To appear in *Journal of Experimental Psychology: Human Perception and Performance*

Wawk on the wild side: Context-dependence of pseudohomophone processing

Vasilena Stefanova

Vasilena.Stefanova@bcu.ac.uk, College of Psychology, Birmingham City University, Birmingham, United Kingdom

Christoph Scheepers

Christoph.Scheepers@glasgow.ac.uk, School of Psychology and Neuroscience, University of Glasgow, Glasgow, United Kingdom

Manuscript word count: 10170 words

CORRESPONDING AUTHORS:

Vasilena Stefanova
College of Psychology
Birmingham City University
4 Cardigan St
Birmingham B4 7RJ
United Kingdom
vasilena.stefanova@bcu.ac.uk

Christoph Scheepers
School of Psychology and Neuroscience
University of Glasgow
62 Hillheas Street
Glasgow G12 8QB
United Kingdom
christoph.scheepers@glasgow.ac.uk

Abstract

The *pseudohomophone effect* refers to an established finding whereby in a visual lexical decision task, non-word letter strings that are pronounced like real words (e.g., WAWK) are harder to reject than non-word strings that are not pronounced like real words (e.g., FLIS). This paper reports three lexical decision experiments that aimed at further exploring the underlying processing mechanisms. In Experiment 1 and 2, we compared pseudohomophones like WAWK with unpronounceable non-words like NRUG and pronounceable non-words like FLIS, making sure that all stimuli (including real-word fillers) were carefully matched in length, bigram frequency, and number of orthographic neighbours. Matching stimuli in this way resulted in the real-word fillers to be of low lexical frequency (lower than for the pseudohomophones' base words). Experiment 1 employed a standard lexical decision task, whereas Experiment 2 used the 2AFC eye-tracking paradigm originally developed in Kunert & Scheepers (2014). Both experiments converged on showing a *reversal* of the classical pseudohomophone effect: while unpronounceable strings like NRUG were correctly rejected relatively quickly, pseudohomophones like WAWK were indeed *easier* to reject than pronounceable non-words like FLIS. Our final Experiment 3, by contrast, confirmed a 'classical' pseudohomophone effect when the same non-word stimuli were tested against high- rather than low frequency words as fillers. We conclude that the direction of the pseudohomophone effect strongly depends on the overall material context.

Keywords: lexical decision; pseudohomophone effect; non-linear modelling

Open Materials and Data: All stimuli, data, and analysis scripts are available from the Open Science Framework (OSF) at https://osf.io/98zas/.

Public Significance Statement: This research employed a standard lexical decision task (Experiment 1 and 3) and an innovative eye-tracking task (Experiment 2) to investigate the processing of pseudohomophones (non-word letter strings that are pronounced like real words, such as WAWK). Experiments 1 and 2 show that pseudohomophones are *easier* to recognise as non-words (compared to pronounceable non-words like FLIS) when using low-frequency words as fillers. Experiment 3 shows that the effect goes in the opposite direction (replicating the 'classical' pseudohomophone effect) when using high-frequency words as fillers. This highlights the importance of the material context in which the non-word stimuli are embedded.

Wawk on the wild side: Context-dependence of pseudohomophone processing

Reading involves information-processing through the joint activation of orthography, phonology and semantics to achieve a transformation from print to meaning (Coltheart, 2005). The interplay between orthography and phonology during reading is a key element in Coltheart et al.'s (2001) Dual Route Cascaded (DRC) model of word recognition. This model builds upon previous accounts of dual route processing of letter strings and highlights the presence of a lexical route and a non-lexical route to word recognition (e.g., Van Orden et al., 1988; Seidenberg & McClelland, 1989). According to the DRC model, the lexical route involves a direct pathway from the orthographic representation of a letter string to its meaning such that exposure to a familiar letter string like MOUSE directly activates a search for the related word meaning in the mental lexicon (Taft & Russel, 1992). In contrast, processing along the non-lexical route occurs when the reader is faced with a non-word string like CLOPE (in English) or a less common word string like DIRGE. This processing route relies on phonological decoding to access word meanings via matching strings of phonemes onto semantic word representations in the mental lexicon (Braun et al., 2009; Grainger & Jacobs, 1996). It is assumed that the non-lexical, phonological route is more effortful than the lexical route as it involves an additional conversion from orthography to phonology (Borowsky et al., 2002). Experienced readers would therefore process familiar words via the more efficient lexical route, but novel or less familiar words, as well as non-words, through the more effortful non-lexical route (Ziegler et al., 2001).

