Exposure to dialect variation in an artificial language prior to literacy training impairs reading of words with competing variants but does not affect decoding skills

Glenn P. Williams
Nikolay Panayotov
Vera Kempe

DOI: https://doi.org/10.1037/xlm0001094

Exposure to dialect variation in an artificial language prior to literacy training impairs reading of words with competing variants but does not affect decoding skills

Glenn P. Williams[1,2], Nikolay Panayotov[1] & Vera Kempe[1]

[1] Abertay University

[2] University of Sunderland

Correspondence concerning this article should be addressed to Vera Kempe, Division of Psychology, School of Applied Sciences, Abertay University. Dundee. DD1 1HG. Scotland, UK. E-mail: v.kempe@abertay.ac.uk. Pre-registration, data and code, and pre-print are available at https://osf.io/7ct9x/.

Acknowledgements:

The authors gratefully acknowledge funding from The Leverhulme Trust (Grant #RPG_2016-039) to Vera Kempe.

Abstract

Many bidialectal children grow up speaking a variety (e.g. a regional dialect) that differs from the variety in which they subsequently acquire literacy. Previous computational simulations and artificial literacy learning experiments with adults demonstrated lower accuracy in reading *contrastive* words for which dialect variants exist compared to *non-contrastive* words without dialect variants. At the same time, exposure to multiple varieties did not affect learners' ability to phonologically decode untrained words; in fact, longer literacy training resulted in a benefit from dialect exposure as competing variants in the input may have increased reliance on grapheme-phoneme conversion. However, these previous experiments interleaved word learning and reading/spelling training, yet children typically acquire substantial oral language knowledge prior to literacy training. Here we used artificial literacy learning with adults to examine whether the previous findings replicate in an ecologically more valid procedure where word learning precedes literacy training. We also manipulated training conditions to explore interventions thought to be beneficial for literacy acquisition, such as providing explicit social cues for variety use and literacy training in both varieties. Our findings replicated the reduced accuracy for reading *contrastive* words in those learners who had successfully acquired the dialect variants prior to literacy training. This effect was exacerbated when literacy training also included dialect variation. Crucially, although no benefits from the interventions were found, dialect exposure did not affect reading and spelling of untrained words suggesting that phonological decoding skills can remain unaffected by the existence of multiple word form variants in a learner's lexicon.

Introduction

Most research on literacy acquisition has focused on investigating how humans learn to read and spell a linguistic variety that they are also using in oral communication. However, many learners the world over are exposed to multiple varieties, often in diglossic situations where a regional dialect spoken at home may be different from the variety that is considered the standard and is introduced in educational settings. In this study, we examine how learning to read and spell is affected when there is a mismatch between varieties (e.g., a dialect variety and a variety considered 'standard') encountered prior to and during literacy training. So far, research has documented poorer literacy outcomes when the language variety used at home did not match that taught in school, compared to situations where there was no such variety mismatch (Charity et al. 2004; Gatlin & Wanzek, 2015; Terry et al., 2016; Washington et al., 2018). However, in naturalistic educational settings, difficulties in literacy attainment may be caused by a host of confounding extra-linguistic factors known to impact educational outcomes, such as quality and quantity of input in the 'standard' variety, general educational provision, or teacher expectations and attitudes, which are only partly explained by a learners' socio-economic status (Goldenberg, Rueda & August, 2007; Van Steensel, 2006). Studying literacy acquisition in naturalistic educational settings makes it difficult to tease apart to what extent learning outcomes in situations of a variety mismatch between a dialect and a 'standard' are affected by these external factors, or whether they can be explained by how the specific input these learners encounter affects the cognitive mechanisms that underpin learning to read and spell. In fact, difficulties may arise because oral dialect use induces greater phonetic, phonological, lexical and morpho-

syntactic variability which can increase the mismatch between print and sound in the standard variety (Labov, 1995). For example, if speakers of Scots dialects typically pronounce ‘house’ as */hu:s/* but in school are encouraged to produce */haʊs/* when reading ‘house’ this introduces a greater mismatch between orthographic and phonological form compared to speakers who only ever use */haʊs/*. Such mismatch between the phonological form of dialect variants and the standardised spelling could potentially impair the acquisition of phonological decoding skills.

To control for external extra-linguistic confounds we used an experimental approach that employed artificial literacy learning. Such experimental scrutiny is paramount considering existing assumptions of a causal relationship between dialect exposure and poor literacy outcomes that seem to have motivated several recent attempts in some regions to restrict the use of dialects in school settings (BBC News, 2013a, b; BBC News, 2020). Furthermore, we aim to experimentally test different training regimens to provide insights that may help educators in diglossic regions to decide which interventions, if any, are best suited to support literacy acquisition in children expected to use multiple language varieties (for a discussion of such interventions see Rickford & Rickford [1995]; McWorther [1997] in the US-context and Costa [2015] in the Scottish context). Although most individuals acquire literacy as children, our study begins to tackle the underlying learning mechanisms by testing adults to disentangle effects of input variability from effects of cognitive immaturity. Testing adult learners provides proof of concept about how variety mismatch arising from dialect exposure as a feature of the input can affect literacy learning in a mature cognitive system. If we observe additional costs associated with the experimental condition of dialect exposure in adults who have already developed literacy skills in their native language this would

provide strong evidence that such costs may be even larger in children encountering a variety mismatch between dialect and 'standard' in experimental and naturalistic settings. If we do not observe additional costs from dialect exposure in adults future research will still have to scrutinise how dialect exposure may interact with a less mature cognitive system.

Some important insights into the mechanisms of literacy learning in situations of dialect exposure while controlling for extra-linguistic factors have been gained from connectionist modelling. Brown et al. (2015) used an attractor network to explore how exposure to African-American English (AAE) word form variants affects learning to read an orthography that represents Mainstream American English (MAE). The network was first trained on phonology-phonology mappings to simulate word learning before training it to map orthographic forms onto phonological forms to simulate the process of learning to read. Crucially, in the variety-mismatch simulations, half of the words in the word learning phase were presented as dialect variants displaying AAE phonological features such as consonant cluster reductions, consonant drops, substitutions, exchanges and devoicing. When the network was subsequently trained to read the MAE counterparts, the cross-entropy error was markedly higher for contrastive words, both compared to non-contrastive words within the same simulation as well as to the counterparts of contrastive words in variety-match simulations with exposure to MAE or AAE variants only. It is important to note that performance was better in both variety-match conditions, thus ruling out any *a priori* differences between the two varieties in learnability of sound-spelling associations. Rather, it was the lower accuracy associated with reading contrastive words that was the main source for impaired overall performance in the variety-mismatch network, implicating interference from competing

variants as the reason behind reduced literacy learning outcomes in situations where learners are exposed to different varieties at home and at school.

Reduced performance with contrastive words is reminiscent of reading difficulties observed for heterophonic homographs like *lead* or *wind* – words for which the context determines which pronunciation to select (Gottlob et al., 1999; Jared et al., 2012). As heterophonic homographs are characteristic for opaque orthographies where graphemes are pronounced differently depending on context, a higher cost of reading contrastive words, i.e. words with dialect variants is consistent with Labov's (1995) idea that dialect exposure might exacerbate any existing mismatch between spelling and sound. However, the connectionist simulations by Brown et al. (2015) left open the question as to whether an increased error rate for contrastive words would also arise if words were linked to semantic representations and if the orthography was entirely transparent. Crucially, it did not investigate whether exposure to different varieties would also lead to impaired decoding of unfamiliar words – a question that is at the heart of the debate about the role of phonological decoding skills in successful literacy acquisition (Castles, Rastle & Nation, 2018).

Subsequent studies using artificial literacy learning sought to clarify these issues. In Williams et al. (2020), we taught adult participants a miniature artificial language consisting of 30 words, interleaved with literacy training using an invented script. In the dialect condition, half of the words were learned as dialect variants instantiating similar phonological variation as in the Brown et al. (2015) simulation. The results of three experiments confirmed the existence of a higher cost of reading contrastive words regardless of whether we had introduced meanings by presenting pictures, whether the orthography was transparent or opaque, and whether participants had only learned to

read or also to spell. Notably, however, the ability to decode untrained words, in analogy to non-word reading tests in naturalistic literacy learning settings, was not adversely affected by exposure to these dialect variants. In fact, when the period of literacy training with an opaque orthography was extended, participants in the dialect condition showed superior reading performance overall, and particularly with untrained words, compared to participants without dialect exposure.

A dialect benefit in reading of untrained words is a counter-intuitive finding that may have arisen because the regimen of interleaved word exposure and literacy training in the artificial literacy learning experiment of Williams et al. (2020) created a situation in which participants received literacy training before stable word form representations had been established in long-term memory. However, such concurrent word and literacy acquisition is not typical for bidialectal children but is more common in adult second language learners where reading of the new variety (i.e. the new language) is recruited for word learning resulting in competition between word forms from the learners' native and the new, in this case artificial, language. In this situation, continuous exposure to variation in word forms arising from interleaved presentation of two variants in the dialect condition may have led learners to prioritise phonological decoding not just to reduce interference from native-language word forms, but also because for contrastive words, no other information was available to determine which one of the competing pronunciation variants to produce. Learners not exposed to different variants encountered less variability in the input and had to learn fewer different words which might have encouraged a strategy of trying to access phonological forms via semantic representations, cued either by the concurrently provided picture cues to meaning, or by decoding just the initial graphemes of the orthographic form. Prioritising access to

phonology via meaning as a strategy, which corresponds to the orthography-semantics-phonology pathway in Harm & Seidenberg's (2004) triangle model, leaves less room for practicing the grapheme-phoneme decoding skills that underpin the orthography-phonology pathway required for reading untrained words for which no phonological representations exist.

While the dialect benefit observed in the Williams et al (2020) study suggested that dialect exposure is unlikely to be a major limitation for the acquisition of phonological decoding skills, the findings were difficult to generalise because a training regimen that interleaves exposure to both varieties is not typical for literacy learning in a first language where phonological forms of most words that children subsequently learn to read and spell are already well entrenched. In fact, recent connectionist modelling suggests that oral language proficiency, i.e. how many words a learner knows and how well entrenched the sound-meaning associations are, shapes the process of literacy acquisition, especially with respect to the success of phonological decoding strategies for written word comprehension (Chang et al., 2020).

The first aim of the present study was therefore to determine whether lower accuracy in reading of contrastive words and a reading benefit for untrained words can still be found under ecologically more valid conditions resembling first language literacy acquisition when learners are trained to read and spell words they already know. While keeping the overall amount of input identical to Experiment 3 from Williams et al. (2020), we re-designed the training procedure so as to front-load exposure to the artificial words to allow learners to gain oral word knowledge, i.e. to establish links between phonology and semantics for the artificial words prior to the onset of literacy training. We compared a condition in which participants encountered the same variants

during word learning and literacy training (No Dialect condition) with a condition in which participants learned different variants for half of the words prior to literacy training (Dialect condition)[1]. Unlike in Williams et al. (2020), establishing links between phonological forms and word meanings prior to literacy learning should enable learners to access these forms more readily and should not require as much reliance on mapping orthography into phonology during reading. If learners acquire dialect word forms prior to literacy training subsequent reading of contrastive words should produce a greater proportion of dialect pronunciations which, when evaluated based on standard pronunciation, will be considered erroneous. This greater cost should arise because accessing the phonology of these words via their meaning would lead to competition between the dialect and the standard phonological form. Moreover, if front-loading word learning and the resulting oral word knowledge de-prioritises phonological decoding skills dialect exposure should no longer lead to a benefit for reading of untrained words compared to the No Dialect condition. In fact, if substantial oral word knowledge is gained prior to literacy training, dialect exposure might actually result in impaired reading of untrained words if, as suspected by Labov (1995), increased overall inconsistency of grapheme-phoneme mappings adversely impacts acquisition of decoding skills.

Word learning in laboratory language studies with adult participants is liable to considerable individual differences, including differences in working memory capacity and existing vocabulary size in the native language (e.g. Kempe et al., 2009; Kaushanskaya et al., 2013). Both of these factors may affect a learner's ability to

[1] We use the terms 'dialect' and 'standard' simply for ease of reference to highlight the ecological motivation behind this study, in analogy to how such varieties are often labelled in the sociolinguistic literature. This is in full acknowledgement that for many languages the 'standard' is a historically conventionalised version of one specific dialect that serves as the basis for the writing system.

establish strong sound-meaning associations for the artificial words. We therefore included a vocabulary test to see how well the artificial words had been learned prior to literacy training. This allowed us to conduct an exploratory analysis of how individual differences in word learning affect the accuracy of processing contrastive words and the acquisition of decoding skills in situations of dialect exposure. Learners who are more successful at learning the artificial words may be more likely to try to retrieve the phonological form via a word's meaning when encountering the orthographic from alongside meaning cues. When successful word learners are subsequently exposed to a different variety during literacy training, prior entrenchment of dialect variants should lead to greater reading difficulties with contrastive words compared to poorer word learners whose less stable word form representations are likely to cause less interference. Better word learning may also present difficulties for reading of untrained words as more entrenched reliance on access to phonology via semantics as the preferred strategy might discourage the more successful learners from practising phonological decoding skills. This may, paradoxically, result in a reduced benefit or even no dialect benefit for reading untrained words in stronger word learners compared to poorer word learners who may attempt sequential phonological decoding of the orthographic forms during reading training more often.

The second aim of this study was to experimentally test the effects of interventions inspired by educational practice designed to help learners to deal with the greater variability in word forms in situations of dialect use. The first intervention involved the introduction of cues that raise awareness of the existence of dialect variants and the situations in which each variety may be used (Brown et al., 2015). Such cues are social in nature linking dialect use to certain groups of speakers that are distinguished

either by social status or region of origin. As a proxy we combined a geographical cue about the locale of two groups of imaginary speakers with the indexical cue of speaker sex to identify a speaker as belonging to one or the other group. The second intervention involved alternating literacy training in both the dialect and the standard variety. This intervention was motivated by applied considerations as we sought to provide empirical evidence for the efficacy of educational practices that have tried to introduce dialect reading and spelling into the classroom to improve children's literacy (Rickford & Rickford, 1995; McWorther, 1997; Costa, 2015, Education Scotland, 2015; 2017). Testing these two interventions under controlled conditions should provide educators with further evidence that can aid them in appraising their efficacy. Below we detail the potential mechanisms by which these two interventions might be beneficial for literacy learning in situations of dialect exposure.

Effects of social information: In their efforts to computationally simulate effects of variety mismatch on learning to read, Brown et al. (2015) designed a second connectionist simulation which implemented explicit dialect coding using context units that were activated depending on whether a specific variant belonged to one or the other variety. Although this network had to learn a larger set of word forms as it was exposed to both variants of contrastive words, overall reading performance was greatly improved compared to the network that switched from learning words in one variety to literacy training in the other without such context units. In fact, when context units started to become activated early in literacy training, overall reading performance in the variety of literacy training was very similar to a network trained on just this variety. Such a network architecture that explicitly cues or tags lexical representations as belonging to one or another variety is reminiscent of bilingual language representation models (e.g.

Green, 1998) which assume that selective activation of a language schema can prevent interference from the non-target language. In a similar vein, the simulation results reported in Brown et al. (2015) suggest that context-dependent tagging of competing variants can alleviate the reduced reading accuracy associated with dialect exposure. These simulation results are also in line with evidence on the role of dialect awareness in bidialectal literacy acquisition (Terry & Scarborough, 2011; Johnson et al., 2017; Yiakoumetti, 2006). For example, Johnson et al. (2017) demonstrated that an explicit dialect awareness intervention administered to bidialectal children resulted in stronger gains in morphosyntactic awareness, reading comprehension, and more frequent use of mainstream forms in narrative writing. Although this evidence suggests that bidialectal learners benefit from explicit knowledge about when it is appropriate to use each variety, it remains unclear whether contextual cues of dialect use increase the accuracy of reading contrastive words because neither for the connectionist simulation nor for the intervention study was performance for such words specifically reported. If contextual information pre-activates a target variety then interference from competing non-target variants may be reduced. We therefore aimed to explore whether explicit social cues increase the accuracy of reading contrastive words, and whether such cues are also beneficial for the ability to decode novel words. To this end, we created a third exposure condition termed Dialect & Social where we linked indexical cues of different speakers (i.e., male and female voice) with a geographical cue designed to help learners to associate different speakers with regional differences in variety use. These cues were introduced to simulate the contextual knowledge that speakers draw on when switching between different varieties in their everyday language use, although we acknowledge that in reality such knowledge is likely to be far more complex.

Effects of dialect literacy training: As indicated above, we introduced the artificial literacy learning methodology to explore an applied question related to the efficacy of explicit dialect literacy teaching. While in some diglossic regions the introduction of dialect literacy as an attempt to improve literacy outcomes in bidialectal children has proven controversial it is actively encouraged by educational authorities in others. Yet the empirical evidence for the outcomes of dialect literacy is scarce. In the United States, AAE dialect literacy has shown some success in a few small-scale studies (Leaverton, 1973; Simpkins & Simpkins, 1981; Rickford & Rickford, 1995) but is difficult to evaluate on its own as it is often introduced in the context of Contrastive Analysis, an approach designed to raise the learners' awareness of contrasting features in both varieties (Rickford et al., 2004) or other approaches that encourage wider dialect use in the classroom beyond literacy learning. In Scotland, on the other hand, the use of Scots dialect in the classroom alongside Standard Scottish English is being encouraged as a means to advance general literacy skills (Education Scotland, 2015; 2017). Because dialect spelling, e.g. in Scots, is often not standardised (Costa, 2015) it is conceivable that having learners read and spell when links between print and sound are flexible may foster phonological awareness in similar ways to invented spelling, i.e. children's self-directed, non-conventionalised and often spontaneous attempts to render words in print (Ouellette & Sénéchal, 2008; 2017). We therefore included a fourth exposure condition termed Dialect Literacy where in addition to the explicit contextual cues for dialect use, literacy training took place in both varieties in equal parts. Using literacy training in both varieties we tested whether increased orthographic variability increases accuracy of reading contrastive words and phonological decoding skills required for reading untrained words.

The present study. The four conditions described above were introduced in an artificial literacy learning study in which participants were first exposed to novel words and novel graphemes before receiving training in both reading and spelling. We included spelling training to maintain ecological validity as children tend to learn both skills from the onset of schooling and because spelling, which requires systematic conversion of phonemes into graphemes, has been shown to benefit reading skills (Caravolas, Hulme & Snowling, 2001). All novel words were presented with a depiction of their meaning to simulate the availability of semantic information, not only during word learning but also during reading, in acknowledgement of the fact that context, be it pictorial or linguistic, provides cues to the meaning of a word. This was intended to simulate teaching literacy through use of picture books and illustrations during literacy learning, but also to maintain methodological consistency with Williams et al. (2020) where pictures were presented alongside words to make a difficult learning task easier. Participants in the No Dialect condition were presented with the same standard words during exposure and literacy training. In the three dialect conditions, half of the words in the exposure phase were phonologically modified 'dialect' variants of the words subsequently presented during literacy training. As in many natural dialects our artificial dialect words were orthographically less consistent than the artificial standard words due to changes like consonant drops at the end. In the Dialect & Social and Dialect Literacy conditions, indexical features of the speaker voice (male vs. female) were linked with explicit information about the speaker's geographical origin to create awareness for the social and situational context of variety use. The use of a male or female voice is not intended to imply that men and women speak different dialects but was simply used as a salient indexical feature that would allow participants to associate

a specific speaker with a specific region. In the Dialect Literacy condition, literacy training was conducted in equal parts in both varieties; the target variety was cued by speaker voice and information about speaker origin.

The study received ethical approval (code EMS-1237). We pre-registered our hypotheses (available at https://osf.io/bxt87) with respect to the accuracy of processing contrastive and untrained words. We hypothesised that, as in Williams et al. (2020), the accuracy of reading contrastive words should be lower compared to non-contrastive words in the Dialect condition, i.e. when learners encountered a different variety prior to literacy learning, due to competition between the two variants when readers attempt to access phonology via semantics. We also predicted that the accuracy of reading contrastive words would be higher in the Dialect & Social condition compared to the Dialect condition as the presence of social and contextual cues for dialect use allows the target variants to be pre-activated. We originally pre-registered no accuracy difference between reading contrastive and non-contrastive words for the Dialect Literacy condition in analogy to a cognate benefit in bilingual processing because increased awareness of the different varieties could allow learners to develop separate schemas for them. However, in bilingual representations, cognates are more similar than the rest of the lexicon whereas in bidialectal representations contrastive words are less so thereby removing their relative advantage. This consideration casts doubt on our original prediction for the Dialect Literacy condition and suggests that predicting accuracy differences between contrastive and non-contrastive words in this condition is not straightforward.

For untrained words we predicted that reading performance would be improved in the Dialect condition compared to the No Dialect condition if, as in Williams et al.

(2020), the increased variability that arises from dialect variants in the input leads learners to prioritise decoding via grapheme-phoneme conversion as opposed to accessing phonology via meaning. We also predicted that reading performance with untrained words would be further improved in the Dialect & Social condition relative to the Dialect condition as the social cue should draw learners' attention to this increased variability thereby prioritising decoding via grapheme-phoneme conversion. Finally, we predicted that reading performance with untrained words should be best in the Dialect Literacy condition. Not only would participants be aware of the different varieties from the contextual cue but, in analogy to non-normative, invented spelling practices, they would be more likely to adopt a more analytical stance toward letter–sound correspondences (Caravolas et al., 2001; Ehri & Wilce, 2006; Ouellette & Sénéchal, 2008, 2017; Ouellette et al., 2008) as they get more opportunities to practice a greater set of phoneme-grapheme conversion rules.

Even though the main emphasis of this study was on how dialect exposure affects learning to read, spelling is an integral component of literacy training and was included into our procedure for reasons of ecological validity. We therefore did not pre-register specific hypotheses about participants' spelling performance and provide exploratory analyses to gain a more complete picture of all aspects of literacy learning. We suspect, however, that no difference in spelling accuracy between contrastive and non-contrastive words should emerge in any of the exposure conditions because spelling at this early stage of learning relies on sequential conversion of phonemes to graphemes (Seymour, 1997), and establishing stable orthographic representations for this entirely novel artificial script is likely to require substantially more time than provided in this experiment.

