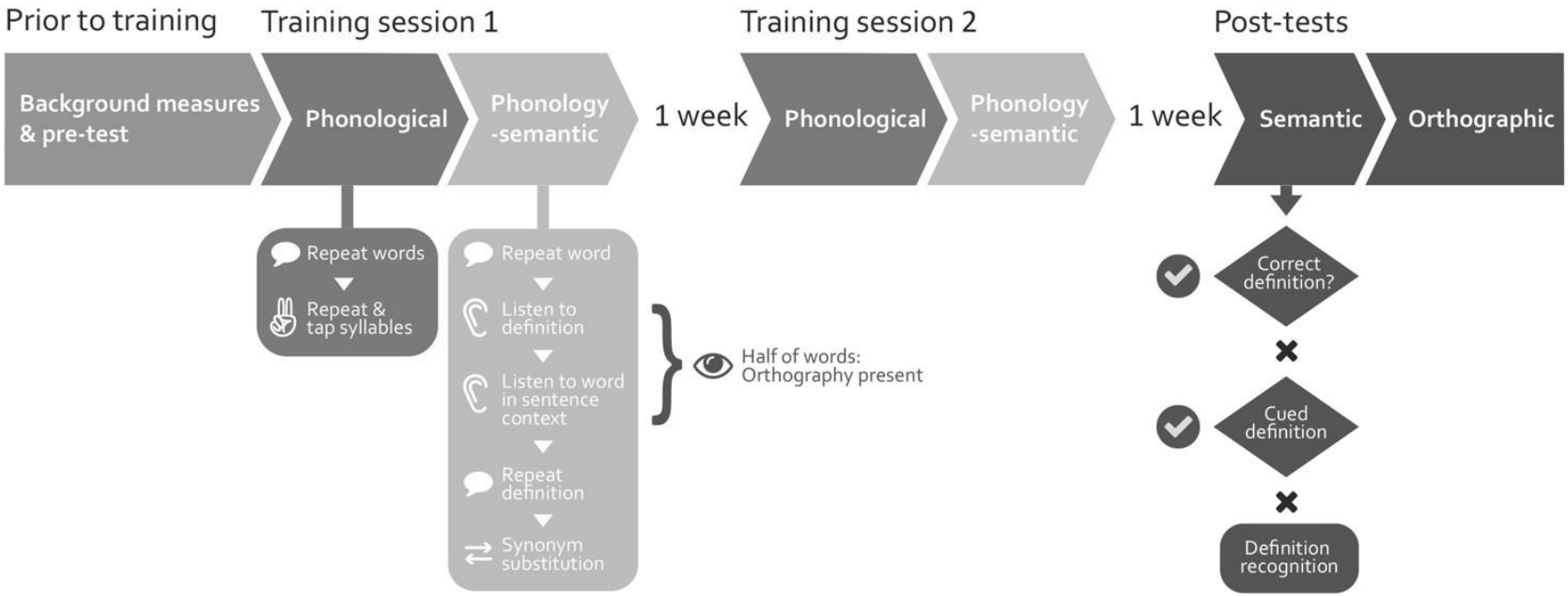

Figure 1. Experimental procedure

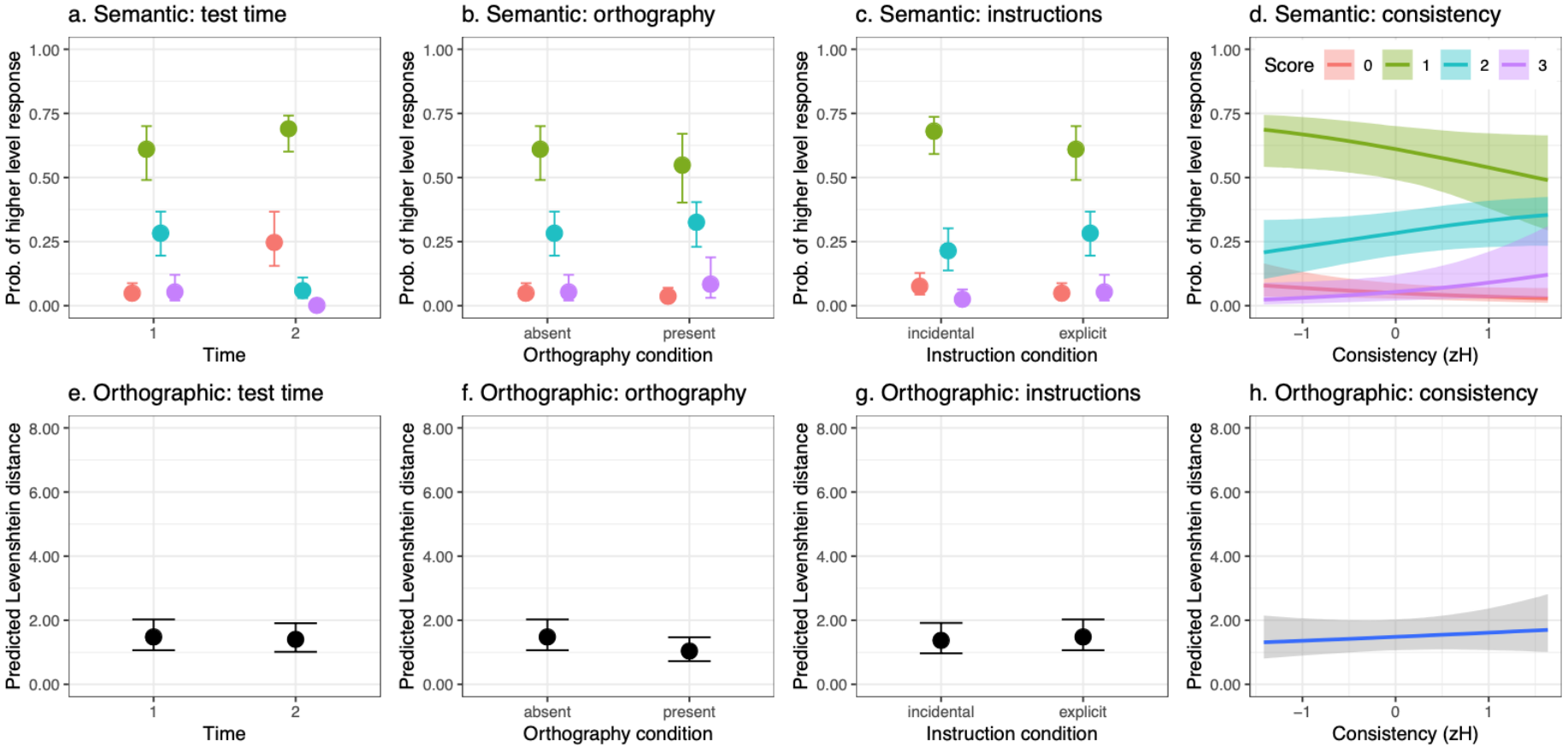

Figure 2. Marginal effects plots for Study 1. The top panel shows semantic post-test main effects (3 = full definition, 2 = partial definition, 1 = definition recognised, 0 = incorrect/no response) and the bottom panel shows orthographic post-test main effects (0 = accurate, otherwise higher scores correspond to less accurate responses).