The relevance of phonology in written letter-string processing is highlighted by the frequently reported *pseudohomophone effect* (e.g., Berent & Perfetti, 1995; Frost, 1998; Underwood et al., 1988; Van Orden et al., 1990; Ziegler et al., 2001). Pseudohomophones are non-word letter strings that are pronounced like real words (e.g., BREIN). Compared to

pronounceable non-words that are non-homophonic to real words (e.g., BRULK), pseudohomophones typically induce longer correct rejection times in a lexical decision task. The standard explanation of the pseudohomophone effect implies that pseudohomophones make contact with a phonological entry in the mental lexicon, thereby giving access to semantic information associated with the corresponding real words (e.g., *brain* in the case of BREIN) via the non-lexical, phonological processing route (Milota et al., 1997; Taft, 1982). The delayed lexical decision for pseudohomophones is assumed to reflect a conflict between phonological and orthographic constraints that are considered in parallel: while the phonological form of the pseudohomophone suggests that it is a real word, its orthographic form suggests that it is not. This results in a kind of *competition*, which takes time to resolve (e.g., Briesemeister et al., 2009; Van Orden et al., 1988; Ziegler et al., 2001; but see, e.g., Seidenberg et al., 1996, for an alternative account). For pronounceable non-words like BRULK, by contrast, there is no such conflict because phonology and orthography agree on the non-word status of such stimuli. Lastly, unpronounceable non-words like BKLUR were consistently found to be most easily identifiable as non-words (e.g., Kunert & Scheepers, 2014; Underwood et al., 1988), as they contain illegal grapheme strings that readers tend to detect quickly during pre-lexical analysis. Unpronounceable non-words aside, the contrast between pronounceable non-words that sound like real words versus those that do not (with lower recognition speed and accuracy for the former) is what we would call the *classical* instantiation of the pseudohmophone effect in lexical decision. It suggests that phonological decoding of an unfamiliar letter string occurs automatically and in parallel to accessing the orthographic lexicon, with the latter leading to a pseudohomophone's eventual identification as a non-word (Goswami et al., 2001; Borowsky & Masson, 1999).

However, it has been shown that processing strategies during reading may vary depending on different properties of the materials. For example, Taft (1982) found no

difference in reaction times for pseudohomophones and pronounceable non-words that were orthographically similar to real words – both yielded slower responses compared to pronounceable non-words that did not approximate real words orthographically. This suggests that orthographic similarity to real words could diminish the pseudohomophone effect. Similarly, Martin (1981) found no significant difference in lexical decision times when comparing pseudohomophones to orthographically matched non-word controls, while 'approximate' orthographic controls and 'distant' orthographic controls were identified as non-words significantly faster. This suggests that readers may flexibly adapt their reliance on phonological vs. orthographic processing in reading, depending on the specifics of the materials. In a study by Kunert and Scheepers (2014), differences between typical versus dyslexic English readers were investigated through the use of a novel two-alternative forced choice (2AFC) lexical decision task measuring participants' eye-movements to pairs of word vs. non-word strings on screen. This task was also employed in our Experiment 2 and will be explained in more detail further below. Regardless of participant group, Kunert and Scheepers (2014) found no clear differences in processing speed or accuracy between pseudohomophone stimuli like `LEPHT` and pronounceable non-word strings like `STOINT`. The absence of a significant pseudohomophone effect resonates with the previously discussed findings by Martin (1981) and Taft (1982). However, in the case of Kunert and Scheepers (2014), the lack of a clear pseudohomophone effect could potentially be attributed to less tightly controlled orthographic neighbourhood characteristics of the materials. Indeed, only a subset of their pseudohomophone stimuli were orthographic neighbours of the words they were derived from (note that Coltheart et al., 2001, previously demonstrated stronger effects for pseudohomophones that were orthographic neighbours of their base words than for pseudohomophones that were not).