Method

Participants

Three hundred and twenty participants were recruited from the crowdsourcing website Prolific Academic and assigned to the four conditions in a pseudo-random way to achieve equal recruitment in each condition (see Procedure). Having English as a first language and self-report of no known cognitive impairments or dementia were participation conditions administered by the crowdsourcing website but in addition we also asked participants to rate their proficiency in English and any other languages[2] on a 1 (elementary proficiency) - 5 (native or fully bilingual proficiency) Likert scale. Self-reported age, sex and English proficiency for participants in each condition are provided in Table 1. Participants were reimbursed £9. An additional 24 participants were tested but not included because they gave the same response on all trials, their responses on most trials repeated the previous trial, their responses were in English rather than in the artificial language or were inaudible, because a technical difficulty had occurred (e.g. losing trials due to poor internet connection), or when recruitment inadvertently extended beyond our pre-registered cut-offs for a given list.

[2] A count of the additional languages known by participants in each condition is provided in the Supplemental Material.

Table 1: Means, standard deviations and range of age, sex and self-reported English proficiency of participants by condition.

	No Dialect	Dialect	Dialect & Social	Dialect Literacy
Age[3] (years)	32.30 (10.99) 18 - 64	31.01 (9.01) 19 - 75	32.33 (11.80) 18 – 63	33.68 (12.86) 18 - 67
Number of women	38	45	54	48
Number of men	40	33	26	32
Number of 'other'	2	2	0	0
English proficiency (1-5)	4.74 (0.73)	4.94 (0.29)	4.84 (0.68)	4.91 (0.36)
Number of multilinguals	32	32	34	48
Additional language proficiency (1-5)	2.07 (1.34)	2.44 (1.40)	2.02 (1.25)	2.16 (1.15)

Materials

Phonemes, graphemes, words and images: We created an artificial language consisting of 14 phonemes and graphemes. The phoneme inventory comprised the vowels [ɑ], [ɛ], [i], [ɔ], [u] and the consonants [m], [n], [s], [k], [b], [d], [f], [l], and [x]. Each grapheme (see Appendix A) was composed of two to four curved or straight strokes as common to most alphabetic writing systems (Changizi & Shimojo, 2005). Using all phonemes except [x], which was reserved for dialect variants, we constructed 42 artificial words distributed across six syllabic templates (3 monosyllabic, 3 bisyllabic) with no more than one consonant cluster per word and no cluster with more than two consonants, to constrain complexity and adhere to English phonotactics (Crystal, 2003; Harley, 2006). Thirty words were presented during exposure and literacy training, while twelve words (two words from each syllable template) were retained for testing only (henceforth: untrained words). The exact method of word construction and all words are provided in

[3] Three participants' self-reported ages of 16 years (No Dialect), 14 years (Dialect & Social), and 2 years (Dialect Literacy) have been removed from this summary under the assumption that these are likely to be typos, primarily due to Prolific Academic's restrictions on participants being 18 years of age or older.

Appendix B. Words were randomly paired with images depicting objects from six categories (body parts; furniture and kitchen utensils; household objects, tools, and instruments; food and clothing; building features and vehicles; animals and plants) from the revised Snodgrass and Vanderwart image set of colourised images provided by Rossion and Pourtois (2004) as detailed in Appendix C. The pictures of the beach and the mountain valley used as a social-geographical cue were taken from the freely available database of images on unsplash.com.

Dialect: To create a dialect version for our vocabulary of 30 artificial training words we created phonological variants for half of the words using the most frequent phonological changes found in another set of prominent British dialects, namely, Scots dialects, to ensure ecological validity. The determination of frequency and type of variation was based on the analysis of a Scots dialect corpus consisting of translations of Standard British English children's books (*The Gruffalo, The Gruffalo's Child*: see Appendix D). The variants for the 15 contrastive words (see Appendix B) were created applying the following changes that were modelled after changes in Scots compared to Standard British English as identified in the analysis of the Gruffalo-corpus: (a) consonant substitutions (e.g. /skub/ changed to /sxub/), (b) consonant drops (e.g. /snid/ changed to /sni/) and (c) vowel changes (e.g. /nɛf/ changed to /nif/) and /nɑl/ changed to /nɔl/). In instances where multiple changes could apply all were implemented in the dialect so that, for example, /skɛfi/ became /sxifi/ and /flɛsɔd/ becomes /flisɔ/. As in Scots dialects, the [x] only appeared in the dialect variants so that its spelling in the Dialect Literacy condition required its own grapheme.

Orthography: We used the same opaque spelling system as in Williams et al. (2020) to simulate the learning of an orthography with inconsistent grapheme-phoneme mappings like English. To disrupt consistent one-to-one mappings between phonemes and graphemes we created a roughly equal amount of feed-forward (spelling-to-sound) and feed-back (sound-to-spelling) inconsistency using two conditional rules: First, the phoneme /l/ was spelled using one grapheme in all instances except if preceded by /b/ or /s/, when it was spelled by another grapheme. Second, the phoneme /s/ was spelled using one grapheme in all instances except when it was preceded by /n/, when it was spelled by another grapheme. As a result, the artificial grapheme for *F* was pronounced as /s/ 27% of times and as /f/ 73% of times, and the artificial grapheme for *N* was pronounced as /n/ 67% of times and as /l/ 33% of times. Conversely, the phoneme /s/ was spelled as (the artificial equivalent of) the letter *S* 74% of times and as *F* 26% of times, and the phoneme /l/ was spelled as *L* 75% of times and as *N* 25% of times. The conditional spelling rules were matched across word types resulting in five contrastive and five non-contrastive words with inconsistent spelling. For one contrastive and one non-contrastive word, both conditional spelling rules applied simultaneously so that /slɔku/ and /slinab/ were spelled as *FNOKU* and *FNINAB*, respectively.

Procedure

Once participants had been admitted to the experiment on the crowdsourcing website Prolific Academic they were instructed that they would learn to read and spell a novel made-up language. In the Dialect & Social and Dialect Literacy conditions, the following explanation was added to alert participants to the contextual cue: '*You will notice that people from different areas speak and sometimes even write the language in*

slightly different ways - you probably know that this happens in many countries like Britain and the USA, for example. In this case, the people from the coast speak slightly differently than the people inland. You will see a little picture, like the one below, that lets you know where a speaker is coming from'.

Participants were told that it was important to perform the task in a quiet environment and to not take any notes for the entire duration. In order to ensure compliance we misled participants into believing that detection of cheating on our part could jeopardise their reward. Next, participants were asked to check the working order of their microphone and headphones/speakers and to rate their English proficiency and any additional known languages on a scale from 1 to 5.

Participants were assigned to one of the four conditions in the following way: Combining condition with literacy task order (reading first vs. spelling first) and speaker voice (male first vs. female first) we created sixteen different procedure sequence variants. We then tested twenty participants in each of these variants, presenting the next, randomly chosen, variant whenever the recruitment target for the preceding variant was fulfilled. All sixteen variants were administered over a period of 76 days. We opted for this method of pseudo-randomisation to prevent over-recruiting in any of the conditions given the cost of participant reimbursement. Repeated participation by the same participants was blocked by the Prolific Academic website. The mean completion time was 110 mins (*SD* = 36.90, range = 38.85 – 300.52). The source code for the experiment can be found at https://osf.io/5mtdj/.

The experiment consisted of four phases: exposure, vocabulary test, literacy training and literacy testing; the last two phases comprised a reading and a spelling task

the order of which was randomised. The phases and associated tasks and stimulus features for each condition are presented in Table 2.

Table 2: Experimental phases and associated stimulus components in the four conditions.

Phase	No Dialect		Dialect		Dialect & Social		Dialect Literacy	
initial instruction	word learning only		word learning only		word learning + cue alert		word learning + cue alert	
exposure	M	F	M	F	M	F	M	F
30 words (3 x)	St	St	Dia	Dia	Dia	Dia	Dia	Dia
vocabulary test 30 words								
training instruction	literacy only		literacy only		literacy + cue alert		literacy + cue alert	
letter training	M	F	M	F	F	M	F	M
R + S training 1	M	F	M	F	F	M	F	M
30 words	St	St	St	St	St	St	St	St
R + S training 2	M	F	M	F	F	M	M	F
30 words	St	St	St	St	St	St	Dia	Dia
R + S testing	M	F	M	F	F	M	F	M
42 words	St	St	St	St	St	St	St	St

Note. (R – reading, S – spelling, M - male voice, F – female voice; St – standard variety, Dia – dialect variety)

In the exposure phase, participants were presented with a picture and heard the phonological form of the word produced either by the female or the male voice. In the Dialect & Social and the Dialect Literacy conditions, the speaker origin (and hence variety) was cued by the picture of either the beach or the mountain valley in the top right corner of the screen so that speaker voice (male vs. female) was always associated with the same geographical cue. Association of speaker voice to geographical cues was counterbalanced across participants. On each trial, participants had to repeat each word once before hearing it again. Participants cycled through the set of thirty training words in random order. This phase was repeated two more times for a total of three exposure trials per word. The exposure phase was followed by a vocabulary test during which

participants saw each picture again and were asked to produce the corresponding word. Their responses were audio-recorded.

In the subsequent letter training phase, participants viewed each grapheme one by one, accompanied by the sound of the isolated phoneme produced by the voice that also produced the words during the testing phase. Each trial started with a fixation point and a prompt to press a button, which revealed the grapheme accompanied by its sound. To discourage participants from taking notes of the graphemes we included time limits (Rodd, 2019) so that each grapheme disappeared after 1,000ms. A second button press triggered a repetition of the grapheme and the sound presentation. The sequence of all graphemes was presented twice in random order, exposing participants to each grapheme-phoneme combination for a total of four times. Crucially, phonemes were randomly assigned to graphemes for each participant to eliminate the impact of any potential differences in how easily graphemes could be associated with certain phonemes.

Next, participants received two blocks of literacy training. Each literacy training block was further divided into a reading and a spelling sub-block. In each sub-block, participants learned to either read or spell the entire set of 30 training words. The order of the reading and spelling sub-blocks was counterbalanced so that half of the participants first completed the reading sub-block first, while the other half first completed the spelling sub-block. Sub-block order was the same in both literacy training blocks for a total of four presentation of each word, twice in reading and twice in spelling training.

In the No Dialect, Dialect and Dialect & Social conditions, participants were simply asked to learn to read the words out loud when presented with the spelling and

the picture, or to spell the words they heard by clicking the fourteen on-screen grapheme keys in the right order when presented with the sound and the picture. In the Dialect Literacy condition, participants additionally received the following instruction: *Just be aware that you may be taught by speakers from different regions*. *You will see the picture to let you know where a speaker is coming from. Please always use the version of the language from the region you see on the picture*. Dialect literacy training always followed standard literacy training (see Table 2). On each reading training trial, participants saw a string of graphemes accompanied by the picture and had to read the target word out loud. To avoid recording long silences participants' responses were timed by presenting a moving hand in a clock indicating the onset, duration and offset of the 2500ms recording window. After participants had recorded their response, they heard the correct word again. On each spelling training trial, participants heard a word accompanied by the corresponding picture, and had to type it by clicking graphemes in the right order using the on-screen keyboard. Once participants had pressed the on-screen 'Enter' key, feedback was provided in the form of the correct spelling, which appeared below their own spelling. The feedback screen was cleared after 1.5 to 3.0 sec to prevent participants from taking notes or obtaining screenshots (the exact presentation time of the feedback was determined dynamically based on the word length, with a duration of 500ms per letter so that, for example, the correct spelling feedback for a 4-letter-word would be presented for 2 sec).

Finally, participants were tested in reading and spelling of the 30 trained and 12 untrained words presented in random order. Test items were always presented by the same speaker, and with the same regional cue as used in the first literacy training block of this condition. For example, for participants in the Dialect & Social and Dialect

Literacy conditions who had been exposed to words presented by a female speaker originating from the coast but had received literacy training in another (the 'standard') variety by a male speaker originating from the valley in the mountains, test items were also presented by the male speaker and accompanied by the picture of the mountains. Participants were instructed to use the variety indicated by speaker voice and picture of speaker origin, which was always the same speaker that had presented the first 'standard' literacy training block, and not the speaker they had heard during dialect word exposure (see Table 2).

Results

All analyses were conducted in R (Version 3.6.3; R Core Team 2020) using the R-packages *bayestestR* (Version 0.5.2; Makowski et al., 2019b), *brms* (Version 2.12.0; Bürkner 2017, 2018), *english* (Version 1.2.5; Fox et al. 2020), *ggforce* (Version 0.3.1; Pedersen 2019), *ggrepel* (Version 0.8.2; Slowikowski 2020), *ggridges* (Version 0.5.2; Wilke 2020), *here* (Version 0.1; Müller 2017), *irr* (Version 0.84.1; Gamer et al. 2019), *kableExtra* (Version 1.1.0; Zhu 2019), *modelr* (Version 0.1.6; Wickham 2020), *papaja* (Version 0.1.0.9942; Aust & Barth 2020), *rlang* (Version 0.4.6; Henry & Wickham 2020), *tidybayes* (Version 2.0.1; Kay 2020), and *tidyverse* (Version 1.3.0; Wickham et al. 2019) for data preparation, analysis, and presentation.

Coding

Responses from the vocabulary test (30 items) and reading responses from the testing phase (42 items) were transcribed and coded by two coders (GPW and VK) blind to each participant's condition. The coding convention, which was based on the CPSAMPA (Marian et al., 2012) simplified notation of IPA characters is described in

detail in Williams et al. (2020). For all coded oral responses as well as for all spellings, length-normalised Levenshtein edit distances (nLEDs) to the target string were computed and used as the dependent variable to assess performance. Such edit distances are computed by dividing the number of insertions, substitutions, and deletions required to transform one string (e.g. a participant's input) into another (e.g. the target word) by the larger of the two string lengths (Levenshtein, 1966). Edit distances constitute a more gradual and fine-grained performance measure than error rates as they can distinguish near-matches from entirely erroneous productions. When literacy training in the Dialect Literacy condition targeted the 'dialect' variety, 'dialect' variants were adopted as targets for computation of nLEDs.

Inter-coder reliability was computed by obtaining intra-class correlations between the two coders' nLEDs for the reading task during the testing phase, using the *irr* R-package (Gamer et al., 2019). We used a single-score, absolute agreement, two-way random effects model based on the summed nLEDs for each participant. The ICC was 0.996, 95% CI = [0.985; 0.998], F(319.000, 8.053) = 752.805, $p < .001$. The 95% confidence interval around the parameter estimate indicates that the ICC falls above the bound of .90, which suggests excellent reliability across coders (Koo & Li, 2016). Whenever there was a discrepancy between the coders further analyses were based on the smaller of the two nLEDs thereby adopting a lenient coding criterion justified by the rationale that a participant response should be regarded acceptable if at least one of the coders can match it to the target as closely as possible.

Five participants had 1 missing trial from the total of 324 trials per participant. Recordings from a further six trials out of a total of 9120 trials in the Dialect & Social condition were invalid due to technical errors. Two of these trials pertained to the

vocabulary test with the remaining four pertaining to the literacy test (corresponding to 0.08% and 0.06% of trials, respectively). Trials that were either inaudible to both coders or had no response were coded as incorrect responses (i.e. with an nLED of 1). This affected an average of 9.80% of trials in the vocabulary test and 0.92% of trials in the literacy test, with the proportion of these errors being relatively equal in all conditions.

Modelling using Bayesian Zero-One-Inflated Beta Distributions

Our dependent variable, nLED, is bounded between 0 and 1, with inflated counts at these bounds, and, unlike proportions, does not arise from a series of independent and identically distributed observations. It is therefore more appropriate to model such data using zero-one-inflated beta (ZOIB) distributions. In contrast to general linear models and linear mixed effects models, which assume a Gaussian data generation process, ZOIB models do not make predictions outside the possible range of values and capture the larger densities at extreme values more accurately. Compared to model fitting that assumes only one underlying distribution, these models account for the multitude of ways in which nLEDs can be generated. For example, with perfect recollection from memory nLEDs are likely to be 0 while with varying levels of decoding ability they are likely to be distributed between values greater than 0 and lower than 1. This is captured by modelling the data as a Beta distribution for nLEDs excluding 0 and 1, and a Bernoulli distribution for nLEDs of 0 and 1. ZOIB models yield four paramters: μ (mu), the mean of the nLEDs excluding 0 and 1 with larger values being associated with higher mean nLEDs in the range excluding 0 and 1; ϕ (phi), the precision (i.e. spread) of the nLEDs excluding 0 and 1 with larger values being associated with tighter distributions of the nLEDs in the range excluding 0 and 1 (i.e. less variance); α (alpha; termed ZOI –

i.e. zero-one inflation – in *brms*) the probability of an nLED of 0 or 1 with larger values being associated with more zero-one inflation in the nLEDs; and γ (gamma; termed COI – i.e. conditional-one inflation – in *brms*), the conditional probability of a 1 given that a 0 or 1 has been observed, with larger values being associated with more one-inflation given zero-inflation in the nLEDs. Predictors in such a model can affect any and all of these four distributional parameters at once. A logit link is used for the μ, α, and γ distributional parameters, and a log link is used for the ϕ distributional parameter.

At the time of writing, distributional models of this nature can only be fitted under a Bayesian framework. We used the *brms* R-package to perform model fitting using this approach. As an additional benefit, Bayesian models do not suffer from the non-convergence problems often associated with fitting complex models using a frequentist framework. Given that these models return estimates for the four distributional parameters – the values of which dependent on one another – drawing inferences from direct inspection of parameter estimates is extremely difficult (if impossible for such complex models). Using Bayesian methods, inferences can be made based on samples from the joint posterior under the conditions of interest. This not only allows for inferences to be made based on simple summaries of the data on the nLED scale, but also allows for uncertainty surrounding all terms to be captured by these summaries.

Model Specification and Analysis. In the interest of brevity, we focus on the most interesting effects, testing for effects of dialect exposure conditions on: (i) reading and spelling performance for contrastive and non-contrastive words, and (ii) reading and spelling performance for untrained words. We additionally preregistered an analysis

testing for any effect of dialect exposure condition on reading and spelling performance across all words. However, this analysis largely replicates the effects shown for the subset of untrained words, which is our primary area of interest for testing implications for decoding ability. For completeness, this latter analysis is reported in the Supplemental Material.

In all models, estimates of population-level (fixed) and group-level (random) effects are computed for all distributional parameters, with group-level effects correlated across all parameters. We fitted three models:

a) The *vocabulary test model* examined vocabulary test performance after word learning and prior to literacy training; nLEDs are predicted by population-level (fixed) effects of Exposure with the four levels of No Dialect, Dialect, Dialect & Social, and Dialect Literacy, Word Type with the two levels of Contrastive and Non-Contrastive words, and the interaction between them, and by group-level (random) effects of random intercepts and slopes of Word Type by participants, and random intercepts and slopes of Exposure by item.

b) The *literacy test model* examined reading and spelling performance during the testing phase after literacy training; nLEDs are predicted by population-level (fixed) effects of Task with the two levels of Reading and Spelling, Exposure and Word Type and the interaction between them, and by group-level (random) effects of random intercepts and slopes of Task and Word Type by participant, and random intercepts and slopes of Exposure by item. In deviation from the pre-registration we had to exclude the by-participant random slopes of the interaction between Task and Word Type in order to keep run time manageable.

c) The *Exploratory Covariate Literacy Test Model* also examined reading and spelling performance in the testing phase but included vocabulary test performance as a further predictor. This model included population-level (fixed) effects of vocabulary test performance, defined as mean nLED by participants, Task, Exposure condition, Word Type, and the interaction between them, and group-level (random) effects of random intercepts and slopes of Task and Word Type by participant, and random intercepts and slopes of vocabulary test performance and Exposure by items. Again, the random slopes by participant did not include the interaction between Task and Word Type, and the random slopes by item did not include the interaction between vocabulary test performance and Exposure in order to reduce model complexity and run time. This model was not pre-registered, but instead serves an exploratory purpose to determine whether or not any effect of dialect exposure is dependent on how well the artificial words are learned.

In the interest of brevity, we do not provide the data for the training phase because in Williams et al. (2020) we observed that other than showing continuous improvement of performance over time the training models did not differ in the main effects from the testing models. This decision was taken in part also due to the substantial amount of time required for transcription of the training reading data. However, the spelling data obtained during the training phase are available at https://osf.io/7ct9x/.

In all models, we used weakly informative, regularising priors. Where divergences were detected during model fitting, these priors were adjusted, typically by placing less prior weight on extreme values. Largely, the priors were selected to allow

the posterior to be determined primarily by the likelihood and to ensure that estimates are more conservative than in typical frequentist analyses, thus allowing better out-of-sample prediction. Full details of the priors and posterior predictive checks are provided in Appendix E. Model summaries for the population-level (fixed) parameter estimates for all fitted models, including details of the coding scheme and how to interpret parameter estimates, can be found in Appendix F.

Fitting complex models such as those involving multilevel data or many distributional assumptions – as we do so here –often produces a posterior that is not multivariate-Normal. Such posterior distributions can often only be estimated using Markov-chain Monte Carlo techniques. This involves taking samples from the posterior to generate implied observations under different conditions. In doing so, more likely parameter values are sampled more often than less likely values. Thus, with an appropriate number of samples we can accurately describe the most likely parameter values and generate estimates of uncertainty around these (assuming all other modelling assumptions are met). To address our questions about how dialect exposure condition affects performance with contrastive words and decoding ability with untrained words, and to generate plots for these effects, we thus present and summarise samples from the posterior for target effects using the *tidybayes* R-package (Kay, 2020).

In all following plots and reported statistics, summaries are provided for the joint posterior of the model taking into account all distributional parameters during sampling. This provides overall nLEDs for any target comparison, rather than separate estimates of nLEDs between the bounds of 0 and 1 and for the extremes of 0 and 1. For reported results in tables, estimates are based on the median and credible interval around the median. The median was selected over the mean as it is a more robust measure of central

tendency for distributions with more than one mode. Thus, we do not provide individual statistics and plots for the individual distributional terms (e.g. for zero-one inflation, or conditional-one inflation) as we did not specify any hypotheses related to these individual terms. Instead, the zero-one-inflated Beta models are used purely to improve model fit and to make more accurate predictions about the overall differences in nLEDs across conditions. Ninety percent credible intervals are used to summarise uncertainty in the estimates as these intervals are more stable than wider intervals when given a limited number of samples from the posterior (Kruschke, 2014). Differences in nLEDs between conditions were obtained using the compare_levels() function from the *tidybayes* package, which provides a more accurate and reliable method of establishing group differences than visual inspection of whether credible intervals overlap from estimates of the individual groups (Schenker & Gentleman, 2001). Again, the posterior was summarised as the median difference and 90% credible interval around that median.