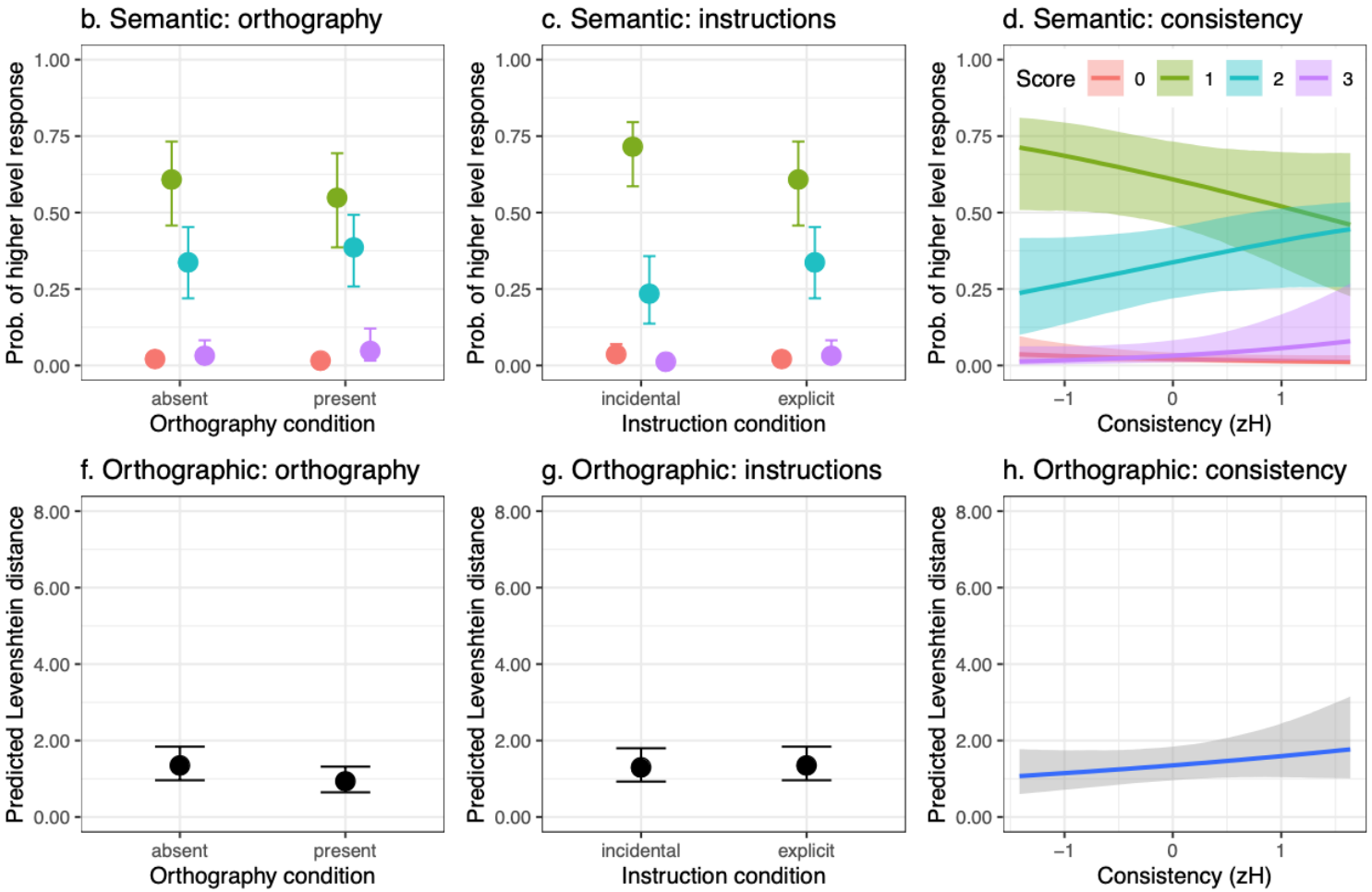

Figure 3. Marginal effects plots for Study 2. The top panel shows semantic post-test main effects (3 = full definition, 2 = partial definition, 1 = definition recognised, 0 = incorrect/no response) and the bottom panel shows orthographic post-test main effects (0 = accurate, higher responses correspond to less accurate responses).

6. Appendix. Stimuli, definitions, sentence contexts and values for H consistency

Stimuli	*Definitions*	*Sentence contexts*	*H*
accolade	a strong compliment	Based on their exam results, the school deserved the accolade of being the best in England.	1.91
cataclysm	a violent event	The village didn't survive the cataclysm of war, but it was re-built after the war ended.	3.51
contrition	a sorry feeling	Ted felt bad for upsetting his parents, and was full of contrition as he apologised.	1.75
debacle	a sudden failure	The England team apologised to their unhappy fans following the debacle of the World Cup.	2.90
dormancy	a sleepy state	During the winter, earwigs go through a long period of dormancy until the warm weather returns.	1.63
epigram	a witty remark	Ed knew how to use a good epigram to keep his friends entertained over dinner.	1.38

Stimuli	*Definitions*	*Sentence contexts*	*H*
foible	a personality weakness	Eve's only foible was that she tended to ignore problems and hope they would go away.	2.71
fracas	a noisy argument	What started out as a small disagreement ended up as an embarrassing fracas at the park.	3.14
lassitude	a tired mood	George was overcome with lassitude, and didn't feel like doing anything other than staying in bed.	0.90
luminary	an inspirational person	Simon was a luminary scientist, and was influential in encouraging people to follow his lead.	1.10
nonentity	an unimportant character	The man was a complete nonentity to Sue; she had never heard of him before.	3.97
platitude	a meaningless comment	Sally begged the new politician to be honest, and not to utter yet another platitude.	0.90

Stimuli	*Definitions*	*Sentence contexts*	*H*
propensity	a predictable behaviour	He had a propensity to lunge into tackles, and as a result received many yellow cards.	1.69
raconteur	a good storyteller	The children sat round and listened eagerly as the raconteur brought the story to life.	3.82
syncopation	a musical pattern	The syncopation of the music made Ryan want to get up and dance to the rhythm.	3.04
veracity	a truthful situation	Lin doubted the veracity of the claim because it seemed too good to be true.	2.87