Another factor of interest in the present paper is the lexical frequency of a pseudohomophone's *base word* (essentially, the word it was derived from). Lexical frequency refers to how often a word occurs in a language and is known to reliably affect response times in lexical decision tasks (e.g., Frederiksen & Kroll, 1976; Monsell et al., 1989; Whaley, 1978). From naming tasks (i.e., tasks that require vocal pronunciation of written letter strings), it is suggested (i) that pseudohomohones derived from high-frequency words evoke shorter latencies than those derived from low-frequency words and (ii) that pseudohomophones are named faster than pronounceable non-words that do not sound like real words (e.g., McCann & Besner, 1987; Taft & Russel, 1992). In the visual lexical decision literature, the picture appears somewhat mixed. Some studies revealed no clear evidence for a base-word frequency effect (McCann et al., 1988; Seidenberg et al., 1996) while others showed a robust effect in the direction of faster and more accurate lexical decisions for pseudohomophones derived from high- rather than low-frequency words, just as in naming (Van Orden, 1991; Van Orden et al., 1992; Ziegler et al., 2001, Tiffin-Richards & Schroeder, 2018). The perhaps most prominent example is Ziegler et al. (2001)'s study on pseudohomophone processing in German. This study confirmed the base-word frequency effect on top of the classical pseudohomophone effect, i.e., pseudohomophones were harder to reject than appropriate pronounceable controls. Interestingly, this latter aspect contrasts with the aforementioned findings from the naming literature (which indicate the reverse), suggesting that different task requirements (naming requires explicit pronunciation of written stimuli, whereas visual lexical decision does not) might modulate the direction of the contrast between pseudohomophones on the one hand and pronounceable non-word controls on the other.

The present paper aims at gaining further insights into the phonological and orthographic processing mechanisms underlying word vs. non-word discrimination in

English. We used two different methodologies. Experiment 1 and 3 implemented a standard visual lexical decision task, whereas Experiment 2 employed Kunert and Scheepers's (2014) 2AFC eye-tracking paradigm. All experiments used the same set of carefully controlled non-word stimuli, including 50 pseudohomophones like WAWK (these were selected from previous papers that revealed 'classical' pseudohomophone effects), 50 unpronounceable non-words like NRUG, and 50 pronounceable non-words like FLIS, which are non-homophonic to existing English words. We also included 150 English word stimuli as filler materials. In Experiment 1 and 2, all stimuli (including the real-word fillers!) were carefully matched in length, bigram frequency, and number of orthographic neighbours. As far as we are aware, matching the filler words with the critical non-word stimuli in this way has not been implemented before. Our main motivation behind this setup was to create conditions whereby word vs. non-word discrimination should primarily rely on phonological features of the stimuli, and less so on orthographic features. One consequence of the rigorous matching was that the word materials in Experiment 1 and 2 were of relatively low lexical frequency. In Experiment 3, however, we tested our non-word stimuli against relatively high frequency words as fillers, which were only matched with the non-words in terms of length.

The research we report here was primarily exploratory in nature, i.e., we refrained from testing specific theoretical hypotheses or models. Rather, we were interested in the question of how rigorous control of orthographic variables like length, bigram frequency, and number of orthographic neighbours across the entire material set (this included the filler words in Experiments 1 and 2, but not in Experiment 3) would affect the size and/or direction of the pseudohomophone effect. With 48 participants per study (about twice as many as in typical studies on pseudohomophone processing) and 50 items per condition, our experiments had reasonable statistical power (see also *Required Sample Size* below).

Experiment 1

The first experiment used a standard visual lexical decision task whereby letter strings were presented one-by-one on screen and participants had to decide (via button response) whether the presented string was an actual English word or not. Apart from words serving as fillers, there were three critical non-word conditions: pseudohomophones like WAWK were compared with unpronounceable non-words like NRUG and pronounceable non-words like FLIS, which are non-homophonic to existing English words. The design was fully within-subjects/between-items, and the dependent variables were recognition accuracy and reaction time for correct responses.

Method

Required Sample Size. Brysbaert and Stevens (2018) pointed out that RT effect sizes can be very small. For example, for data from Adelman et al.'s (2014) large scale lexical decision study on orthographic priming (1020 participants, 420 items), they found that a population effect of ca. 16 ms translates into to a Cohen's *d* of only 0.087 when random variation at participant, item, and trial level are taken into account. We used this as a benchmark for our own power analyses, which were performed in J. Westfall's power calculation app for designs with crossed random effects (Westfall, Kenny, & Judd, 2014). We aimed to be able to detect a cross-condition contrast of 20-40 ms (reflecting the range of previously reported pseudohomophone effects) and so we assumed a minimum population *d* of 0.15, which is still a conservatively low estimate. Since our design was within-participants but between-items (or "Stimuli-within-Condition" in Westfall et al.'s terminology), we had to specify standardized random variance components (in proportions of 1) for the participant and item random intercepts, the by-participant random slope, and the trial-level residual term. After considering various random effect estimates for RT experiments that are available in