To determine support for hypotheses using these estimates, the probability of direction *P*(*direction*), or *pd*, is provided as calculated using the *bayestestR* R-package (Makowski et al., 2019b). This is defined as the proportion of the posterior that is of the same sign as the median. In previous simulations, the *pd* has been found to be linearly related to the frequentist *p*-value (Makowski et al., 2019a). The *pd* therefore provides an index of the existence of an effect and the degree of certainty in whether the effect is positive or negative. This can be used to ultimately reject the null hypothesis, but like the frequentist *p*-value, does not give a reliable estimate of evidence in support of the null hypothesis. Unlike the frequentist *p*-value, a ‘significant’ effect here is typically associated with a larger proportion of the posterior being of the same sign as the median (e.g. a *p*-value of <.05 is akin to a *pd* of >.95).

Vocabulary Test

We compared vocabulary test performance across conditions to make sure that participants in the four conditions did not differ in word learning ability. We also compared vocabulary test performance for non-contrastive and contrastive words to check whether inherent differences in learnability between non-contrastive words and the standard variants of contrastive words might be responsible for lower accuracy in processing contrastive words after literacy training. Recall that participants had not encountered any different variants at this point. If contrastive words are inherently as learnable as non-contrastive ones there should be no difference between them in vocabulary test performance in the No Dialect condition. For the dialect conditions, a comparison between non-contrastive words and the dialect variants of the contrastive words would reveal whether dialect variants themselves are inherently more difficult to learn which would have repercussions for the Dialect Literacy training condition. Figure 1 provides a descriptive summary of the nLEDs for each word type within each condition. These are presented as empirical means and standard errors (adjusted for within-subjects effects using the Morey (2008) correction) along with densities and points representing mean scores for each participant (panel A) and posterior medians with 80% and 90% credible intervals (panel B).

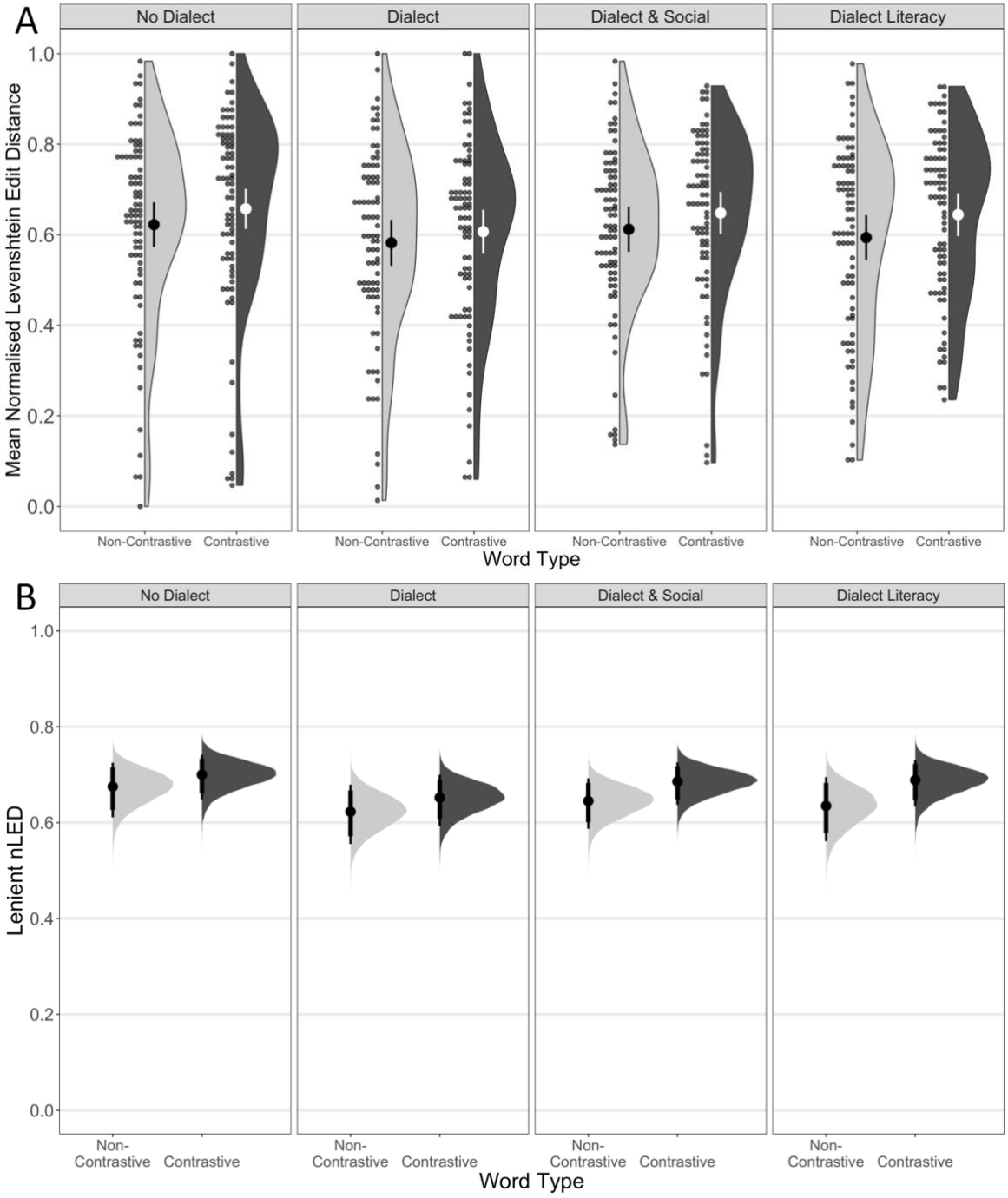

Figure 1: Vocabulary test as a function of Word Type and Exposure condition. Panel A shows mean nLEDs; small dots indicate by-participant means; large dots with whiskers indicate by-condition means ± 1 *SEM*. Panel B shows joint posteriors; point ranges indicate median ± 80% and 90% credible intervals.

Although the medians in Figure 1, Panel B lie slightly above the means in Figure 1, Panel A both plots show a similar trend, demonstrating that the choice of priors does not substantially skew the estimates. All estimated median nLEDs lie at or above 0.623 suggesting that word learning is generally fairly poor in all exposure conditions. To test whether the performance differences between non-contrastive and contrastive words were reliable, we obtained posterior medians with 80% and 90% credible intervals for the difference between contrastive and non-contrastive words within each exposure condition (see Figure 2 and Table 3).

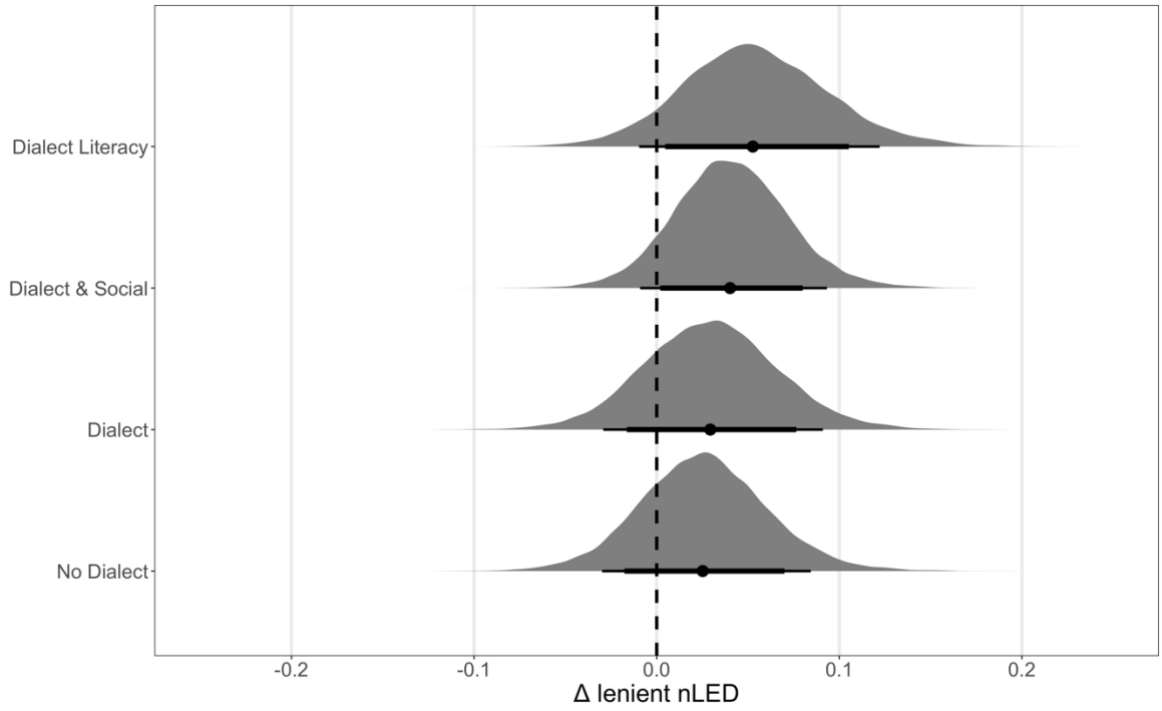

Figure 2: Joint posterior median nLEDs for non-contrastive words subtracted from contrastive words within each Exposure condition for the vocabulary test. Point ranges show posterior median ± 80% and 90% credible intervals.

Table 3. Joint posterior medians, 90% credible intervals and probability of direction for median Vocabulary Test nLEDs for non-contrastive words subtracted from contrastive words by Exposure condition. Larger values indicate worse performance with contrastive words.

Exposure Condition	Median nLED	90% Credible Interval	*P*(Direction)
No Dialect	0.025	[-0.03, 0.08]	0.774
Dialect	0.029	[-0.03, 0.09]	0.794
Dialect & Social	0.040	[-0.01, 0.09]	0.911
Dialect Literacy	0.053	[-0.01, 0.12]	0.920

Although performance in all conditions was slightly better (i.e. nLEDs were slightly lower) for non-contrastive compared to contrastive words, all 90% credible intervals included zero indicating that there is insufficient evidence to rule out an effect in the opposite direction. Crucially, a *pd* of 0.774 in the No Dialect condition does not constitute sufficient evidence for inherent difficulties with the standard variants of the contrastive words. Even in the three dialect conditions there was not enough evidence to suggest that the dialect variants were inherently more difficult to learn although a trend in this direction may have been due to difficulties with learning words with the phoneme /x/ which is unfamiliar to many speakers of English outside of Scotland where it forms part of the phonological repertoire of Scots dialects.

Literacy Testing

To answer questions pertaining to our pre-registered hypotheses, and to generate plots for the relevant summaries, we obtained samples from the posterior for the comparisons of interest and test our hypotheses using Probability of Direction *pd*.

Contrastive vs. non-contrastive words. We compared nLEDs for non-contrastive and contrastive words across tasks in the four exposure conditions. Figure 3 presents the

descriptive results for the condition means (panel A) and the associated posterior medians with 80% and 90% credible intervals (panel B).

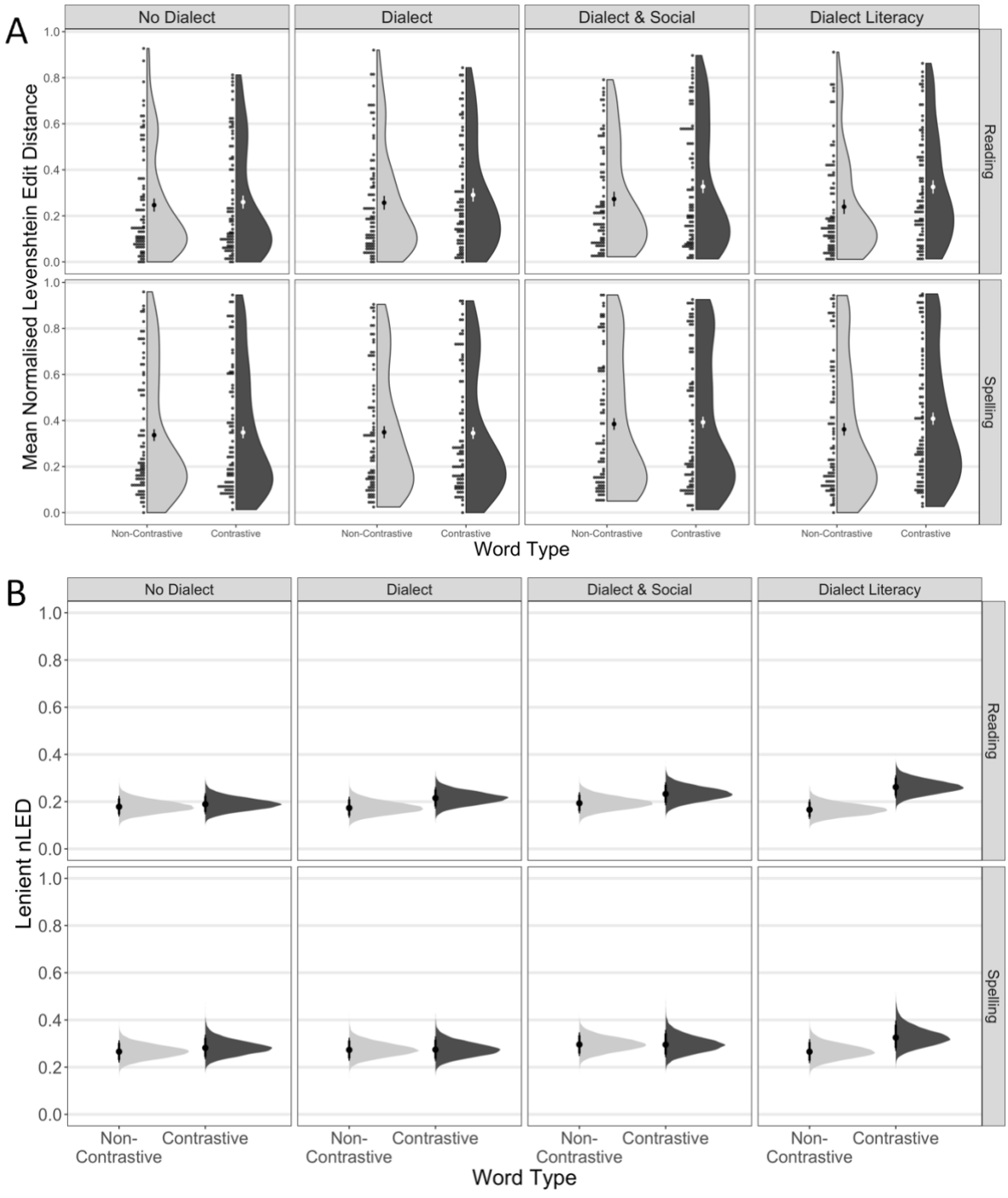

Figure 3: Literacy testing performance as a function of Word Type and Exposure condition. Panel A shows mean nLEDs; small dots indicate by-participant means; large dots with whiskers indicate by-condition means ± 1 *SEM*. Panel B shows joint posteriors; point ranges indicate median ± 80% and 90% credible intervals.

Reading and spelling performance (the highest median nLED was 0.325) was better than vocabulary test performance prior to literacy training. To directly compare contrastive and non-contrastive words we obtained posterior medians and 80% and 90% credible intervals of their difference for each task and exposure condition (see Figure 4 and Table 4).

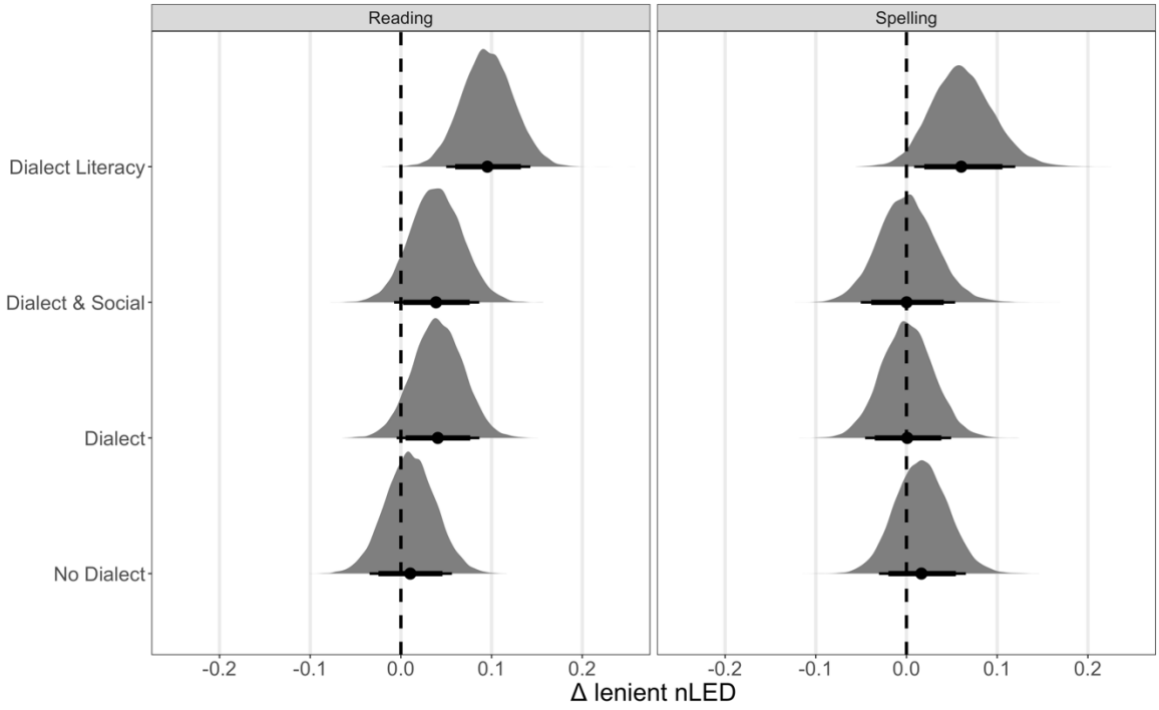

Figure 4: Joint posterior median differences between the nLEDs for reading (left panel) and spelling (right panel) of contrastive and non-contrastive words within each Exposure condition in the literacy testing phase. Point ranges show posterior medians ± 80% and 90% credible intervals.

Table 4. Joint posterior medians, 90% credible intervals and probability of direction for median nLEDs for non-contrastive words subtracted from those for contrastive within each Task and Exposure condition in the literacy testing phase.

Task	Variety Exposure	Median	90% Credible Interval	*P*(Direction)
Reading	No Dialect	0.010	[-0.04, 0.06]	0.648
Reading	Dialect	0.041	[-0.00, 0.09]	0.931
Reading	Dialect & Social	0.039	[-0.01, 0.09]	0.914
Reading	Dialect Literacy	0.095	[0.05, 0.14]	0.999
Spelling	No Dialect	0.016	[-0.03, 0.06]	0.714
Spelling	Dialect	0.001	[-0.05, 0.05]	0.512
Spelling	Dialect & Social	0.000	[-0.05, 0.05]	0.502
Spelling	Dialect Literacy	0.060	[0.01, 0.12]	0.972

While nLEDs appear generally to be higher for contrastive compared to non-contrastive words for reading and spelling in all exposure conditions, this difference was smallest in the No Dialect condition and largest in the Dialect Literacy condition. A direct comparison of contrastive and non-contrastive words for each exposure condition showed that in the Dialect Literacy condition, the 90% credible intervals around the nLED difference scores did not contain zero and the *pd*s are 0.999 and 0.972 for reading and spelling, respectively. This suggests that when literacy training comprised both standard and dialect orthographic forms, contrastive words reliably incurred an additional accuracy cost both in reading and in spelling. In the Dialect and Dialect & Social conditions, there was only weak evidence for lower accuracy in reading contrastive, compared to non-contrastive, words. In these two conditions, the 80% credible interval for the effect of Word Type contained zero for spelling but not for reading (see Figure 4), and *pd*s indicated that over 90% of the posterior exhibits the median's sign. For all other contrasts the 90% credible intervals around difference scores for nLEDs contained zero and the *pd*s indicate that there is a less than a 72% probability

of exhibiting the same sign as the median. Counter to our pre-registered predictions, this suggests that there is only convincing evidence of lower accuracy for contrastive words in the Dialect Literacy condition, not just in reading but also in spelling.

Untrained Words. Figure 5 shows the condition means for reading and spelling of untrained words, which for illustrative purposes are shown in comparison to trained words. The estimated posterior medians with 80% and 90% credible intervals are presented in Figure 6.

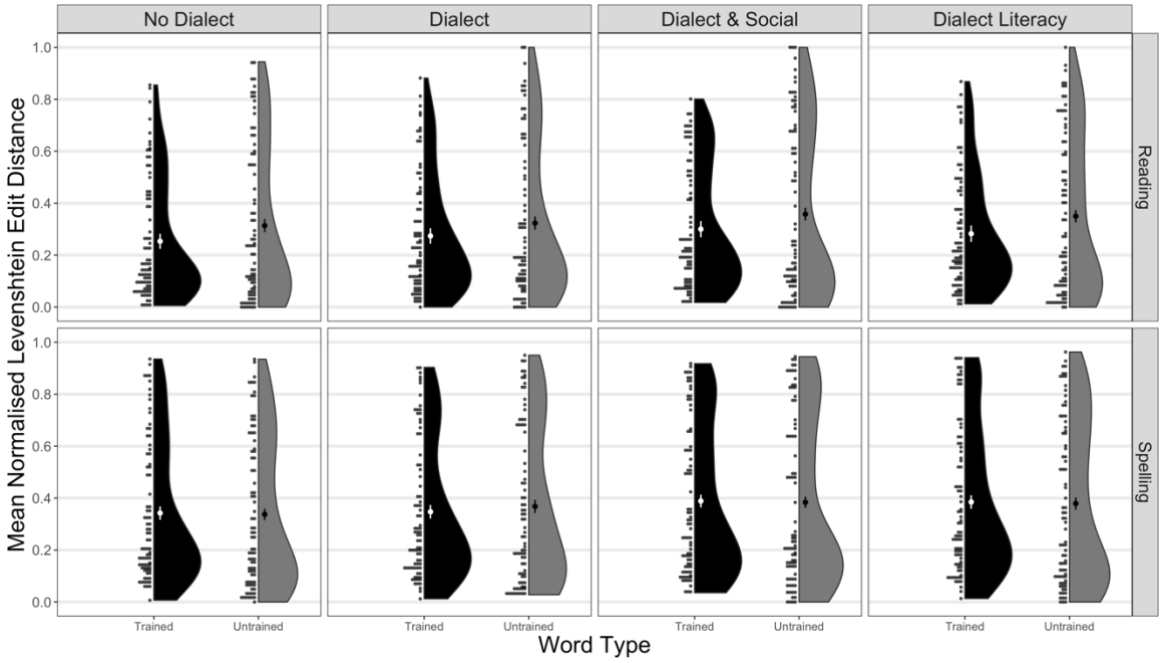

Figure 5: Mean nLEDs for trained and untrained words for reading (upper panels) and spelling (lower panels) in each exposure condition in the testing phase. Small dots indicate by-participant means. Large dots and whiskers indicate by-condition means ±1 *SEM*.