the literature (e.g., Brysbaert & Stevens, 2018; Barr et al., 2013's appendix on random effects in real data sets) we entered the following plausible values into the app: participant intercept = 0.298, by-participant slope = 0.002, item intercept = 0.05, and residual variance = 0.65. Keeping the number of stimuli fixed at 100 (contrasting two conditions with 50 items each), we then increased the number of participants from 20 to 50 by increments of 1. This gave us an estimated minimum requirement of 43 participants to be able to detect a cross-condition contrast of $d = 0.15$ at $p < .05$ (2-tailed) with at least 80% power. Since this figure does not take potential data loss due to outliers and/or incorrect responses into account, we set our actual participant sample size to N = 48 (10.4% higher than the required minimum).

Participants. Forty-eight native English speakers (age range: 18-36 years; 65% females) took part in this study in exchange for course credits. All participants were right-handed and had normal or corrected-to-normal vision. None of the participants reported having any visual or reading-related impairments. Ethical approval for this study was obtained from the College of Science and Engineering Ethics Committee at the University of Glasgow. All participants gave written informed consent.

Materials. The complete list of materials is available at https://osf.io/98zas/, sub-folder *Stimuli_and_Norms*. For the pseudohomophone condition (PH), we selected 50 items (e.g., `WAWK`) from previous 'classical' demonstrations of the pseudohomophone effect in written word processing (Lupker & Pexman, 2010; Reynolds & Besner, 2005; Borowsky et al., 2002; Seidenberg et al., 1996; Taft & Russel, 1992). Each of these was closely matched with an unpronounceable non-word like `NRUG` (UP condition), and a pronounceable non-word like `FLIS` (PN condition), creating 150 critical non-word stimuli altogether. In addition, we selected 150 monosyllabic English words which acted as decoys for the lexical decision task. Each word was closely matched with one of the 150 critical non-word stimuli.

Primary matching criteria were length in characters, log cumulative bigram frequency, and number of real-word orthographic neighbours, extracted from the English Lexicon Project database (Balota et al., 2007). All pronounceable stimuli (PH, PN, and the filler words) were monosyllabic. For the filler words, as well as the words to which the 50 PH stimuli were phonologically related, we also extracted lexical frequency data from the British National Corpus (Version 3, BNC XML Edition, 2007). Table 1 provides a summary of the item norms across the different subgroups of stimuli. As can be seen, there were no appreciable cross-condition differences in the means per variable. However, note that lexical frequencies were considerably higher for the base words of the PH stimuli than for the filler words. Indeed, the PH stimuli were phonologically related to relatively high frequency base words (with a mean occurrence of ca. 43 per million in the BNC corpus), whereas the filler words were from a much lower frequency band (ca. 3 occurrences per million on average). The latter was likely due to the close match in the words' orthographic features to the non-words.

Table 1. Means (*SE*s in brackets) for the control variables length (in characters), log cumulative bigram frequency (Bigram Frequency), and number of orthographic neighbours (Orthogr. Neighbours) in each of the critical non-word conditions (PH, UP, and PN). Also shown are the relevant figures for the filler words used. The last column (Lexical Frequency) shows mean $\log_{10}$ lexical frequencies per million for the PH non-words' *base words* and the three subgroups of matched filler words, respectively.

	N	Length	Bigram Frequency	Orthogr. Neighbours	Lexical Frequency
PH Non-words	50	4.66 (.1)	3.12 (.05)	2.84 (.39)	1.63 (.11)
UP Non-words	50	4.66 (.1)	3.12 (.05)	2.84 (.37)	NA
PN Non-words	50	4.66 (.1)	3.12 (.05)	2.82 (.38)	NA
PH-matched Words	50	4.66 (.1)	3.13 (.05)	2.82 (.38)	0.41 (.12)
UP-matched Words	50	4.66 (.1)	3.13 (.05)	2.88 (.39)	0.44 (.14)
PN-matched Words	50	4.66 (.1)	3.13 (.04)	2.78 (.37)	0.53 (.14)

Length, *Bigram Frequency*, and *Orthographic Neighbors* for the critical non-word stimuli were entered into a Principal Component Analysis (PCA) with *Varimax* rotation in order to condense them into a smaller set of uncorrelated factors. The PCA showed that ca. 92% of the original control predictor variability could be preserved by extracting two principal components: the first (henceforth labelled *LBF*) explained 52% of the original variance and showed a strong positive loading from *Length* (.82) and a strong negative loading from *Bigram Frequency* (−.95) (i.e., longer stimuli had lower cumulative bigram frequencies); the second component (henceforth *ON*) explained 39% of the variance and received a strong positive loading from *Orthographic Neighbours* only (.84). The standardized factor scores of the two principal components were used as covariates in subsequent analysis. Using principal components instead of the original variables not only

reduced the number of control predictors, but also prevented collinearity issues in the analyses (principal components are uncorrelated).

Procedure. A computer-based lexical decision task was implemented in DMDX (Forster & Forster, 2003). Prior to testing, all participants took part in a short training session comprising 20 trials (10 pseudohomophones and 10 words in random succession), using stimuli different from those described in the Materials section. This not only acquainted participants with the task, but also served to increase the detectability of potential pseudohomophone effects (research by Underwood et al., 1988, suggested that pseudohomophone effects become stronger after previous exposure to pseudohomophone examples). After the training session, each participant completed 300 trials involving the presentation of four types of stimuli – pseudohomophones (PH), unpronounceable non-words (UP), pronounceable non-words (PN), and actual English words (fillers) – in an individually determined random order. On each trial, participants had to indicate whether the presented stimulus was a word or not by pressing either the left (index finger) or right (middle finger) mouse button, for 'yes' and 'no' respectively. Reaction times (in ms) and accuracy of the button-responses were recorded. We did not counterbalance 'yes'/'no' button assignments because our primary focus was on the non-word trials only (counterbalancing might have added noise related to motoric preferences). The stimuli appeared centred on screen, printed in black 14 point Courier font (all capitals) on a white background. After providing a button response, participants were given feedback, i.e. either a green tick (for correct) or a red X (for incorrect) was presented, which stayed on screen for 400 ms and was followed by a 100 ms blank screen before the next trial was initiated. If the participant did not respond within 3.5 seconds, the message "too late" (printed in red) appeared and the next trial was initiated. The task took ca. 10-15 minutes to complete and was carried out on a PC in a dimmed experimental room. Every item appeared only once during the experiment.

Analysis. Only non-word trials (PH, UP, and PN condition) were considered, resulting in 48 (participants) × 150 (items) = 7200 trial-level observations for analysis (2400 data points per condition). Response times below 250 ms or above 2500 ms were extremely rare, accounting for less than 1% of the data. We excluded the corresponding trials from further analysis as they were unlikely to reflect normal processing.

Inferential analyses were based on Generalized Linear Mixed Models (GLMMs) in R (R Core Team, 2018) using the package `lme4` (Bates et al., 2015). Two types of analyses were performed, one focusing on response accuracy and one on RTs for correct responses (excluding trials where participants responded incorrectly).

Since response accuracy is dichotomous, we analyzed this variable using binary logistic GLMMs predicting log odds of correct responses from a combination of various fixed and random effect parameters (see below). Correct RT was analyzed via GLMMs combining the *Gamma* distribution with *Identity* link in the family argument (see Lo & Andrews, 2015), thereby accounting for a strong positive skew in the distribution of correct RT while still assuming linear relations between predictor and outcome variables.

The fixed effects of the GLMMs comprised the main effect of condition (PH, UP, and PN), the main effects of the two covariates (principal components *LBF* and *ON*, see Materials section), as well as two-way interactions between condition and each of the covariates (the latter to account for potentially different slopes of the covariates dependent on condition). Interactions between covariates were not considered, as their inclusion was theoretically difficult to justify (exploratory analyses did not justify their inclusion either). Where appropriate, we will also report analyses without the covariates. All predictors were entered into the models in mean-centred form (deviation coding), such that model intercepts were aligned with the grand means. The three-level predictor *condition*, as well as the *covariate by*

condition interactions, required two contrast-parameters each. We treated the PH condition as comparison baseline.