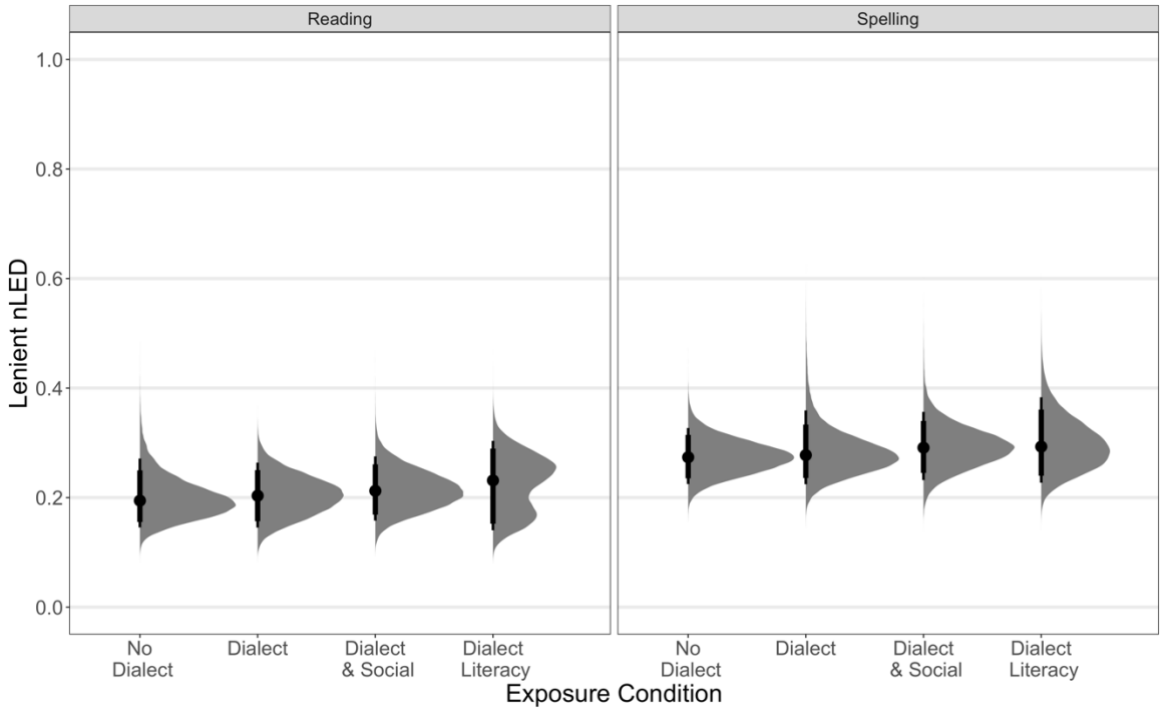

Figure 6: Joint posterior nLEDs for the effect of Exposure condition for reading (left panel) and spelling (right panel) on untrained words in the literacy testing phase. Point ranges show posterior median ± 80% and 90% credible intervals.

Overall, performance was better for reading than for spelling. Of all exposure conditions, the highest median nLEDs were found in the Dialect Literacy condition (0.244 and 0.292 for reading and spelling, respectively). To compare reading and spelling performance between the different exposure conditions we examined posterior medians with 80% and 90% credible intervals of the differences between the median nLEDs for all combinations of differences between exposure conditions (see Figure 7 and Table 5).

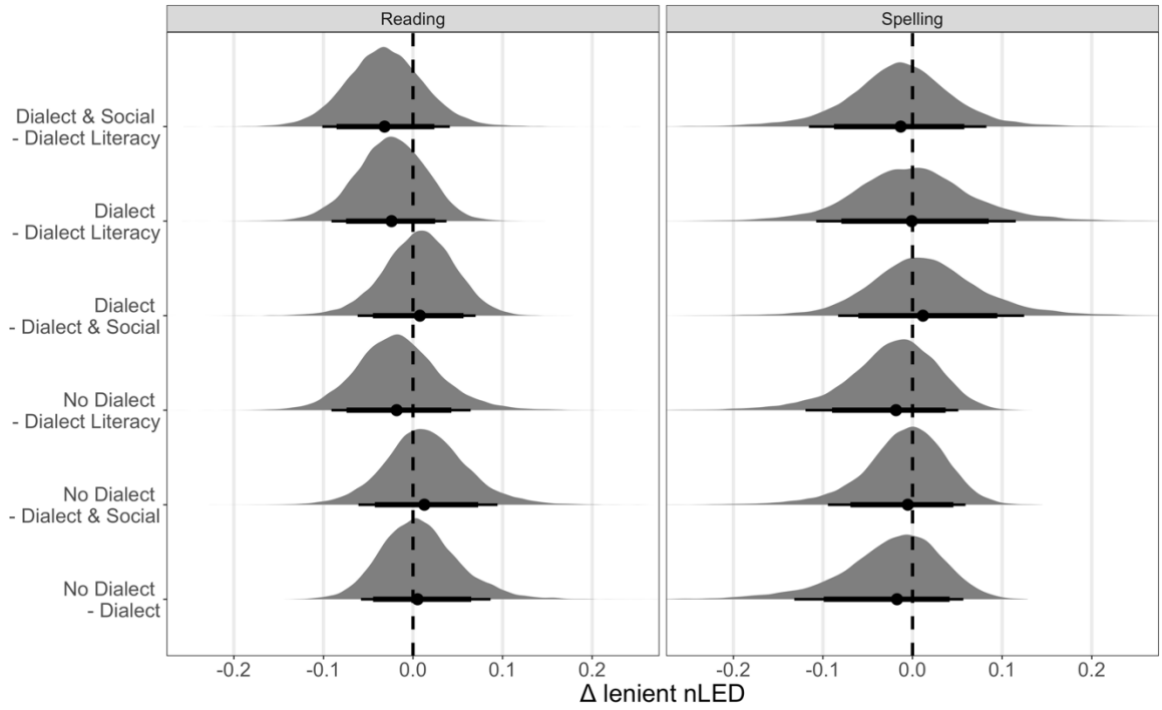

Figure 7: Joint posterior median differences between the nLEDs for all comparisons between exposure conditions for reading (left panel) and spelling (right panel) of untrained words in the literacy testing phase. Point ranges show posterior median ± 80% and 90% credible intervals.

Table 5. Joint posterior medians, 90% credible intervals and probability of direction for the median nLEDs for all differences between exposure conditions for reading and spelling of untrained words in the literacy testing phase. Larger values indicate better performance in the second condition.

Task	Variety Exposure Contrast	Median	90% Credible Interval	*P*(Direction)
Reading	No Dialect - Dialect	0.005	[-0.06, 0.09]	0.553
Reading	No Dialect - Dialect & Social	0.013	[-0.06, 0.09]	0.619
Reading	No Dialect - Dialect Literacy	-0.018	[-0.09, 0.06]	0.662
Reading	Dialect - Dialect & Social	0.008	[-0.06, 0.07]	0.580
Reading	Dialect - Dialect Literacy	-0.024	[-0.09, 0.04]	0.731
Reading	Dialect & Social - Dialect Literacy	-0.032	[-0.10, 0.04]	0.772
Spelling	No Dialect - Dialect	-0.017	[-0.13, 0.06]	0.636
Spelling	No Dialect - Dialect & Social	-0.005	[-0.09, 0.06]	0.551
Spelling	No Dialect - Dialect Literacy	-0.018	[-0.12, 0.05]	0.658
Spelling	Dialect - Dialect & Social	0.012	[-0.08, 0.12]	0.583
Spelling	Dialect - Dialect Literacy	-0.001	[-0.11, 0.12]	0.506
Spelling	Dialect & Social - Dialect Literacy	-0.013	[-0.12, 0.08]	0.600

All 90% credible intervals included zero and all *pd*s were less than or equal to 0.772. Thus, counter to our pre-registered predictions, we found no reliable differences across conditions in the ability to read and spell untrained words. This indicates that exposure to a dialect did not have a negative impact on decoding ability.

Exploratory Analyses with Vocabulary Test Performance as a Covariate

The low performance during Vocabulary Test with a mean nLED of 0.62 across all conditions suggests that it may have been difficult for many participants to form stable links between a word's phonology and its meaning. As a result, in the conditions containing dialect variants especially the poorer vocabulary learners there may not have had sufficiently stable representations of competing variants available when attempting to access the meaning of the standard variants during literacy testing. If this was the case, only strong vocabulary learners should exhibit lower accuracy with contrastive

words. To explore this possibility, the final analysis examined whether vocabulary test performance interacted with the effect of Word Type in reading and spelling in the three dialect conditions. Moreover, if successful word learning ensures more stable representations of dialect variants then there is less pressure to rely on phoneme-grapheme conversion rules, and, hence, a phonological decoding strategy may be invoked less frequently resulting in greater difficulties when trying to read untrained words. These possibilities are explored in the models presented below, which are fitted in the same way as the pre-registered models. We note that this exploratory analysis was not pre-registered but was deemed informative as vocabulary learning was far from ceiling. Moreover, caution is needed when interpreting the probability of direction (*pd*) in these exploratory models.

Contrastive vs. non-contrastive words as a function of individual differences in word learning. We first explored whether performance on the vocabulary test after the exposure phase predicted both overall reading and spelling performance as well as the size of the accuracy cost for contrastive words. Figure 8 shows the median nLED for reading and spelling performance for non-contrastive and contrastive words based on samples from the posterior as a function of mean nLED in the vocabulary test (panel A) and a median split of vocabulary test performance, dichotomising participants into better or worse word learners (panel B).

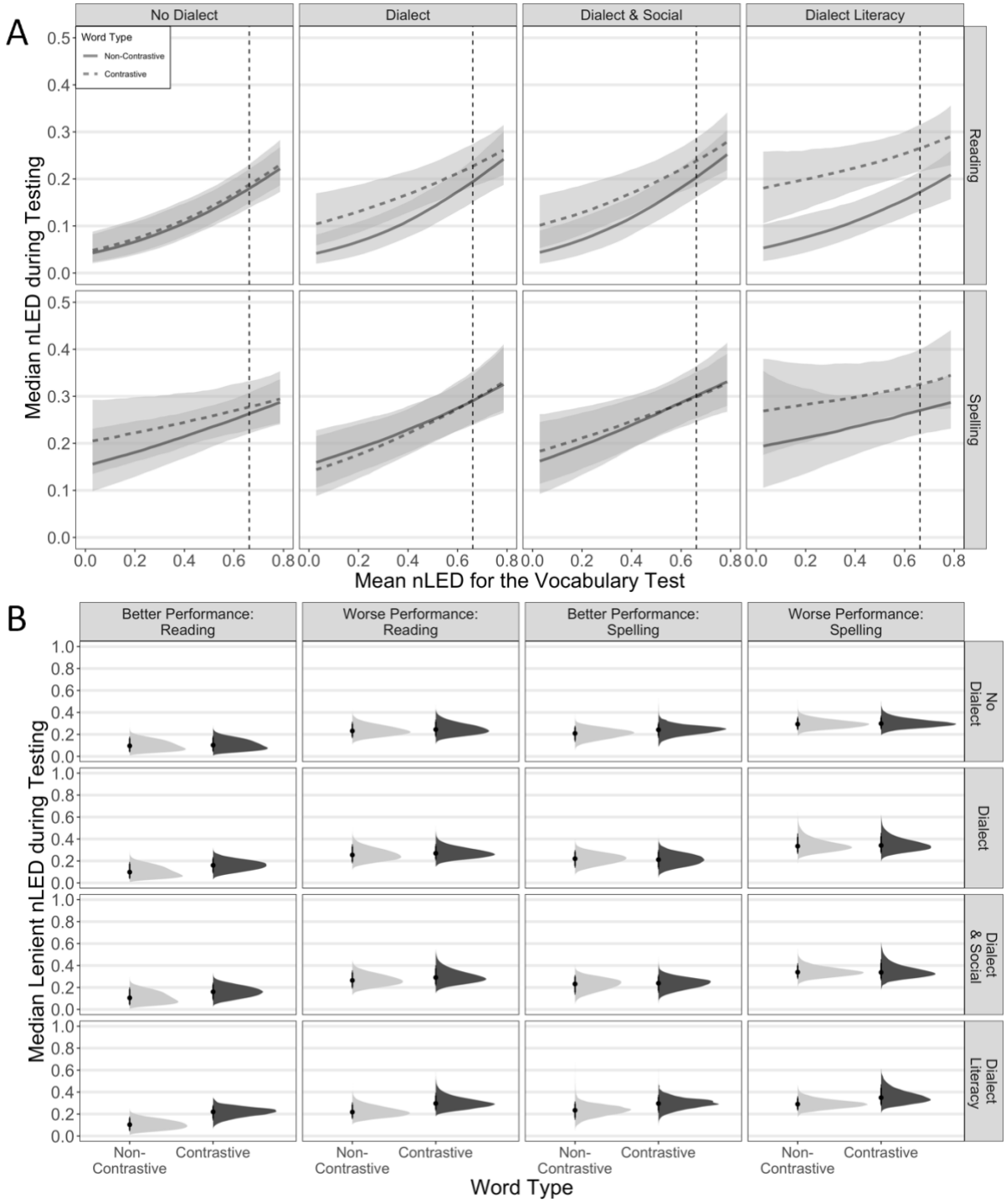

Figure 8: Joint posterior nLEDs for reading of non-contrastive and contrastive words in all exposure conditions as a function of mean vocabulary test performance for reading (top row of graphs in panel A) and spelling (bottom row of graphs in panel A). Note that for both measures low values indicate good performance and high values indicate poor

performance. Lines and ribbons show posterior median ± 90% credible intervals. The dotted vertical line indicates median vocabulary test performance. Panel B shows the joint posterior nLEDs in all Exposure conditions as a function of median-split vocabulary test performance into better and worse; point ranges show posterior median ± 80% and 90% credible intervals.

The results depicted in Figure 8, panel A suggest that poorer word learning was associated with overall poorer reading and spelling performance, as indicated by the association between nLEDs for the vocabulary test and for reading and spelling. In the Dialect and Dialect & Social conditions, the reading accuracy difference between contrastive and non-contrastive words was higher in successful word learners. Even in the Dialect Literacy condition, where reading accuracy for contrastive words was consistently below that for non-contrastive words across the entire spectrum of vocabulary test performance, the accuracy difference was nonetheless somewhat greater in successful word learners. As expected, in the No Dialect condition, there was no accuracy reduction for reading contrastive words because no competing variants existed. It is also noteworthy that reading performance for non-contrastive words was comparable across conditions showing that dialect exposure incurred a cost for reading contrastive words rather than a benefit for reading non-contrastive words. These observations qualify the findings of weak evidence for reduced reading accuracy for contrastive words presented above by showing that the hypothesised cost was only incurred when dialect word representations were sufficiently stable and entrenched to cause interference. It is noteworthy that counter to our hypothesis there was no attenuation of the contrastive word accuracy cost in the Dialect & Social condition

compared to the Dialect condition. The results also confirm that in the Dialect Literacy condition the contrastive word accuracy cost was ubiquitous not just in reading but also in spelling.

To estimate the evidence for the interaction between vocabulary test performance and the contrastive word accuracy cost, we computed and summarised samples from the posterior based on a median split of vocabulary test performance, dichotomising participants into better or worse word learners (see Figure 8, panel B), and compared samples from the posterior for differences between contrastive and non-contrastive words (see Figure 9).

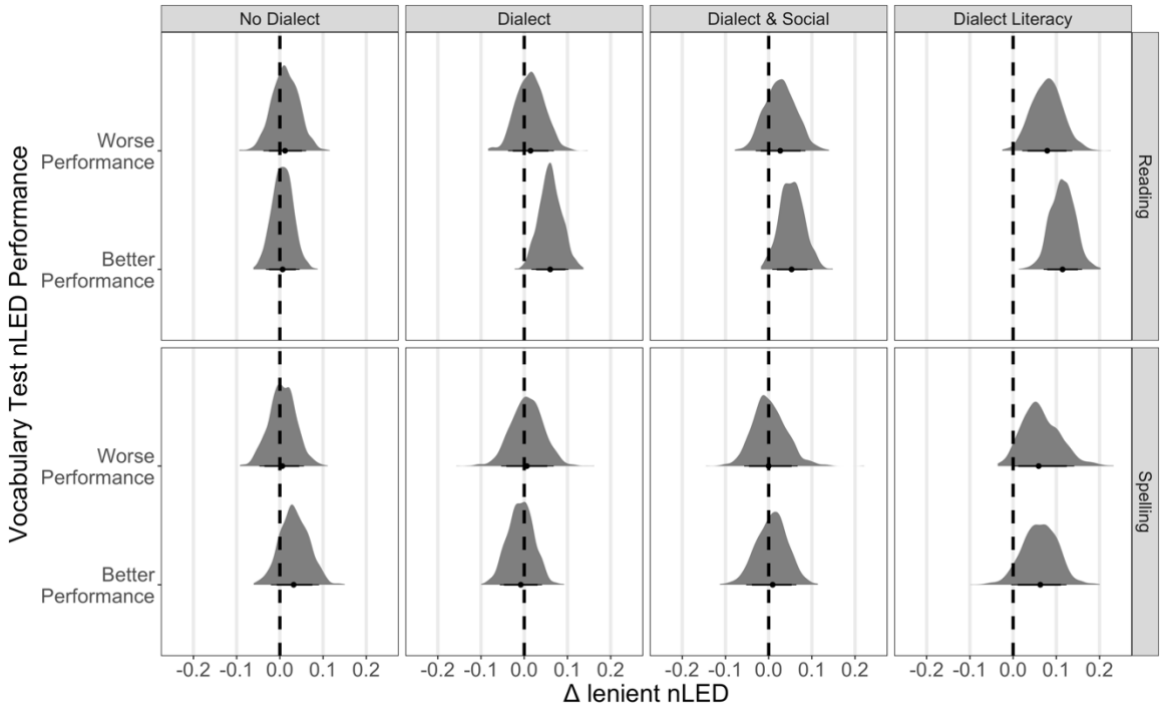

Figure 9: Joint posterior median differences between the nLEDs for the comparison between each level of Word Type (contrastive − non-contrastive) by Task, Exposure condition, and median-split vocabulary test performance. Point ranges show posterior median ± 80% and 90% credible intervals.

The data depicted in Figure 9 (for the table of full results please see the Supplemental Material) show that the 90% credible intervals the nLED difference scores between contrastive and non-contrastive words for the successful word learners in the Dialect and Dialect & Social conditions did not include zero confirming that a greater accuracy cost for contrastive words was contingent upon sufficiently strong knowledge of dialect variants. Only those learners who managed to acquire more stable word form representations during the exposure phase incurred an accuracy cost in reading contrastive words. The data show also that in the Dialect Literacy condition, which comprised literacy training for both standard and dialect words, an accuracy cost was evident for reading and for spelling of contrastive words in poor and in successful word learners.

Performance with untrained words as a function of individual differences in word learning. Next, we explored whether a dialect benefit arose in poorer word learners who had less stable word form representations and hence less access to phonological forms via word meanings cued by the picture, and who would therefore have been more likely to gain practice with sequential decoding which, in turn, could benefit the processing of untrained words (see Figure 10).

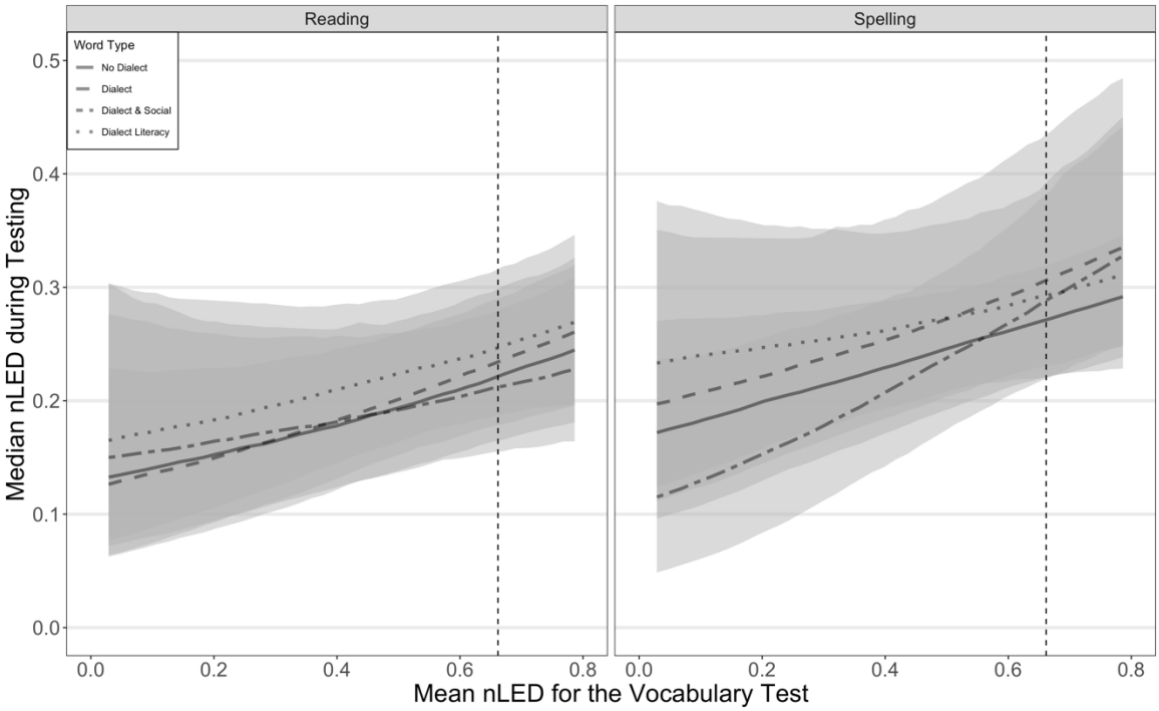

Figure 10: Joint posterior nLEDs for untrained words by Exposure condition for reading (left panel) and spelling (right panel) as a function of mean nLED in the vocabulary test. Lines and ribbons show posterior median ± 90% credible intervals. The dotted vertical line indicates the median vocabulary test performance.

The results show that word learning predicted overall performance with untrained words. To see whether vocabulary test performance interacted with the effect of exposure condition we again computed and summarised samples from the posterior based on a median split of participants based on vocabulary test performance (see Figure 11).