To ensure generalizability of results across participants and items, all analyses employed the maximal random effects structure justified by the design (Barr et al., 2013), including random correlations. Since all predictors (covariates included) were *within-subjects*, we not only added by-subject random intercepts (accounting for inter-individual differences in overall responses) but also by-subject random slopes for every fixed effect term in the GLMMs (accounting for participant-specific variation in the fixed effects). By-item random intercepts were also included (accounting for stimulus-related variation in overall responses), but no by-item random slopes since all predictors were *between-items*.

In a supplementary set of GLMM analyses, we focused on PH trials only. Specifically, we assessed whether accuracy and correct RT for the processing of PH stimuli were predictable from the lexical frequencies of their phonologically related base words, using the relevant BNC corpus data as a continuous predictor. Again, these analyses employed mean-centred predictor variables and maximal random effect structures.

Collinearity between predictor terms was negligible in all of the reported GLMMs (*VIF*s < 1.5).

Results

The raw descriptive figures per condition are shown in **Table 2**. In terms of both speed and accuracy, UP letter-strings were most easily recognised as non-words. Interestingly, the PH vs. PN contrasts were in the *opposite* direction to the 'classical' pseudohomophone effect, indicating lower accuracy and higher correct RTs for PN than for PH stimuli.

Table 2. Observed probability of correct responses (Accuracy) and correct RT (in ms) across all valid trials per condition in Experiment 1 (PH = pseudohomophone, UP = unpronounceable non-word, PN = pronounceable non-word). The last row shows corresponding figures for the filler words (these were of relatively low lexical frequency and were not considered in subsequent inferential analyses). Means and *SE*s are not adjusted for covariate influences and do not take participant or item variability into account.

	Accuracy		Correct RT	
	Mean	SE	Mean	SE
PH	.880	.007	824	6.6
UP	.973	.003	739	5.2
PN	.862	.007	855	6.8
Filler Words	.867	.004	710	3.2

The inferential results corroborated this pattern. **Table 3** shows GLMM parameter estimates and **Figure 1** plots covariate-adjusted model estimates per condition, separately for response accuracy and correct RT. As can be seen, UP stimuli had a very clear speed and accuracy advantage over both PH and PN stimuli. More critically, there was a significant PH vs. PN contrast in correct RT, indicating that PH stimuli were correctly identified *faster* than PN stimuli (a *reversal* of the classical pseudohomophone effect!). The corresponding contrast was not significant in response accuracy, but pointed in the direction of reduced accuracy for PN compared to PH stimuli.

When the control predictors (*LBF* and *ON*) and their interactions with experimental condition were removed from the GLMMs, significance patterns stayed the same. The UP condition differed significantly from the PH baseline, both in terms of response accuracy (b = 2.095, SE = 0.302, z = 6.946, $p < 0.001$) and in terms of correct RT (b = –101.21, SE = 4.26, z = –23.769, $p < 0.001$). The PN condition did not differ reliably from the PH baseline in accuracy (b = –0.202, SE = 0.230, z = –0.879, p = 0.380), but very clearly so in correct RT (b

= 25.10, *SE* = 4.96, *z* = 5.056, *p* < .001), again indicating a reversal of the classical pseudohomophone effect.

Table 3. Fixed effects estimates from the full GLMMs of *response accuracy* (left; corresponding *b*s and *SE*s are in log odds units), and *correct RT* (right; *b*s and *SE*s in millisecond units) in Experiment 1. The first three rows show estimated grand averages (intercepts), estimates for the effect of UP relative to PH, and estimates for the effect of PN relative to PH (the PH condition always served as a comparison baseline). The next two rows show estimated main effects of the two covariates (*LBF* and *ON*), and the last four rows show estimates for the *condition by covariate* interactions (again, relative to the PH baseline). Significant effect parameters ($p < .05$) are highlighted in bold.