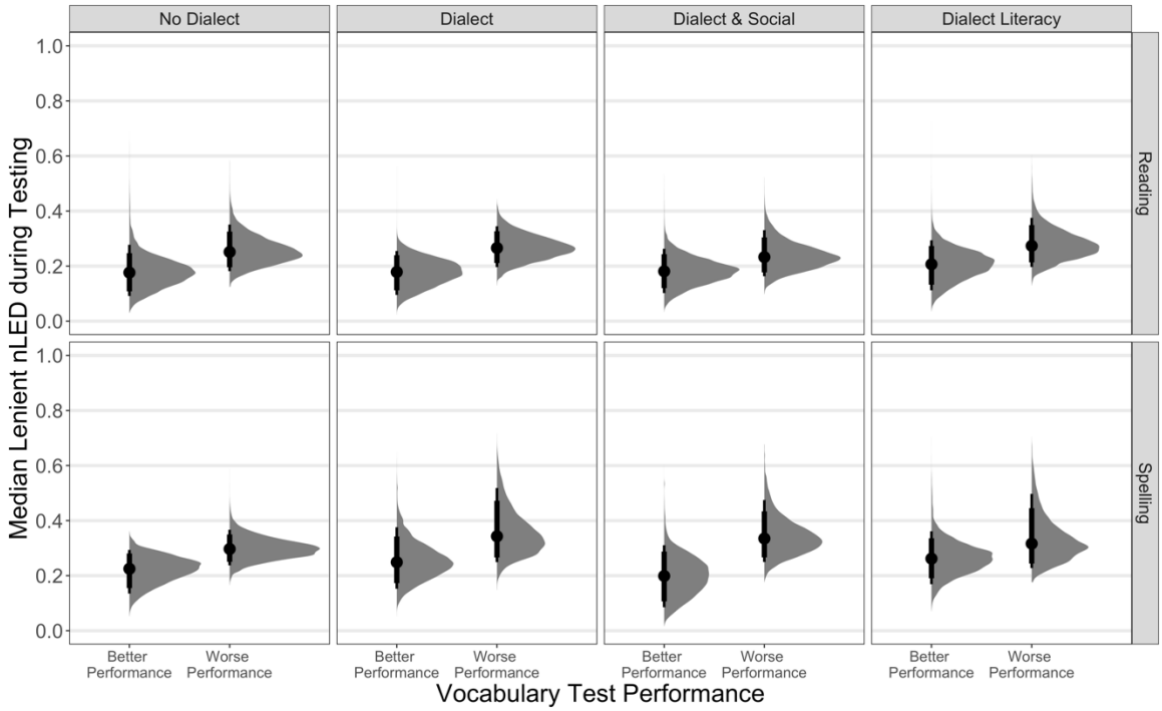

Figure 11: Joint posterior nLEDs for the effect of exposure condition on performance with untrained words in reading (top panels) and spelling (bottom panels) as a function of better or worse vocabulary test performance. Point ranges show posterior median ± 80% and 90% credible intervals.

Figure 11 shows a clear overall link between vocabulary test performance and reading and spelling of untrained words, such that better word learners also tended to perform better with reading and spelling of untrained words. However, the results did not indicate any differences between exposure conditions. To estimate the reliability of an effect of exposure condition we again obtained posterior medians with 80% and 90% credible intervals for all differences between exposure conditions for each task and for each level of vocabulary test performance, which are shown in Figure 12.

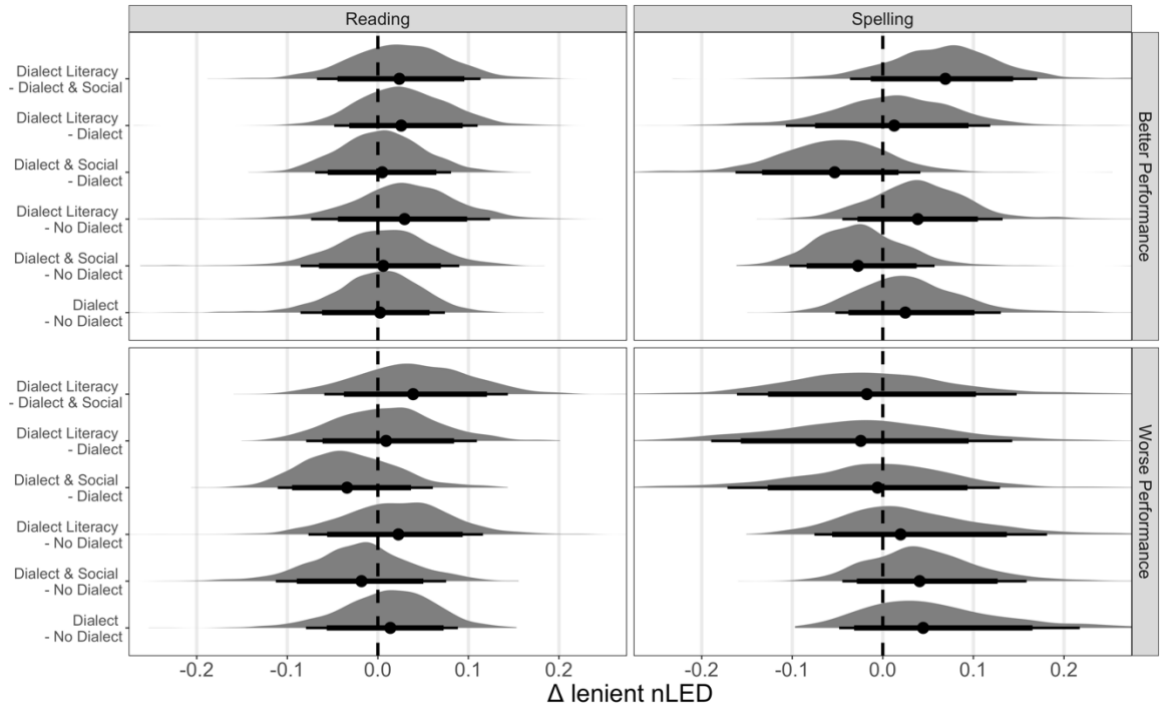

Figure 12: Joint posterior median differences between the nLEDs for the comparison of performance with untrained words between all exposure conditions for reading (left panels) and spelling (right panels) by median-split vocabulary test performance. Point ranges show posterior median ± 80% and 90% credible intervals.

Figure 12 shows that all 90% credible intervals include zero for both levels of word learning ability suggesting that regardless of how well entrenched the word forms were, there were no differences between exposure conditions in participants' ability to decode untrained words using grapheme-phoneme conversion for reading and phoneme-grapheme conversion for spelling.

Description of error types. Finally, to gain a better understanding of the kinds of errors participants produced we used automatic string comparison for an exploratory description of error types. We wanted to see whether the errors with contrastive words, especially in the Dialect Literacy condition, were indeed due to more frequent

production of the competing dialect variants. Overall, reading and spelling errors clustered into four types (examples below are provided for reading assuming similar patterns for spelling reflected in the associated graphemes): (a) *Dialect Word Match* where a dialect variant was produced in response to its standard counterpart (e.g., target /kublɛ/, response /xublɛ/), (b) *Dialect Word Mismatch* where a dialect variant was produced in response to another standard contrastive word (e.g., target: /skɛfi/, response: /xublɛ/), (c) *Standard Word Mismatch* where a standard word was produced in response to another standard word (e.g., target: /skɛfi/, response: /kublɛ/) and (d) any *Other Mismatch* which did not include word substitutions. Figure 13 presents the distribution of correct responses and error types for all conditions. While *Dialect Word Match* errors did not constitute the most frequent error type for contrastive words, they were indeed more frequent in all three Dialect exposure conditions than in the No Dialect condition. This confirms that interference from competing dialect variants makes a small but measurable contribution to the difficulty of learning to read in situations of dialect exposure prior to literacy learning. The error profiles of words also confirmed that accuracy in reading and spelling of untrained words was unaffected by dialect exposure condition.

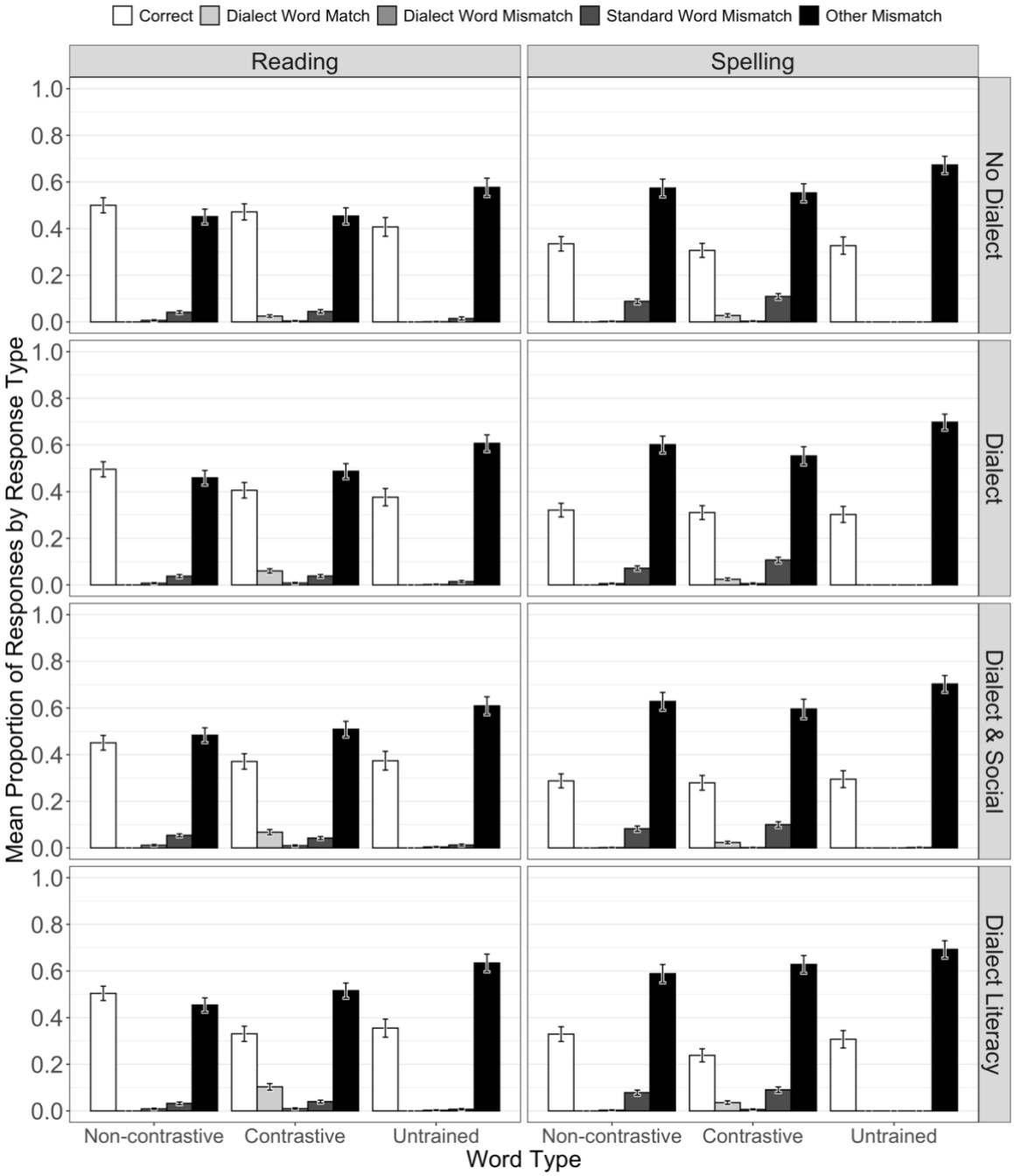

Figure 13: Distribution of response types (correct responses and four types of errors) for reading and spelling in all exposure conditions. Error bars show ± 1 *SEM*.

Discussion

Using adult artificial literacy learning to control for extra-linguistic confounds, we examined effects of dialect exposure on learning to read and spell artificial words. In contrast to recent connectionist simulations (Brown et al., 2015) and previous artificial literacy learning experiments where word learning of the dialect variety was interleaved with literacy training in a standard variety (Williams et al., 2020), this study introduced word learning before literacy training to simulate a learning situation where children acquire one variety (e.g. a regional dialect) at home before starting literacy training in a different variety (e.g. a 'standard'/mainstream variety) at school. We hypothesised that consolidating representations of dialect variants before literacy training commences might exacerbate interference from these dialect variants when reading contrastive words compared to a training regimen where word learning and literacy training are interleaved and dialect variants are not as well entrenched. Greater interference was expected because during reading learners might attempt to access phonological representations of words via their semantic representations, a strategy that might be prioritised when word knowledge is fairly stable. Our second goal was to explore the effects of information about the social context of dialect use (Dialect & Social condition) and of concurrent dialect literacy training (Dialect Literacy condition) on the reading accuracy for contrastive words and on the ability to decode untrained words, the latter designed in analogy to non-word reading tests. We operationalised accuracy cost as length-normalised Levenshtein edit distances (nLEDs) between participants' responses and the phonological or orthographic target forms, modelled as zero-one-inflated Beta distributions, and analysed these data using multilevel distributional Bayesian models. Vocabulary test performance prior to literacy training revealed no inherent difficulties

with learning of the standard variants of contrastive words but showed fairly low performance and considerable individual differences in participants' overall ability to learn artificial words. Note that identifying which factors may have been responsible for these individual differences, such as working memory capacity or native language vocabulary size, was beyond the scope of this study; instead, we explored how these individual differences in word knowledge acquired prior to literacy training affected learning to read and spell under conditions of dialect exposure. Below we discuss our findings with respect to each of the three dialect exposure conditions.

Effects of dialect exposure. Counter to our hypotheses, we found only weak evidence for an additional accuracy cost associated with reading contrastive words in the Dialect condition. This is in contrast to our previous findings (Williams et al., 2020) where participants received exactly the same amount of input as in this study but dialect word learning was interleaved with standard literacy training. When word learning is front-loaded, exposure to the new standard variants during literacy training can potentially override the earlier dialect representations, a process similar to catastrophic interference in neural networks (McCloskey and Cohen, 1989) where a change in the input-output mappings results in the replacement of old activation patterns by new ones. This may be the reason why no overall accuracy cost for contrastive words was observed. However, adding vocabulary test performance as a covariate to an exploratory model revealed that lower accuracy in reading contrastive words was predicted by participants' word learning ability: Only for strong, but not for poor word learners, was there evidence for an accuracy cost associated with reading contrastive words. This shows that if representations of dialect variants become sufficiently entrenched early on, they continue to interfere with reading of the standard variants of contrastive words.

This finding may appear at odds with findings of the supporting role of oral word knowledge in literacy acquisition (for a review see Taylor et al., 2015); however, such findings have been obtained in monolingual literacy learning situations where the word forms acquired prior to literacy learning are identical to the ones accessed through reading. In contrast, Bühler et al. (2018) showed that greater exposure to Swiss German prior to literacy learning in Standard German had a negative effect on early literacy skills. This was largely due to dialect-specific spelling errors akin to the errors observed here for spelling of contrastive words in the Dialect Literacy condition. At the same time, greater exposure to Swiss German prior to literacy was positively associated with phonological awareness. The accuracy cost for contrastive words observed in strong word learners also aligns with results from recent connectionist simulations showing that oral word knowledge acquired prior to literacy modulates effectiveness of phonology-based vs. semantics-based reading instructions, especially for written word comprehension (Chang et al., 2020). Our data extend these insights by suggesting that oral word knowledge acquired prior to literacy training modulates the local cost of dialect exposure on the processing of contrastive words when participants attempt to access the sound of a word from its spelling via its meaning and the encoding of an alternative variant of semantics-phonology mappings interferes with phonological output. It is unlikely that this finding is confined to inconsistent orthographies like the one employed here, as in Williams et al. (2020) we observed a higher cost for reading contrastive words for a consistent orthography as well.

In the Dialect and Dialect & Social conditions, the word-knowledge-dependent local accuracy cost was found only for reading but not for spelling of contrastive words, most likely because early stages of spelling acquisition are much more reliant on

sequential conversion of phonemes into graphemes rather than some form of more direct retrieval of orthographic forms based on meaning. Apparently, in these conditions the dialect representations were not strong enough to interfere with maintenance of the word form in working memory while spelling was executed.

In contrast to our previous training regimen of word learning interleaved with literacy training (Williams et al., 2020), this study did not find a dialect benefit for the processing of untrained words. We had assumed that the dialect benefit in that earlier study arose from greater reliance on phonological decoding when participants were faced with increased variability as they encountered both standard and dialect variants in relatively close temporal proximity. In the Dialect condition of the present study, dialect variants for half of the words appeared in the first part of the experiment and standard variants in the second part. Thus, once participants reached the literacy training phase there may not have been a sustained perception of input variation and hence no prioritisation of a phonological decoding strategy. It is important to note, however, that we found no detrimental effect of dialect exposure on the ability to decode untrained words either.

Effects of contextual information. Previously, implementing explicit coding for what variety a word belongs to in a neural network (Brown et al., 2015) and training bidialectal children explicitly in dialect awareness (Johnson et al., 2017) had been shown to result in improved reading in a standard variety. We had therefore hypothesised that explicit information about the social context of variety use would facilitate pre-activation of the target variety, which should reduce the likelihood of errors in processing contrastive words and thereby improve literacy learning. However, our results showed no difference with respect to the cost associated with processing

contrastive words between the Dialect and the Dialect & Social condition compared to the No Dialect condition: For strong word learners in the Dialect & Social condition the accuracy cost associated with processing contrastive words was similar to that in the Dialect condition. Thus, the contextual information that linked the voice of the speaker to a geographical locale was not sufficient to pre-activate the target variety sufficiently so as to alleviate interference from dialect variants in reading.

This result seems to contradict findings of a benefit from explicit contextual information about dialect use (Johnson et al., 2017). We argue that this discrepancy is likely due to task difficulty and insufficient opportunity for learning the social cue. Our failure to find a benefit from social context may simply reflect the need for more time required to associate lexical variants with specific contexts. Even though we presented a relatively simple contextual cue – an association of speaker voice with a depiction of geographical locale, to mimic information about regional dialects – learning this seemingly arbitrary cue may have proved too difficult during this already challenging artificial language learning task presented within one experimental session. The notion that acquiring such contextual information is difficult aligns well with developmental evidence for the acquisition of sociolinguistic competence. For example, the ability to discriminate and identify regional US dialects develops only by adolescence (McCullough, Clopper & Wagner, 2019) or even early adulthood (Dossey, Clopper & Wagner, 2020) indicating that learning to link variety-specific linguistic features with the associated social-contextual information takes considerable time. It is thus doubtful that a conceptualisation of contextual information as activation of a single context unit in a neural network (Brown et al., 2015) does justice to the protracted development of socio-linguistic competence. In recognition of this, the dialect awareness intervention

described in Johnson et al. (2017) was administered over a period of 4 weeks, which stands in sharp contrast to our training duration of approximately 60 minutes. Given how difficult it is to link social context with variety-specific linguistic features, even our fairly simple operationalisation may have proved too difficult for our participants to learn in such a short time period. However, processing of untrained words was very similar to the other conditions suggesting that again, receiving contextual information about dialect use neither hinders nor benefits acquisition of decoding skills.

Effects of dialect literacy training. When literacy training in the two varieties was combined such that participants first learned to read and spell in the standard and then in the dialect variety, before being tested in the standard variety, we found a significant accuracy cost associated with processing of contrastive words. This suggests that the introduction of dialect literacy training kept representations of both variants active thereby exacerbating interference. What is noteworthy is that in this condition there was also an accuracy cost associated with the spelling of contrastive words. This is an unexpected result if one assumes that early spelling relies mainly on sequential grapheme-phoneme conversion. It could be explained by the fact that spelling requires to some extent the short-term maintenance of word's phonology: As participants attempt to assemble a word's spelling by successively clicking the associated graphemes on the virtual keyboard in the right order, they need to keep the phonological representation of the word active in working memory. This may require engagement of the phonological loop (Baddeley, 2003) and subsequent re-retrieval of the phonological form from working memory as the trace decays. If the competing dialect variant remained highly accessible in the Dialect Literacy condition it would have been more likely to cause interference with this process.

Although the Dialect Literacy condition was inspired by attempts to teach dialect literacy in educational settings, we would urge extreme caution in extrapolating from these findings to raise doubt in the viability of such educational interventions. First, although our participants were trained in both varieties, we tested them only in one, to maintain comparability with the other conditions. However, such a test does not do justice to the potential learning gains these participants may have made with respect to reading and spelling dialect variants. Secondly, the accuracy cost, although reliable, is small and confined just to contrastive words, i.e. words with dialect variants. Because the training duration of our artificial literacy learning experiment was very limited for reasons of feasibility, our observations pertain only to the very early stages of literacy learning. It is conceivable that with continuous reading and spelling practice in both varieties interference effects will reduce even further and become as manageable as they are for biliteral bilinguals. Thirdly, many real-world literacy interventions that incorporate non-standard language varieties into the curriculum convey indirect benefits to literacy and general educational attainment in addition to improving historical, social, and cultural awareness. For example, the role of Scots language in the Curriculum for Excellence (Education Scotland, 2017) notes that the use of Scots in the classroom removes barriers to learning in the home environment and develops confidence in individuals who otherwise may be penalised for using their home language in class. Finally, in the Dialect Literacy condition time spent training in the nonstandard variety came at a cost to the standard variety, while in the real world any literacy training in non-stantard varieties is likely to be in addition to, rather than instead of, training in the standard variety.

In conjunction with our previous findings (Williams, 2020), the present findings suggest that dialect exposure can incur a local accuracy cost for reading of contrastive words but only if the representations of the associated dialect variants are sufficiently strong, be it because dialect variants are continuously presented in interleaved training or because dialect word knowledge prior to onset of literacy training have been well entrenched. Most importantly, when extra-linguistic factors were controlled in these artificial literacy learning studies no evidence was found that dialect exposure had any adverse effects on the acquisition of decoding skills. In other words, dialect exposure is unlikely to reduce the ability to associate print to sound, as suggested by Labov (1995). We would argue that the route that is considered to be most beneficial for reading acquisition and vocabulary expansion – phonological decoding (Castles, Rastle, & Nation, 2018) – is likely to be unimpaired by dialect exposure. Taken together, our findings suggest that dialect exposure in itself need not be considered an obstacle to literacy and that further research needs to scrutinise the role of extra-linguistic factors in literacy learning outcomes.

In closing we would like to note an important caveat. Most adults who are asked to learn to read and spell in an artificial language are probably able to flexibly deploy different routes to reading in a strategic way: they might attempt to access the phonological form of familiar words based on semantic cues or partially decoded initial phonemes, but may favour application of grapheme-phoneme conversion rules when words appear unfamiliar. The situation may be different for children who are just beginning to acquire the alphabetic principle that underpins phonological decoding and therefore may lack the capacity for strategic use of different routes to reading. We therefore urge caution in generalising our findings to literacy acquisition in bidialectal

children, and view these artificial literacy learning studies with adults as a first step towards trying to disentangle linguistic and extra-linguistic factors in this domain. Hopefully, future research will be able to adapt the methodology employed here to gain a better understanding of whether and when bidialectal children are able to deploy different reading strategies in a flexible manner when encountering input variability that arises when encountering multiple linguistic varieties.

References

Apfelbaum, K. S., Hazeltine, E., & McMurray, B. (2013). Statistical learning in reading: Variability in irrelevant letters helps children learn phonics skills. *Developmental Psychology*, *49*(7), 1348–1365. https://doi.org/10.1037/a0029839

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189-208. doi: https://doi.org/10.1016/S0021-9924(03)00019-4

Baratz, J. C. & Shuy, R. W. (Eds.) (1969). *Teaching Black children to read*. Washington, D.C.: Center for Applied Linguistics.

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

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

BBC News (2013a) Middlesbrough school bans use of local dialect. Retrieved from https://www.bbc.co.uk/news/av/uk-21361324

BBC News (2013b). Colley Lane school in Halesowen bans Black Country dialect. Retrieved from https://www.bbc.co.uk/news/uk-england-birmingham-24941692

BBC News (2020) Should schools be allowed to ban slang words like 'peng'? Retrieved from https://www.bbc.co.uk/news/education-51064279

Brown, M. C., Sibley, D. E., Washington, J. A., Rogers, T. T., Edwards, J. R., MacDonald, M. C., & Seidenberg, M. S. (2015). Impact of dialect use on a basic component of learning to read. *Frontiers in Psychology*, *6*, 1–17. doi:10.3389/fpsyg.2015.00196

Bühler, J. C., Oertzen, T. von, McBride, C. A., Stoll, S., & Maurer, U. (2018). Influence of dialect use on early reading and spelling acquisition in German-speaking children in Grade 1. *Journal of Cognitive Psychology*, *30*(3), 336–360. doi:10.1080/20445911.2018.1444614

Bürkner, P.-C. (2017). brms: 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: https://doi.org/10.32614/RJ-2018-017

Caravolas, M., Hulme, C., & Snowling, M. J. (2001). The foundations of spelling ability: Evidence from a 3- year longitudinal study. *Journal of Memory and Language*, *45*(4), 751–774. doi: https://doi.org/10.1006/jmla.2000.2785

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. doi:10.1177/1529100618772271

Chang, Y. N., Taylor, J. S., Rastle, K., & Monaghan, P. (2020). The relationships between oral language and reading instruction: Evidence from a computational

model of reading. *Cognitive Psychology*, *123*, 101336. doi: https://doi.org/10.1016/j.cogpsych.2020.101336

Changizi, M. A., & Shimojo, S. (2005). Character complexity and redundancy in writing systems over human history. *Proceedings of the Royal Society B*, *272*, 267–75. doi:10.1098/rspb.2004.2942

Charity, A. H., Scarborough, H. S., & Griffin, D. M. (2004). Familiarity with school English in African American children and its relation to early reading achievement. *Child Development*, *75*(5), 1340–1356. doi: https://doi.org/10.1111/j.1467-8624.2004.00744.x

Costa, J. (2015). Can Schools Dispense with Standard Language? Some Unintended Consequences of Introducing Scots in a Scottish Primary School. *Journal of Linguistic Anthropology*, *25*(1), 25-42. doi: https://doi.org/10.1111/jola.12069

Crystal, D. (2003). *The Cambridge Encyclopedia of the English language* (2nd ed.). Cambridge, UK: Cambridge University Press.

Donaldson, J. (1999). *The Gruffalo*. Pan Macmillan.

Donaldson, J. (2005). *The Gruffalo's Child*. Pan Macmillan.

Dossey, E., Clopper, C. G., & Wagner, L. (2020). The Development of Sociolinguistic Competence across the Lifespan: Three Domains of Regional Dialect Perception. *Language Learning and Development*, 1-21. doi: https://doi.org/10.1080/15475441.2020.1784736

Education Scotland (2015) Scottish Curriculum for Excellence Briefing: Scots Language Retrieved from https://education.gov.scot/Documents/cfe-briefing-17.pdf

Education Scotland (2017) *Scots Language in Curriculum for Excellence: enhancing skills in literacy, developing successful learners and confident individuals.* Retrieved from https://education.gov.scot/improvement/Documents/ScotsLanguageinCfEAug17.pdf

Fox, J., Venables, B., Damico, A., & Salverda, A. P. (2020). *English: Translate integers into english*. Retrieved from https://CRAN.R-project.org/package=english

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *Irr: Various coefficients of interrater reliability and agreement*. Retrieved from https://CRAN.R-project.org/package=irr

Gatlin, B., & Wanzek, J. (2015). Relations among children's use of dialect and literacy skills: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *58*, 1306–1318. doi:10.1044/2015

Goldenberg, C., Rueda, R. S., & August, D. (2007). Sociocultural contexts and literacy development. *D. August & T. Shanahan, Developing Reading and writing in second-language learners: Lessons from the Report of the National Literacy Panel on Language-Minority Children and Youth*, 95-130.

Gottlob, L. R., Goldinger, S. D., Stone, G. O., & Van Orden, G. C. (1999). Reading homographs: Orthographic, phonologic, and semantic dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(2), 561–574. doi:10.1037/00961523.25.2.561

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*(2), 67-81. doi: https://doi.org/10.1017/S1366728998000133

Harley, H. (2006). *English words: A linguistic introduction*. Oxford, UK: Blackwell Publishing.

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. doi: https://doi.org/10.1037/0033-295X.111.3.662

Henry, L., & Wickham, H. (2020). *Rlang: Functions for base types and core r and 'tidyverse' features*. Retrieved from https://CRAN.R-project.org/package=rlang

Jared, D., Cormier, P., Levy, B. A., & Wade-Woolley, L. (2012). Cross-language activation of phonology in young bilingual readers. *Reading and Writing*, *25*(6), 1327–1343. doi:10.1007/s11145-011-9320-0

Johnson, L., Terry, N. P., Connor, C. M., & Thomas-Tate, S. (2017). The effects of dialect awareness instruction on nonmainstream American English speakers. *Reading and Writing, 30*(9), 2009-2038. doi: https://doi.org/10.1007/s11145-017-9764-y

Kaushanskaya, M., Yoo, J., & Van Hecke, S. (2013). Word learning in adults with second-language experience: Effects of phonological and referent familiarity. *Journal of Speech, Language, and Hearing Research*. doi: https://doi.org/10.1044/1092-4388(2012/11-0084)

Kay, M. (2020). *tidybayes: Tidy data and geoms for Bayesian models*. doi:10.5281/zenodo.1308151

Kempe, V., Brooks, P. J., & Christman, S. D. (2009). Inconsistent handedness is linked to more successful foreign language vocabulary learning. *Psychonomic Bulletin & Review*, *16*(3), 480-485. doi: https://doi.org/10.3758/PBR.16.3.480

Koo, T. K., & Li, M. Y. (2016). A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163. doi:10.1016/j.jcm.2016.02.012

Kruschke, J. (2014). *Doing Bayesian Data Analysis: A tutorial with R, JAGS, and Stan*. New York, NY: Academic Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi:10.18637/jss.v082.i13

Labov, W. (1995). Can reading failure be reversed? A linguistic approach to the question. In V. L. Gadsden & D. A. Wagner (Eds.), *Literacy among African-American youth: Issues in learning, teaching, and schooling* (pp. 39–68). Cresskill, NJ: Hampton Press.

Leaverton, L. (1973). Dialectal readers: Rationale, use and value. In J. L. Laffey & Roger. Shuy (Eds.), *Language differences: Do they interfere*? (pp. 114-126). Newark, Delaware: International Reading Association.

Makowski, D., Ben-Shachar, M. S., Chen, S., & Lüdecke, D. (2019a). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, *10*, 2767. doi: https://doi.org/10.3389/fpsyg.2019.02767

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019b). BayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. doi:10.21105/joss.01541

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLoS ONE*, *7*(8), e43230. doi:10.1371/journal.pone.0043230

McCloskey, M., and Cohen, N. J. (1989). "Catastrophic interference in connectionist networks: the sequential learning problem," in *The Psychology of Learning and Motivation Vol. 24*, ed G. H. Bower (San Diego, CA: Academic Press), 109–164.

McCullough, E. A., Clopper, C. G., & Wagner, L. (2019). Regional dialect perception across the lifespan: Identification and discrimination. *Language and Speech*, *62*(1), 115-136. doi: https://doi.org/10.1177/0023830917743277

McWhorter, J. H. (1997). Wasting energy on an illusion: six months later. *The Black Scholar*, *27*(2), 2-5.

Milde, B. (2011). Shapecatcher: Unicode character recognition. Retrieved from http://shapecatcher.com/

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*(2), 61–64.

Müller, K. (2017). *Here: A simpler way to find your files*. Retrieved from https://CRAN.R-project.org/package=here

Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609-1623. doi: https://doi.org/10.1016/j.csda.2011.10.005

Ouellette, G. P., & Sénéchal, M. (2008). A window into early literacy: Exploring the cognitive and linguistic underpinnings of invented spelling. *Scientific Studies of Reading*, *12*(2), 195–219. doi:10.1080/10888430801917324

Ouellette, G., & Sénéchal, M. (2017). Invented spelling in kindergarten as a predictor of reading and spelling in grade 1: A new pathway to literacy, or just the same road, less known? *Developmental Psychology*, *53*(1), 77–88. doi:10.1037/dev0000179

Ouellette, G. P., Sénéchal, M., & August, J. (2008). Path- ways to Literacy: A Study of Invented Spelling Its Role in Learning to Read and. *Child Development*, *79*(4), 899–913. doi: https://doi.org/10.1111/j.1467-8624.2008.01166.x

Pedersen, T. L. (2019). *Ggforce: Accelerating 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggforce

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rickford, J. R., & Rickford, A. E. (1995). Dialect readers revisited. *Linguistics and Education*, *7*(2), 107-128.

Rickford, J. R., Sweetland, J., & Rickford, A. E. (2004). African American English and other vernaculars in education: A topic-coded bibliography. *Journal of English Linguistics*, *32*(3), 230-320. doi: https://doi.org/10.1177/0075424204268226

Rodd, J. (2019). How to maintain data quality when you can't see your participants. *APS Observer*, *32*(3).

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217–236. doi:10.1068/p5117

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, *55*(3), 182–186. doi: https://doi.org/10.1198/000313001317097960

Seymour, P. H. (1997). Foundations of orthographic development. *Learning to spell: Research, theory, and practice across languages*, 319-337.

Simpkins, G. A. & Simpkins, C. (1981). Cross cultural approach to curriculum development. In G. Smitherman (Ed.), *Detroit: Center for Black Studies, Wayne State University* (pp. 221-40).

Slowikowski, K. (2020). *Ggrepel: Automatically position non-overlapping text labels with 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggrepel

Stoller, P. (Ed.) (1975). *Black American English: Its background and its usage in schools and in literature*. New York City: Delta.

Taylor, J. S. H., Duff, F. J., Woollams, A. M., Monaghan, P., & Ricketts, J. (2015). How word meaning influences word reading. *Current Directions in Psychological Science*, *24*(4), 322–328.

Terry, N. P., & Scarborough, H. S. (2011). The phonological hypothesis as a valuable framework for studying the relation of dialect variation to early reading skills. In S. A. Brady, D. Braze, & C. A. Fowler (Eds.), *Explaining individual differences in reading: Theory and Evidence* (pp. 97–117). New York, NY: Taylor & Francis.

Terry, N. P., Connor, C. M., Johnson, L., Stuckey, A., & Tani, N. (2016). Dialect variation, dialect-shifting, and reading comprehension in second grade. *Reading and Writing*, *29*(2), 267-295. doi: https://doi.org/10.1007/s11145-015-9593-9

Van Steensel, R. (2006). Relations between socio-cultural factors, the home literacy environment and children's literacy development in the first years of primary education. *Journal of Research in Reading*, *29*(4), 367-382. doi: https://doi.org/10.1111/j.1467-9817.2006.00301.x

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. doi:10.1016/j.jml.2018.07.004

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020, May 29). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. doi: arXiv:1903.08008v3

Washington, J. A., Branum-Martin, L., Sun, C., & Lee-James, R. (2018). The impact of dialect density on the growth of language and reading in African American children. *Language, Speech, and Hearing Services in Schools*, *49*(2), 232-247. doi: https://doi.org/10.1044/2018_LSHSS-17-0063

Wickham, H. (2020). *Modelr: Modelling functions that work with the pipe*. Retrieved from https://CRAN.R-project.org/package=modelr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: https://10.21105/joss.01686

Wilke, C. O. (2020). *Ggridges: Ridgeline plots in 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggridges

Williams, G. P., Panayotov, N. & Kempe, V. (2020). How Does Dialect Exposure Affect Learning to Read and Spell? An Artificial Orthography Study. *Journal of Experimental Psychology: General* (Advance online publication.) doi: https://doi.org/10.1037/xge0000778

Yiakoumetti, A. (2006). A bidialectal programme for the learning of Standard Modern Greek in Cyprus. *Applied Linguistics*, *27*(2), 295-317. doi: https://doi.org/10.1093/applin/aml012

Zhu, H. (2019). *KableExtra: Construct complex table with 'kable' and pipe syntax*. Retrieved from https://CRAN.R-project.org/package=kableExtra

Appendix A: Graphemes used to render phonemes in all experiments

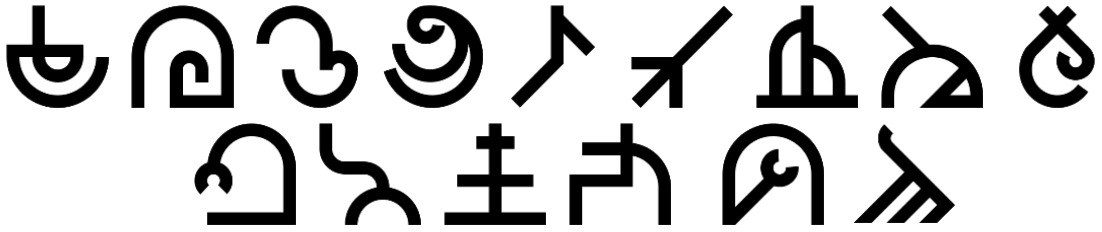

Figure A1: Invented graphemes used to represent each phoneme in all experiments. Note: The last grapheme was created as a spare but not used in this experiment. To prevent participants from memorising the novel graphemes based on resemblance to known graphemes we controlled for similarity to characters of extant writing systems by comparing each invented grapheme against the database of 11,817 characters (excluding Chinese, Korean, and Japanese) on the Shapecatcher website (Milde, 2011). If visual inspection indicated a resemblance, we modified the grapheme to minimise that resemblance.

Appendix B: Spelling and pronunciation of experimental words

We used a string generation algorithm from Williams et al. (2020), accessible at https://osf.io/5mtdj/, to produce all possible permutations of phoneme combinations for each template of syllables. We then selected seven strings from each template ensuring a similar distribution of phonemes across strings. This selection of strings was also applied to have an equal amount of high and low English phonological neighbourhood densities according to the Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities (Marian et al., 2012) database based on the total neighbor metric (i.e. including substitutions, additions, and deletions). The final selection of strings with a mean neighbourhood density of 2.88 was the filtered such that each word differed from all others by a length-normalised Levenshtein edit distance of at least 0.5, with an overall mean of 0.86. This ensured that the items in the study were sufficiently different from one another to limit confusability and to support learning of grapheme-phoneme correspondences (Apfelbaum et al., 2013). The final list of items is displayed in table B1.

Table B1. Spellings and pronunciations for untrained and trained (contrastive and noncontrastive) words in the standard variety (i.e. No Dialect) and the corresponding dialect variants used in the Dialect, Dialect & Social and Dialect Literacy conditions.

Word Type	Standard Variety		Dialect Variants	
	Spelling	Pronunciation	Spelling	Pronunciation
Untrained	MAB	/mab/		
	SKUB	/skub/		
	KLEB	/klɛb/		
	DOLK	/dɔlk/		
	SULD	/suld/		
	DIKLA	/dikla/		
	LUSKO	/luskɔ/		
	KLUFE	/klufɛ/		
	KLODA	/klɔda/		
	SKONEF	/skɔnɛf		
	KLUSIM	/klusim/		
	FLABUN	/flabun/		
Contrastive	NESK	/nɛsk/	NISX	/nisx/
	SKEFI	/skɛfi/	SXIFI	/sxifi/
	BNEKUS	/blɛkus/	BNIXUS	/blixus/
	FLESOD	/flɛsɔd/	FLISO	/flisɔ/
	NEF	/nɛf/	NIF	/nif/
	BESMI	/bɛsmi/	BISMI	/bismi/
	NAL	/nal/	NOL	/nɔl/
	DAF	/daf/	DOF	/dɔf/
	BNAF	/blaf/	BNOF	/blɔf/
	BALF	/balf/	BOLF	/bɔlf/
	DASMU	/dasmu/	DOSMU	/dɔsmu/
	SMADU	/smadu/	SMODU	/smɔdu/
	KUBNE	/kublɛ/	XUBNE	/xublɛ/
	FNOKU	/slɔku/	FNOXU	/slɔxu/
	FNID	/snid/	FNI	/sni/
Noncontrastive	FUB	/fub/		
	MIF	/mif/		
	LOM	/lɔm/		
	FNOF	/snɔf/		
	BNIM	/blim/		
	FLOB	/flɔb/		
	MOLS	/mɔls/		
	FONS	/fɔns/		
	NIFS	/nifs/		
	NOFLE	/nɔflɛ/		
	DEFNA	/dɛsna/		
	SMIBA	/smiba/		
	FLIDU	/flidu/		
	FNIBOL	/snibɔl/		
	FNINAB	/slinab/		

Note. Spellings are represented by Latin characters in the table but were mapped onto the corresponding artificial graphemes in the experiment.

Appendix C: Image Norms

Images used from the (Rossion & Pourtois, 2004) colourised Vanderwart picture set:

- Body part: Finger, foot, eye, hand, nose, arm, ear.
- Furniture and kitchen utensils: Chair, glass, bed, fork, spoon, pot, desk.
- Household objects, tools, and instruments: Television, toothbrush, book, pen, refrigerator, watch, pencil.
- Food and clothing: Pants, socks, shirt, sweater, apple, tomato, potato.
- Buildings, building features, and vehicles: Door, house, window, car, doorknob, truck, bicycle. Animals and plants: Tree, dog, cat, flower, rabbit, duck, chicken.

The subset of pictures and their associated norms are provided in the supplemental material to Williams, Panayotov, & Kempe (2020) at: https://osf.io/5mtdj/.

Appendix D: Gruffalo Index of Language Variation

Because the present study focused on phonological dialect variation it was necessary to obtain an ecologically valid quantitative estimate of the amount of phonological in a naturally occurring dialect which, like our miniature artificial language, also adheres to English phonotactics. To our knowledge, to date no parallel corpora for standard and dialect variants exist. We therefore created such a corpus from the children's books The Gruffalo and The Gruffalo's Child (Donaldson 1999; Donaldson, 2005) in their translations into various Scots dialects. This approach treats the translators of these books as native dialect informants with the benefit of providing estimates of linguistic content that is appropriate for the age group at which literacy is acquired. The books included in this corpus analysis are listed below.

The Gruffalo

- The Doric Gruffalo (translated by Sheena Blackhall)
- Thi Dundee Gruffalo (translated by Matthew Fitt)
- The Glasgow Gruffalo (translated by Elaine C. Smith)
- The Gruffalo in Scots (translated by James Robertson)

The Gruffalo's Child

- The Doric Gruffalo's Bairn (translated by Sheena Blackhall)
- Thi Dundee Gruffalo's Bairn (translated by Matthew Fitt)
- The Gruffalo's Wean (Scots; translated by James Robertson)

[Note: The Gruffalo's Child was not available in Glaswegian at the time of this corpus analysis.]

This corpus consisted of 310 word types. In each book, the Scots translations were aligned with the Standard British English equivalent and coded for no change (e.g. *snake – snake*), lexical change (e.g. *child – bairn*), or phonological change (e.g. *foot – fit*). Words which could not be categorised (N = 26) and words in Scots that arose from paraphrasing the Standard British English phrases were excluded from our analyses. Mean length-normalised Levenshtein edit distances between the Scots and Standard British English translations were calculated to confirm the assignment of categories. As expected, on average phonological changes were associated with smaller nLEDs (M = .40) than lexical changes (M = .80). Overall, we found 92.23% of word types and 53.01% of word tokens to have a dialect variant. In these cases, phonological variation was involved in 49.48% of types and 63.94% of tokens. The most frequent phonological variation involved phoneme substitution (e.g. *bright – bricht*; 79.91% of all tokens) and phoneme drops (e.g. *and – an*; 24.87% of tokens). We thus modelled our dialect variation to simulate Scots phonological changes in which phonemes were substituted or dropped from words.

Appendix E: Model Priors and Posterior Predictive Checks

Here, priors are described first by their expected distribution, and the parameters that define that distribution. For example, a prior of $N(0,1)$ describes a normal distribution with a mean of 0 and a standard deviation of 1. Similarly, a prior of logistic(0,1) describes a logistic distribution with a mean of 0 and a standard deviation of 1. (By default, brms restricts priors on the *SD* to be positive.) The priors used in the vocabulary test model are listed in Table E1.

Table E1. Priors used in the vocabulary test model.

Term	Intercept	Slope	*SD*	*SD* by participant number	*SD* by item
μ	$N(0,5)$	$N(0,0.5)$	$N(0,1)$	$N(0,1)$	$N(0,1)$
ϕ	$N(0,3)$	$N(0,1)$	$N(0,1)$	$N(0,5)$	$N(0,5)$
α	logistic(0,1)	$N(0,5)$	$N(0,5)$	$N(0,10)$	$N(0,10)$
γ	logistic(0,1)	$N(0,0.5)$	$N(0,5)$	$N(0,10)$	$N(0,10)$

Weakly informative regularising priors were used for all terms. All priors were centred on 0, with standard deviations ranging from 0.5 to 10, thus allowing for a range of values with less prior probability places on extreme responses. These priors allow the posterior to be determined primarily by the data. For the slope terms, the priors assume no effect to small effects for each parameter in either direction. Weakly informative regularising priors were also used for all standard deviation terms. Finally, an $LKJ(2)$ prior was used for the correlation between terms, which acts to down-weight perfect correlations (Vasishth et al., 2018). These priors are in some cases more informative than those initially planned in our pre-registration (available at https://osf.io/bxt87, which used very weakly informative priors) to improve model fit (i.e. accounting for divergences

during fitting). For example, the μ intercept and slope, and γ slope have standard deviations half as large as planned, while the standard deviation for the ϕ intercept is three times as large as initially planned.

During fitting 8000 iterations were used instead of the planned 1000 and 6 were used rather than 4 chains to improve estimates in response to warnings about bulk and tail effective sample size, totalling 48,000 samples rather than 4000. The priors for the literacy test models are outlined in Table E2.

Table E2. Priors used in the literacy test models.

Term	Intercept	Slope	SD	SD by participant number	SD by item
μ	*N*(0,5)	*N*(0,1)	*N*(0,1)	*N*(0,1)	*N*(0,1)
ϕ	*N*(0,3)	*N*(0,1)	*N*(0,1)	*N*(0,5)	*N*(0,5)
α	logistic(0,1)	*N*(0,5)	*N*(0,5)	*N*(0,10)	*N*(0,10)
γ	logistic(0,1)	*N*(0,1)	*N*(0,5)	*N*(0,10)	*N*(0,10)

Due to having more observations for analyses during the literacy test, both the μ and γ slope terms used more weakly informative priors than the vocabulary test model. This allowed the data to have a larger impact on parameter estimates while having no impact on model convergence. As with the vocabulary test model, an *LKJ*(2) prior was used for the correlation between terms.

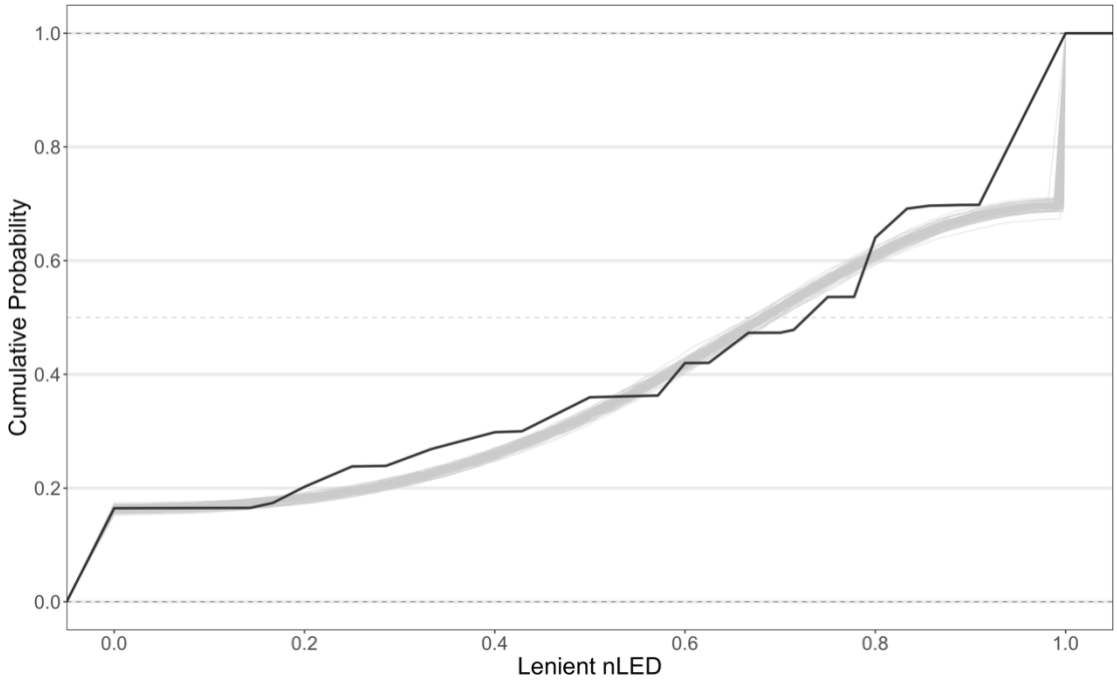

Figure E1: Posterior predictive check for the model fitted to the vocabulary test data. Lines indicate the empirical cumulative distribution function of the observations (black) and samples from the posterior (grey).

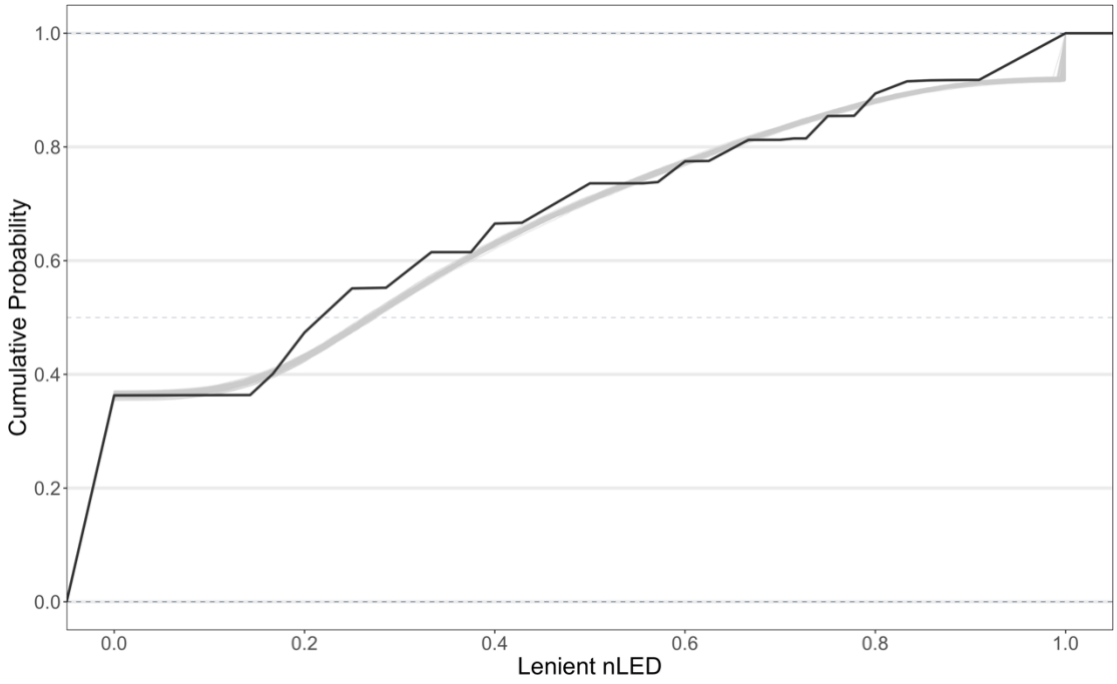

Figure E2: Posterior predictive check for the model fitted to the literacy test data. Lines indicate the empirical cumulative distribution function of the observations (black) and samples from the posterior (grey).

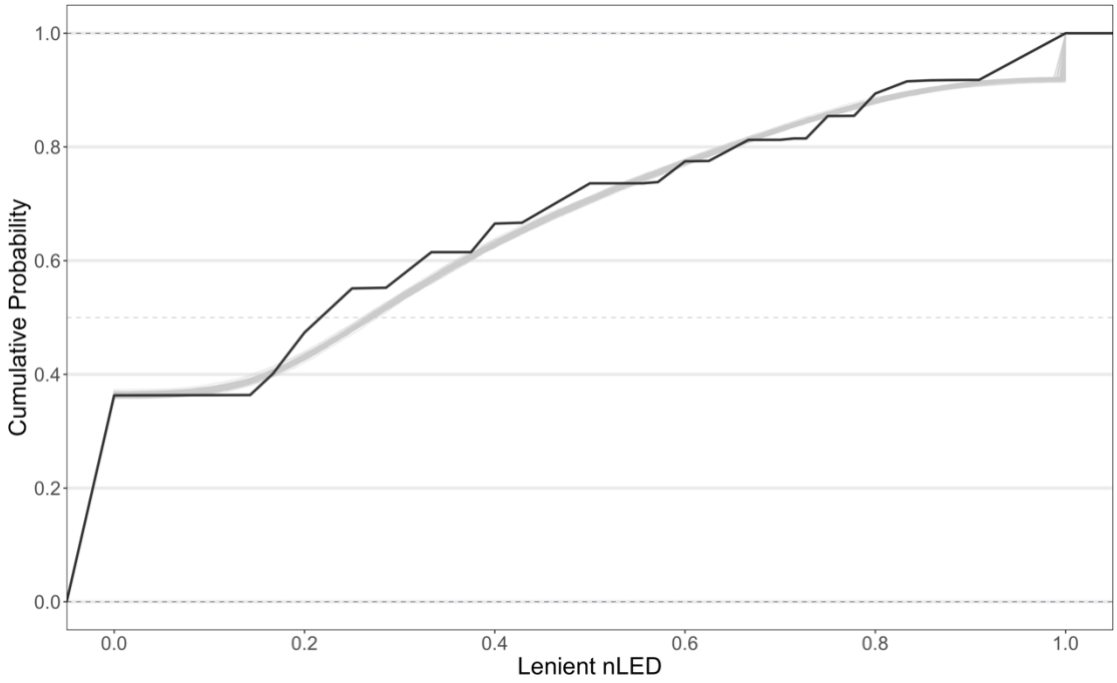

Figure E3: Posterior predictive check for the model fitted to the literacy test data with vocabulary test performance as a covariate. Lines indicate the empirical cumulative distribution function of the observations (black) and samples from the posterior (grey). As can be seen from the plots, the posterior predictive checks indicate a generally good model fit in all instances, such that the model largely captures the shape of the data.

Appendix F: Fitted Model Summaries

In the tables of population level (fixed) effects, $\hat{R}$ is a measure of convergence for within- and between-chain estimates, with values closer to 1 being preferable. The bulk and tail effective sample sizes give diagnostics of the number of samples which contain the same amount of information as the dependent sample (Vehtari et al., 2020), with higher values being preferable. The tail effective sample size is determined at the 5% and 95% quantiles, while the bulk is determined at values in between these quantiles.

Vocabulary Test Model

A summary of the population-level (fixed) effects for the vocabulary test model is provided below. This can be used to determine model diagnostics, coefficients, and estimates around these coefficients using 95% credible intervals. Note that the model was fitted with the above priors and sum coded effects of Exposure condition, Word Type, and Task. As a result, the intercept represents the grand mean (i.e. the mean of the means of the dependent variable at each level of the categorical variables). The regression coefficients then represent the difference between the grand mean and the following three exposure conditions: No Dialect, Dialect, Dialect & Social. To obtain parameter estimates for the No Dialect, Dialect, and Dialect & Social conditions their regression coefficients need to be added to the intercept. To obtain regression estimates for the Dialect Literacy condition, all three Exposure condition coefficients must be subtracted from the intercept. To obtain parameter estimates for contrastive words, the coefficient needs to be added to the intercept. To obtain the estimates for non-contrastive words, the coefficient needs to be subtracted from the intercept. To obtain parameter

estimates for the Reading task, the coefficient needs to be added to the intercept. To obtain the estimates for the spelling task, the coefficient needs to be subtracted from the intercept.

Table F1: Parameter estimates, standard errors, 95% credible intervals, and model diagnostics for the model fitted to nLEDs from the vocabulary test.

	Parameter	*Est.*	*SE*	95% CI	*Bulk ESS*	*Tail ESS*
μ						
	Intercept	0.34	0.03	[0.28, 0.40]	7824	12363
	No Dialect	0.03	0.04	[-0.05, 0.10]	6736	11607
	Dialect	-0.04	0.04	[-0.12, 0.03]	7308	12508
	Dialect & Social	0	0.04	[-0.07, 0.07]	6644	11987
	Word Type	0.03	0.02	[-0.01, 0.07]	10320	15405
	No Dialect × Word Type	0	0.02	[-0.04, 0.04]	30108	19005
	Dialect × Word Type	0	0.02	[-0.04, 0.04]	28808	19137
	Dialect & Social × Word Type	0	0.02	[-0.04, 0.04]	31508	19390
ϕ						
	Intercept	1.82	0.03	[1.76, 1.88]	5350	13365
	No Dialect	-0.01	0.04	[-0.09, 0.07]	12336	18285
	Dialect	-0.06	0.04	[-0.14, 0.03]	12602	16852
	Dialect & Social	0.02	0.04	[-0.06, 0.09]	13794	17082
	Word Type	0	0.03	[-0.05, 0.06]	10054	17796
	No Dialect × Word Type	-0.03	0.04	[-0.10, 0.05]	29030	19408
	Dialect × Word Type	0.01	0.04	[-0.06, 0.08]	29955	19532
	Dialect & Social × Word Type	0.04	0.04	[-0.03, 0.11]	31022	20460
α						
	Intercept	-0.14	0.1	[-0.34, 0.05]	15525	17181
	No Dialect	0.09	0.09	[-0.08, 0.26]	13530	17076
	Dialect	0.05	0.08	[-0.12, 0.21]	13206	16313
	Dialect & Social	-0.1	0.08	[-0.26, 0.07]	13104	15637
	Word Type	-0.08	0.09	[-0.25, 0.09]	14507	15504
	No Dialect × Word Type	-0.01	0.05	[-0.11, 0.09]	25967	18558
	Dialect × Word Type	0	0.04	[-0.09, 0.08]	31103	19435
	Dialect & Social × Word Type	0.05	0.04	[-0.04, 0.14]	29087	18243
γ						
	Intercept	1.15	0.2	[0.76, 1.55]	11742	16173
	No Dialect	0.19	0.22	[-0.23, 0.61]	11829	16574
	Dialect	-0.26	0.19	[-0.64, 0.12]	9681	14910
	Dialect & Social	0.06	0.2	[-0.32, 0.45]	9621	14270
	Word Type	0.22	0.16	[-0.10, 0.53]	15056	16693
	No Dialect × Word Type	-0.05	0.14	[-0.33, 0.23]	21307	18614
	Dialect × Word Type	-0.07	0.09	[-0.24, 0.09]	31341	17516
	Dialect & Social × Word Type	0.01	0.09	[-0.18, 0.20]	25910	19286

Note. *All* $\hat{R} = 1$.

Literacy Test Model

A summary of the literacy test model is provided below. This can be used to determine model diagnostics and coefficients.

The same coding was used here as in the vocabulary test model, with the exception that Word Type has three levels in this phase of the experiment. Words can be contrastive, non-contrastive, or untrained. Thus, Word Type was Helmert coded in R (R defaults to what is traditionally called reverse Helmert coding) such that the first estimate (Word Type in the below table) represents half the difference in scores between trained non-contrastive words and trained contrastive words. The second estimate (Word Familiarity in the below table) represents the difference in scores between the mean of the trained (non-contrastive and contrastive words) and untrained words and the mean of the trained words. Thus, it estimates the effect of words being untrained vs. trained.

As before the intercept represents the grand mean. To obtain parameter estimates for contrastive words, the parameter estimate for Word Type and the parameter estimate for Word Familiarity need to be subtracted from the intercept. To obtain parameter estimates for non-contrastive words, the parameter estimate for Word Type needs to be added, and the parameter estimate for Word Familiarity needs to be subtracted from the intercept. To obtain parameter estimates for unfamiliar words, both parameter estimates for Word Type and Word Familiarity need to be added to the intercept.

Table F2: Parameter estimates, standard errors, 95% credible intervals, and model diagnostics for the model fitted to nLEDs from the literacy test.

	Parameter	*Est.*	*SE*	95% CI	*Bulk ESS*	*Tail ESS*
μ						
	Intercept*	-0.39	0.04	[-0.47, -.32]	872	2362
	Task	-0.06	0.01	[-0.08, -.04]	4293	10471
	No Dialect*	-0.03	0.06	[-0.14, 0.08]	760	1537
	Dialect*	-0.03	0.06	[-0.13, 0.08]	976	1969
	Dialect & Social*	0.03	0.05	[-0.08, 0.13]	882	1478
	Word Type	0	0.02	[-0.05, 0.04]	3844	7230
	Word Familiarity	0	0.01	[-0.03, 0.03]	3580	7359
	Task × No Dialect	0.01	0.02	[-0.02, 0.04]	4333	9246
	Task × Dialect	0.02	0.02	[-0.01, 0.05]	5496	11559
	Task × Dialect & Social	-0.01	0.02	[-0.04, 0.02]	4634	11692
	Task × Word Type	-0.01	0.01	[-0.02, 0.01]	27098	16538
	Task × Word Familiarity	0.01	0	[0.00, 0.02]	23157	17466
	No Dialect × Word Type	0.01	0.01	[-0.01, 0.03]	20350	18569
	Dialect × Word Type	0.02	0.01	[-0.00, 0.04]	19552	16539
	Dialect & Social × Word Type	-0.01	0.01	[-0.03, 0.02]	21319	19307
	No Dialect × Word Familiarity	-0.01	0.01	[-0.02, 0.01]	2782	7124
	Dialect × Word Familiarity	0	0.01	[-0.02, 0.02]	3279	8546
	Dialect & Social × Word Familiarity	0	0.01	[-0.02, 0.02]	3022	7788
	Task × No Dialect × Word Type	0.01	0.01	[-0.02, 0.03]	22200	18142
	Task × Dialect × Word Type	0.01	0.01	[-0.02, 0.03]	20607	18319
	Task × Dialect & Social × Word Type	-0.01	0.01	[-0.03, 0.01]	23328	18455
	Task × No Dialect × Word Familiarity	-0.01	0.01	[-0.02, 0.01]	22972	18686
	Task × Dialect × Word Familiarity	0	0.01	[-0.02, 0.01]	21835	19248
	Task × Dialect & Social × Word Familiarity	0	0.01	[-0.01, 0.01]	21982	19288
ϕ						
	Intercept	2.64	0.04	[2.56, 2.73]	2173	5803
	Task	-0.18	0.02	[-0.22, -.15]	10289	17192
	No Dialect	0.07	0.06	[-0.05, 0.18]	2015	3984
	Dialect	-0.03	0.06	[-0.15, 0.08]	2006	4967
	Dialect & Social	0.01	0.06	[-0.10, 0.13]	2415	5978
	Word Type	0	0.03	[-0.07, 0.06]	8299	12907
	Word Familiarity	0.09	0.02	[0.05, 0.14]	7498	12119
	Task × No Dialect	0.02	0.03	[-0.04, 0.08]	9083	13062
	Task × Dialect	0.01	0.03	[-0.05, 0.07]	8361	14215
	Task × Dialect & Social	-0.04	0.03	[-0.10, 0.02]	7657	13387
	Task × Word Type	-0.02	0.01	[-0.04, 0.01]	24341	18144
	Task × Word Familiarity	0.01	0.01	[-0.01, 0.03]	20434	16978
	No Dialect × Word Type	0.02	0.03	[-0.03, 0.07]	19798	17214
	Dialect × Word Type	-0.02	0.03	[-0.07, 0.03]	20696	17910
	Dialect & Social × Word Type	-0.02	0.03	[-0.08, 0.03]	17569	17857
	No Dialect × Word Familiarity	0.02	0.02	[-0.02, 0.06]	12286	16528
	Dialect × Word Familiarity	-0.03	0.02	[-0.06, 0.01]	12205	16205
	Dialect & Social × Word Familiarity	0	0.02	[-0.04, 0.04]	12006	14273
	Task × No Dialect × Word Type	0.05	0.02	[-0.00, 0.09]	22267	18108
	Task × Dialect × Word Type	0	0.02	[-0.05, 0.05]	21631	18927
	Task × Dialect & Social × Word Type	0	0.02	[-0.05, 0.05]	22569	18027
	Task × No Dialect × Word Familiarity	0	0.02	[-0.03, 0.03]	22592	18688
	Task × Dialect × Word Familiarity	0	0.02	[-0.03, 0.03]	19848	17845
	Task × Dialect & Social × Word Familiarity	-0.01	0.02	[-0.04, 0.02]	21878	18709

α					
Intercept	-0.32	0.13	[-0.57, -.07]	4357	9644
Task	0.28	0.03	[0.22, 0.33]	5887	12534
No Dialect	0.03	0.11	[-0.19, 0.25]	2390	4549
Dialect	0.03	0.11	[-0.19, 0.25]	2437	6047
Dialect & Social	0.01	0.11	[-0.21, 0.23]	3048	6900
Word Type	0.13	0.13	[-0.12, 0.38]	8811	12646
Word Familiarity	-0.02	0.08	[-0.18, 0.14]	8079	13276
Task × No Dialect	0.08	0.05	[-0.02, 0.17]	6314	12914
Task × Dialect	0	0.05	[-0.09, 0.09]	6271	12575
Task × Dialect & Social	0	0.05	[-0.09, 0.09]	6217	10116
Task × Word Type	0.08	0.02	[0.04, 0.11]	33338	18081
Task × Word Familiarity	-0.03	0.01	[-0.05, -.00]	32607	18646
No Dialect × Word Type	-0.07	0.04	[-0.14, 0.00]	15363	16866
Dialect × Word Type	-0.01	0.04	[-0.09, 0.06]	16414	17588
Dialect & Social × Word Type	-0.08	0.04	[-0.16, -.01]	15749	16738
No Dialect × Word Familiarity	-0.05	0.03	[-0.10, 0.01]	10467	14463
Dialect × Word Familiarity	-0.02	0.03	[-0.07, 0.04]	9434	14577
Dialect & Social × Word Familiarity	0.07	0.03	[0.01, 0.12]	9835	15193
Task × No Dialect × Word Type	-0.08	0.03	[-0.14, -.02]	24660	18592
Task × Dialect × Word Type	0.03	0.03	[-0.03, 0.09]	25700	18775
Task × Dialect & Social × Word Type	0.01	0.03	[-0.05, 0.07]	24559	18031
Task × No Dialect × Word Familiarity	0.01	0.02	[-0.02, 0.05]	26206	18406
Task × Dialect × Word Familiarity	-0.03	0.02	[-0.07, 0.01]	26062	18774
Task × Dialect & Social × Word Familiarity	0.02	0.02	[-0.01, 0.06]	26449	18581
γ					
Intercept	-3.87	0.4	[-4.66, -.13]	1395	4218
Task	-0.34	0.22	[-0.74, 0.11]	7653	11945
No Dialect*	-0.5	0.48	[-1.43, 0.45]	998	2146
Dialect	0.07	0.48	[-0.87, 0.99]	1216	2812
Dialect & Social*	0.21	0.47	[-0.71, 1.12]	1131	2431
Word Type	-0.45	0.21	[-0.86, -.03]	12521	15233
Word Familiarity	0.1	0.19	[-0.30, 0.45]	7840	11954
Task × No Dialect	0.45	0.23	[0.01, 0.91]	5693	14185
Task × Dialect	-0.3	0.22	[-0.73, 0.14]	5655	13101
Task × Dialect & Social	0.04	0.23	[-0.41, 0.49]	6168	13407
Task × Word Type	0	0.1	[-0.19, 0.18]	20633	17428
Task × Word Familiarity	0.12	0.09	[-0.05, 0.29]	20958	18994
No Dialect × Word Type	0.16	0.18	[-0.19, 0.50]	16019	16929
Dialect × Word Type	0.21	0.17	[-0.13, 0.55]	15381	17442
Dialect & Social × Word Type	0.01	0.18	[-0.34, 0.37]	15063	16743
No Dialect × Word Familiarity	-0.12	0.18	[-0.48, 0.24]	8137	13815
Dialect × Word Familiarity	0.05	0.17	[-0.29, 0.38]	7207	12831
Dialect & Social × Word Familiarity	0.04	0.19	[-0.33, 0.42]	7974	13449
Task × No Dialect × Word Type	0.17	0.15	[-0.12, 0.47]	20221	18302
Task × Dialect × Word Type	0.02	0.14	[-0.27, 0.30]	18800	18348
Task × Dialect & Social × Word Type	-0.06	0.15	[-0.36, 0.24]	17793	17422
Task × No Dialect × Word Familiarity	0.5	0.15	[0.21, 0.80]	15651	15454
Task × Dialect × Word Familiarity	-0.39	0.13	[-0.66, -.14]	16184	17639
Task × Dialect & Social × Word Familiarity	-0.03	0.15	[-0.33, 0.27]	16146	16892

Note. All $\hat{R}$ = 1 except those denoted by * which are equal to 1.01.

Exploratory Covariate Literacy Test Model

A summary of the literacy test model incorporating mean vocabulary test scores as a covariate is provided below. This can be used to determine model diagnostics and coefficients. This model used the same coding scheme as the literacy test model but included a continuous numerical predictor of mean vocabulary test performance (ranging from 0-1).

Table F3: Parameter estimates, standard errors, 95% credible intervals, and model diagnostics for the model fitted to nLEDs from the literacy test with mean vocabulary test performance as a covariate.

	Parameter	*Est.*	*SE*	95% CI	*Bulk ESS*	*Tail ESS*
μ						
	Intercept	-0.69	0.09	[-0.86, -0.52]	1916	3997
	Mean Voc. Test nLED	0.48	0.12	[0.24, 0.72]	2498	5228
	Task	-0.1	0.03	[-0.16, -0.04]	6745	11373
	No Dialect	-0.08	0.1	[-0.28, 0.12]	2563	5344
	Dialect	-0.06	0.1	[-0.26, 0.14]	2621	5524
	Dialect & Social	-0.03	0.11	[-0.24, 0.18]	3048	6633
	Word Type	0.05	0.04	[-0.03, 0.13]	6909	11300
	Word Familiarity	0.04	0.03	[-0.02, 0.10]	5576	9887
	Mean Voc. Test nLED × Task	0.06	0.05	[-0.03, 0.16]	6841	11173
	Mean Voc. Test nLED × No Dialect	0.04	0.15	[-0.25, 0.33]	2947	6340
	Mean Voc. Test nLED × Dialect	0.09	0.15	[-0.21, 0.39]	3080	6242
	Mean Voc. Test nLED × Dialect & Social	0.09	0.16	[-0.22, 0.39]	3756	6996
	Task × No Dialect	-0.01	0.05	[-0.11, 0.09]	7520	12366
	Task × Dialect	0.13	0.05	[0.04, 0.22]	7490	12189
	Task × Dialect & Social	-0.02	0.05	[-0.12, 0.09]	7289	11955
	Mean Voc. Test nLED × Word Type	-0.07	0.05	[-0.17, 0.03]	8413	13230
	Mean Voc. Test nLED × Word Familiarity	-0.07	0.04	[-0.14, 0.00]	6391	11737
	Task × Word Type	0.03	0.02	[-0.02, 0.07]	13711	16671
	Task × Word Familiarity	0.03	0.01	[-0.00, 0.05]	13588	16217
	No Dialect × Word Type	0.04	0.04	[-0.04, 0.13]	8974	14041
	Dialect × Word Type	0.01	0.04	[-0.06, 0.08]	9970	13728
	Dialect & Social × Word Type	0.06	0.04	[-0.03, 0.14]	10027	15117
	No Dialect × Word Familiarity	-0.01	0.03	[-0.07, 0.05]	6754	12410
	Dialect × Word Familiarity	-0.02	0.03	[-0.08, 0.03]	7171	12105
	Dialect & Social × Word Familiarity	0.03	0.03	[-0.03, 0.09]	7252	13006
	Mean Voc. Test nLED × Task × No Dialect	0.03	0.07	[-0.11, 0.18]	7556	12263
	Mean Voc. Test nLED × Task × Dialect	-0.17	0.07	[-0.31, -0.03]	7498	12616

Mean Voc. Test nLED × Task × Dialect & Social	0.01	0.08	[-0.15, 0.17]	7221	11443
Mean Voc. Test nLED × Task × Word Type	-0.05	0.04	[-0.12, 0.02]	14242	16701
Mean Voc. Test nLED × Task × Word Familiarity	-0.02	0.02	[-0.06, 0.02]	14006	16024
Mean Voc. Test nLED × No Dialect × Word Type	-0.05	0.06	[-0.17, 0.07]	9228	14135
Mean Voc. Test nLED × Dialect × Word Type	0.01	0.06	[-0.10, 0.13]	10637	14973
Mean Voc. Test nLED × Dialect & Social × Word Type	-0.1	0.07	[-0.23, 0.03]	10299	15074
Mean Voc. Test nLED × No Dialect × Word Familiarity	0.01	0.04	[-0.08, 0.09]	7328	12247
Mean Voc. Test nLED × Dialect × Word Familiarity	0.04	0.04	[-0.05, 0.12]	7262	12548
Mean Voc. Test nLED × Dialect & Social × Word Familiarity	-0.05	0.05	[-0.15, 0.04]	7571	13132
Task × No Dialect × Word Type	0.06	0.04	[-0.02, 0.15]	8799	13967
Task × Dialect × Word Type	0.02	0.04	[-0.06, 0.09]	9765	14421
Task × Dialect & Social × Word Type	-0.06	0.04	[-0.15, 0.02]	10159	14892
Task × No Dialect × Word Familiarity	-0.02	0.02	[-0.06, 0.03]	9589	13660
Task × Dialect × Word Familiarity	-0.01	0.02	[-0.06, 0.03]	10556	14908
Task × Dialect & Social × Word Familiarity	-0.01	0.03	[-0.06, 0.04]	10195	15080
Mean Voc. Test nLED × Task × No Dialect × Word Type	-0.09	0.06	[-0.21, 0.04]	9137	13855
Mean Voc. Test nLED × Task × Dialect × Word Type	-0.02	0.06	[-0.13, 0.09]	10359	14624
Mean Voc. Test nLED × Task × Dialect & Social × Word Type	0.08	0.06	[-0.05, 0.20]	10506	15284
Mean Voc. Test nLED × Task × No Dialect × Word Familiarity	0.02	0.03	[-0.05, 0.09]	9598	14866
Mean Voc. Test nLED × Task × Dialect × Word Familiarity	0.02	0.03	[-0.05, 0.08]	10972	15294
Mean Voc. Test nLED × Task × Dialect & Social × Word Familiarity	0.01	0.04	[-0.07, 0.09]	10233	14349
ϕ					
Intercept	2.88	0.11	[2.66, 3.11]	5142	9843
Mean Voc. Test nLED	-0.36	0.16	[-0.68, -0.04]	5370	10626
Task	-0.32	0.07	[-0.45, -0.18]	8696	13997
No Dialect	0.04	0.15	[-0.25, 0.32]	5402	9940
Dialect	0.1	0.14	[-0.18, 0.38]	6107	10795
Dialect & Social	-0.18	0.15	[-0.48, 0.12]	5305	9197
Word Type	-0.02	0.06	[-0.14, 0.11]	12885	15759
Word Familiarity	0.13	0.05	[0.04, 0.23]	9741	14882
Mean Voc. Test nLED × Task	0.21	0.1	[0.00, 0.41]	8274	13721
Mean Voc. Test nLED × No Dialect	0.06	0.22	[-0.36, 0.49]	5342	10110
Mean Voc. Test nLED × Dialect	-0.23	0.21	[-0.65, 0.19]	6385	10913
Mean Voc. Test nLED × Dialect & Social	0.31	0.23	[-0.15, 0.75]	5711	9797
Task × No Dialect	-0.01	0.11	[-0.22, 0.19]	7508	13845
Task × Dialect	-0.2	0.1	[-0.40, 0.00]	8000	13761
Task × Dialect & Social	0.06	0.11	[-0.15, 0.28]	7938	13728
Mean Voc. Test nLED × Word Type	0.01	0.08	[-0.16, 0.17]	15805	17652
Mean Voc. Test nLED × Word Familiarity	-0.05	0.07	[-0.18, 0.07]	10730	14473
Task × Word Type	-0.04	0.06	[-0.15, 0.08]	14295	16702
Task × Word Familiarity	0	0.04	[-0.07, 0.07]	13503	15904
No Dialect × Word Type	0.07	0.09	[-0.11, 0.25]	12546	15671
Dialect × Word Type	-0.05	0.09	[-0.23, 0.13]	11664	14006
Dialect & Social × Word Type	-0.18	0.1	[-0.37, 0.01]	12025	15427

No Dialect × Word Familiarity	0	0.07	[-0.14, 0.13]	8795	14221
Dialect × Word Familiarity	-0.09	0.07	[-0.22, 0.05]	9496	13632
Dialect & Social × Word Familiarity	0	0.08	[-0.15, 0.15]	9006	14427
Mean Voc. Test nLED × Task × No Dialect	0.04	0.15	[-0.26, 0.35]	7454	13363
Mean Voc. Test nLED × Task × Dialect	0.33	0.15	[0.03, 0.63]	8094	14049
Mean Voc. Test nLED × Task × Dialect & Social	-0.16	0.16	[-0.48, 0.15]	7950	13077
Mean Voc. Test nLED × Task × Word Type	0.03	0.08	[-0.14, 0.19]	14613	16737
Mean Voc. Test nLED × Task × Word Familiarity	0.01	0.05	[-0.09, 0.12]	13943	16692
Mean Voc. Test nLED × No Dialect × Word Type	-0.07	0.13	[-0.34, 0.19]	12628	16385
Mean Voc. Test nLED × Dialect × Word Type	0.04	0.13	[-0.22, 0.30]	11930	15432
Mean Voc. Test nLED × Dialect & Social × Word Type	0.23	0.14	[-0.04, 0.51]	12148	15760
Mean Voc. Test nLED × No Dialect × Word Familiarity	0.03	0.1	[-0.17, 0.23]	8819	14203
Mean Voc. Test nLED × Dialect × Word Familiarity	0.1	0.1	[-0.11, 0.30]	9577	14850
Mean Voc. Test nLED × Dialect & Social × Word Familiarity	0	0.11	[-0.21, 0.22]	8931	14479
Task × No Dialect × Word Type	-0.02	0.09	[-0.19, 0.17]	10766	14940
Task × Dialect × Word Type	-0.07	0.09	[-0.24, 0.11]	11973	15099
Task × Dialect & Social × Word Type	0.15	0.1	[-0.03, 0.34]	10897	15422
Task × No Dialect × Word Familiarity	0.03	0.06	[-0.09, 0.15]	11580	15050
Task × Dialect × Word Familiarity	-0.05	0.06	[-0.17, 0.07]	10879	15650
Task × Dialect & Social × Word Familiarity	0.11	0.06	[-0.02, 0.23]	11116	15296
Mean Voc. Test nLED × Task × No Dialect × Word Type	0.09	0.13	[-0.18, 0.34]	10823	15344
Mean Voc. Test nLED × Task × Dialect × Word Type	0.1	0.13	[-0.16, 0.36]	12272	15436
Mean Voc. Test nLED × Task × Dialect & Social × Word Type	-0.24	0.14	[-0.51, 0.04]	11041	15011
Mean Voc. Test nLED × Task × No Dialect × Word Familiarity	-0.04	0.09	[-0.21, 0.13]	11680	15176
Mean Voc. Test nLED × Task × Dialect × Word Familiarity	0.08	0.09	[-0.10, 0.25]	11182	15369
Mean Voc. Test nLED × Task × Dialect & Social × Word Familiarity	-0.18	0.09	[-0.36, 0.01]	11092	15198
α					
Intercept	0.58	0.24	[0.12, 1.04]	6207	11122
Mean Voc. Test nLED	-1.4	0.31	[-2.00, -0.79]	6445	11522
Task	0.48	0.09	[0.29, 0.66]	8022	13257
No Dialect	0.16	0.27	[-0.37, 0.70]	5973	9887
Dialect	0.04	0.26	[-0.48, 0.55]	6309	10686
Dialect & Social	0.12	0.29	[-0.44, 0.68]	5831	8551
Word Type	0.4	0.17	[0.07, 0.73]	10057	13912
Word Familiarity	-0.08	0.11	[-0.30, 0.14]	8832	12986
Mean Voc. Test nLED × Task	-0.32	0.14	[-0.60, -0.03]	8085	13079
Mean Voc. Test nLED × No Dialect	-0.14	0.4	[-0.93, 0.64]	6133	11371
Mean Voc. Test nLED × Dialect	-0.09	0.41	[-0.88, 0.72]	6540	11330
Mean Voc. Test nLED × Dialect & Social	-0.17	0.43	[-1.03, 0.67]	5886	10085
Task × No Dialect	0.23	0.15	[-0.06, 0.52]	7808	13332
Task × Dialect	0.08	0.14	[-0.20, 0.35]	7830	13202
Task × Dialect & Social	-0.17	0.16	[-0.48, 0.13]	8239	12882
Mean Voc. Test nLED × Word Type	-0.44	0.14	[-0.71, -0.17]	13598	16988
Mean Voc. Test nLED × Word Familiarity	0.1	0.1	[-0.10, 0.30]	9506	14408

Task × Word Type	0.21	0.06	[0.08, 0.33]	16890	18011
Task × Word Familiarity	-0.13	0.04	[-0.21, -0.06]	16488	17445
No Dialect × Word Type	-0.17	0.12	[-0.40, 0.06]	12073	15275
Dialect × Word Type	-0.11	0.11	[-0.33, 0.11]	12304	15479
Dialect & Social × Word Type	0.06	0.12	[-0.18, 0.31]	11352	15705
No Dialect × Word Familiarity	-0.07	0.09	[-0.25, 0.11]	8382	13367
Dialect × Word Familiarity	-0.07	0.09	[-0.24, 0.10]	9191	13857
Dialect & Social × Word Familiarity	0.12	0.1	[-0.07, 0.32]	9004	13690
Mean Voc. Test nLED × Task × No Dialect	-0.23	0.22	[-0.66, 0.20]	7853	12887
Mean Voc. Test nLED × Task × Dialect	-0.14	0.22	[-0.57, 0.29]	8249	13657
Mean Voc. Test nLED × Task × Dialect & Social	0.28	0.24	[-0.19, 0.75]	8353	13431
Mean Voc. Test nLED × Task × Word Type	-0.21	0.09	[-0.39, -0.02]	16241	17929
Mean Voc. Test nLED × Task × Word Familiarity	0.16	0.06	[0.05, 0.28]	16659	17691
Mean Voc. Test nLED × No Dialect × Word Type	0.17	0.18	[-0.18, 0.51]	11821	15283
Mean Voc. Test nLED × Dialect × Word Type	0.15	0.17	[-0.19, 0.49]	12112	15809
Mean Voc. Test nLED × Dialect & Social × Word Type	-0.23	0.19	[-0.61, 0.13]	11389	16455
Mean Voc. Test nLED × No Dialect × Word Familiarity	0.04	0.14	[-0.24, 0.31]	8574	14145
Mean Voc. Test nLED × Dialect × Word Familiarity	0.09	0.14	[-0.18, 0.36]	9181	13600
Mean Voc. Test nLED × Dialect & Social × Word Familiarity	-0.09	0.15	[-0.40, 0.20]	8822	13551
Task × No Dialect × Word Type	-0.22	0.1	[-0.43, -0.02]	12942	15620
Task × Dialect × Word Type	0.19	0.1	[-0.00, 0.39]	13698	17378
Task × Dialect & Social × Word Type	-0.03	0.11	[-0.25, 0.18]	12826	16411
Task × No Dialect × Word Familiarity	-0.04	0.07	[-0.17, 0.09]	13930	16447
Task × Dialect × Word Familiarity	0.07	0.06	[-0.06, 0.19]	13637	16390
Task × Dialect & Social × Word Familiarity	-0.14	0.07	[-0.29, -0.00]	12290	15462
Mean Voc. Test nLED × Task × No Dialect × Word Type	0.23	0.16	[-0.07, 0.54]	12904	15165
Mean Voc. Test nLED × Task × Dialect × Word Type	-0.26	0.15	[-0.56, 0.04]	13840	17030
Mean Voc. Test nLED × Task × Dialect & Social × Word Type	0.07	0.17	[-0.26, 0.41]	12847	16596
Mean Voc. Test nLED × Task × No Dialect × Word Familiarity	0.09	0.1	[-0.11, 0.28]	13951	15637
Mean Voc. Test nLED × Task × Dialect × Word Familiarity	-0.15	0.1	[-0.34, 0.04]	13658	16651
Mean Voc. Test nLED × Task × Dialect & Social × Word Familiarity	0.27	0.11	[0.05, 0.48]	12338	15805
γ					
Intercept	-5.89	0.64	[-7.16, -4.66]	2657	6409
Mean Voc. Test nLED	3.15	0.84	[1.49, 4.81]	4206	9626
Task	-0.51	0.39	[-1.28, 0.27]	10075	14987
No Dialect	-0.24	0.61	[-1.44, 0.94]	3262	7838
Dialect	0.05	0.61	[-1.14, 1.24]	3262	7675
Dialect & Social	0.13	0.62	[-1.10, 1.33]	3126	7861
Word Type	-0.42	0.36	[-1.13, 0.30]	14415	16572
Word Familiarity	0.37	0.32	[-0.28, 0.98]	10289	15030
Mean Voc. Test nLED × Task	0.25	0.54	[-0.81, 1.32]	12258	15498
Mean Voc. Test nLED × No Dialect	-0.66	0.8	[-2.22, 0.90]	5961	11940
Mean Voc. Test nLED × Dialect	0.37	0.81	[-1.23, 1.97]	5441	11479
Mean Voc. Test nLED × Dialect & Social	0.11	0.81	[-1.48, 1.69]	6310	12317

Task × No Dialect	0.29	0.47	[-0.63, 1.20]	12681	16437
Task × Dialect	-0.05	0.47	[-0.97, 0.85]	13472	17028
Task × Dialect & Social	0.18	0.46	[-0.73, 1.09]	12095	16536
Mean Voc. Test nLED × Word Type	-0.04	0.51	[-1.07, 0.96]	14472	16650
Mean Voc. Test nLED × Word Familiarity	-0.5	0.47	[-1.43, 0.42]	12431	15540
Task × Word Type	-0.16	0.3	[-0.74, 0.43]	15499	17441
Task × Word Familiarity	0.61	0.26	[0.10, 1.12]	13680	16359
No Dialect × Word Type	0.12	0.43	[-0.73, 0.96]	14184	15990
Dialect × Word Type	-0.24	0.43	[-1.08, 0.60]	14480	16847
Dialect & Social × Word Type	0.15	0.44	[-0.71, 1.01]	14962	16635
No Dialect × Word Familiarity	-0.38	0.41	[-1.18, 0.42]	11711	16109
Dialect × Word Familiarity	0.49	0.4	[-0.29, 1.27]	12021	14108
Dialect & Social × Word Familiarity	0.12	0.41	[-0.68, 0.92]	12684	15885
Mean Voc. Test nLED × Task × No Dialect	0.31	0.66	[-0.96, 1.60]	14261	17238
Mean Voc. Test nLED × Task × Dialect	-0.47	0.67	[-1.79, 0.84]	14551	16737
Mean Voc. Test nLED × Task × Dialect & Social	-0.22	0.67	[-1.54, 1.08]	14374	17169
Mean Voc. Test nLED × Task × Word Type	0.24	0.43	[-0.60, 1.09]	15568	17338
Mean Voc. Test nLED × Task × Word Familiarity	-0.75	0.39	[-1.54, 0.01]	13804	16781
Mean Voc. Test nLED × No Dialect × Word Type	0.04	0.6	[-1.15, 1.22]	14573	15758
Mean Voc. Test nLED × Dialect × Word Type	0.7	0.61	[-0.49, 1.90]	14716	16716
Mean Voc. Test nLED × Dialect & Social × Word Type	-0.2	0.63	[-1.43, 1.03]	15105	16416
Mean Voc. Test nLED × No Dialect × Word Familiarity	0.4	0.59	[-0.76, 1.55]	13039	15757
Mean Voc. Test nLED × Dialect × Word Familiarity	-0.67	0.58	[-1.81, 0.46]	13289	15567
Mean Voc. Test nLED × Dialect & Social × Word Familiarity	-0.12	0.6	[-1.31, 1.05]	13911	16592
Task × No Dialect × Word Type	0.5	0.41	[-0.30, 1.30]	14738	17135
Task × Dialect × Word Type	0.01	0.39	[-0.75, 0.77]	14793	16713
Task × Dialect & Social × Word Type	0.33	0.41	[-0.47, 1.13]	14522	17347
Task × No Dialect × Word Familiarity	0.78	0.37	[0.07, 1.52]	12302	15574
Task × Dialect × Word Familiarity	-0.68	0.37	[-1.40, 0.03]	13911	16526
Task × Dialect & Social × Word Familiarity	-0.1	0.35	[-0.79, 0.59]	14012	16597
Mean Voc. Test nLED × Task × No Dialect × Word Type	-0.44	0.56	[-1.55, 0.67]	14690	16983
Mean Voc. Test nLED × Task × Dialect × Word Type	-0.06	0.57	[-1.17, 1.05]	14979	17154
Mean Voc. Test nLED × Task × Dialect & Social × Word Type	-0.56	0.59	[-1.72, 0.59]	14644	17017
Mean Voc. Test nLED × Task × No Dialect × Word Familiarity	-0.39	0.54	[-1.44, 0.66]	12744	16179
Mean Voc. Test nLED × Task × Dialect × Word Familiarity	0.44	0.55	[-0.63, 1.50]	14418	17411
Mean Voc. Test nLED × Task × Dialect & Social × Word Familiarity	0.06	0.54	[-1.00, 1.12]	14104	17489

Note. All $\hat{R} = 1$.