	Accuracy				Correct RT			
Source	*b*	*SE*	*z*	*P*	*b*	*SE*	*z*	*p*
Intercept	3.312	0.187	17.700	< .001	842.08	10.92	77.085	< .001
UP	**2.321**	**0.331**	**7.006**	**< .001**	**−97.79**	**12.26**	**-7.977**	**< .001**
PN	−0.208	0.217	−0.961	.337	**28.24**	**13.81**	**2.045**	**.041**
LBF	**−0.383**	**0.126**	**−3.041**	**.002**	**34.91**	**5.60**	**6.233**	**< .001**
ON	**−0.361**	**0.103**	**−3.517**	**< .001**	**13.93**	**5.76**	**2.421**	**.016**
UP × *LBF*	0.080	0.305	0.262	.794	−23.79	12.36	−1.924	.054
UP × *ON*	−0.231	0.266	−0.868	.386	11.17	13.59	0.822	.411
PN × *LBF*	−0.255	0.235	−1.085	.278	3.31	14.09	0.235	.815
PN × *ON*	**0.490**	**0.208**	**2.358**	**.018**	−17.64	14.43	−1.222	.222

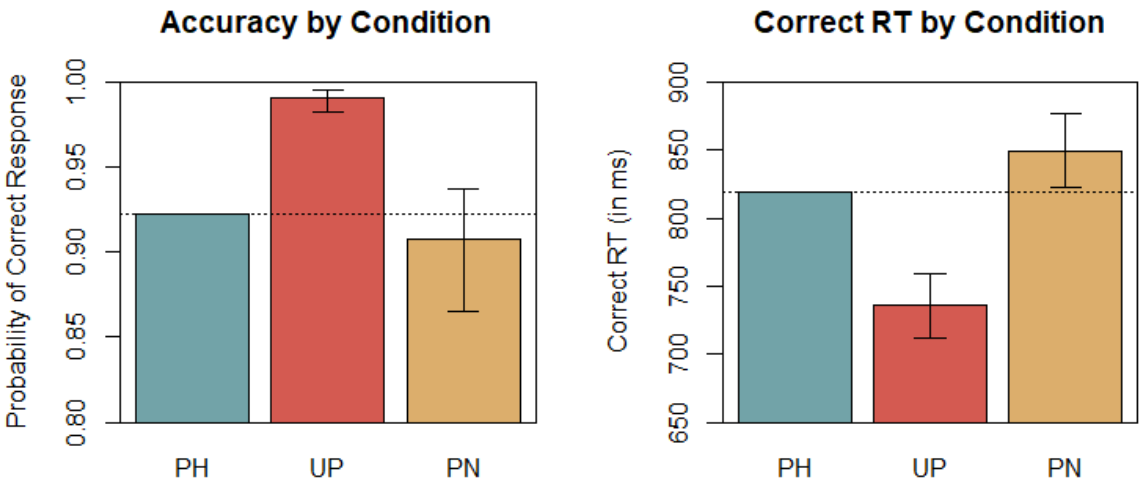

Figure 1. GLMM-estimated (covariate-adjusted) probabilities of correct responses (left) and correct RTs (right) as a function of experimental condition (PH = pseudohomophone, UP = unpronounceable non-word, PN = pronounceable non-word) in Experiment 1. Error bars represent 95% CIs for contrasts with the PH condition (comparison baseline, as indexed by horizontal dotted lines). The asymmetries in the CIs on the left result from converting log odds model predictions into probabilities.

In terms of control predictor influences (which are not of primary interest) **Table 3** shows reliably negative main effects of *LBF* and *ON* in response accuracy (longer stimuli with lower cumulative bigram frequencies, and stimuli with more orthographic real-word neighbours, were less likely to be correctly identified as non-words) and reliably positive main effects in correct RT (longer stimuli with lower cumulative bigram frequencies, and stimuli with more orthographic real-word neighbours, took longer to be correctly identified as non-words). The only significant interaction term showed up in response accuracy: relative to the PH reference condition, numbers of orthographic real-word neighbours had a less negative influence on response accuracy in the PN condition. Finally, the marginal UP × *LBF*

term in correct RT suggests that, relative to the PH reference condition, *LBF* had a somewhat less positive effect on correct RT in the UP condition.

Pseudohomophone processing as a function of base-word frequency. Results from the complementary GLMMs indicated that increasing lexical frequencies of the pseudohomophones' base words led to a marginal increase in response accuracy and to a significant decrease in correct RT (**Figure 2**). These effects held true regardless of whether the control predictors (*LBF* and *ON*) were included in the analysis or not. Accuracy: b = 0.366, SE = 0.207, z = 1.767, p = 0.077 (including control predictors); b = 0.351, SE = 0.207, z = 1.695, p = 0.090 (excluding control predictors); correct RT: b = –27.77, SE = 8.67, z = –3.205, p = .001 (including control predictors); b = –49.03, SE = 11.89, z = –4.122, p < .001 (excluding control predictors).

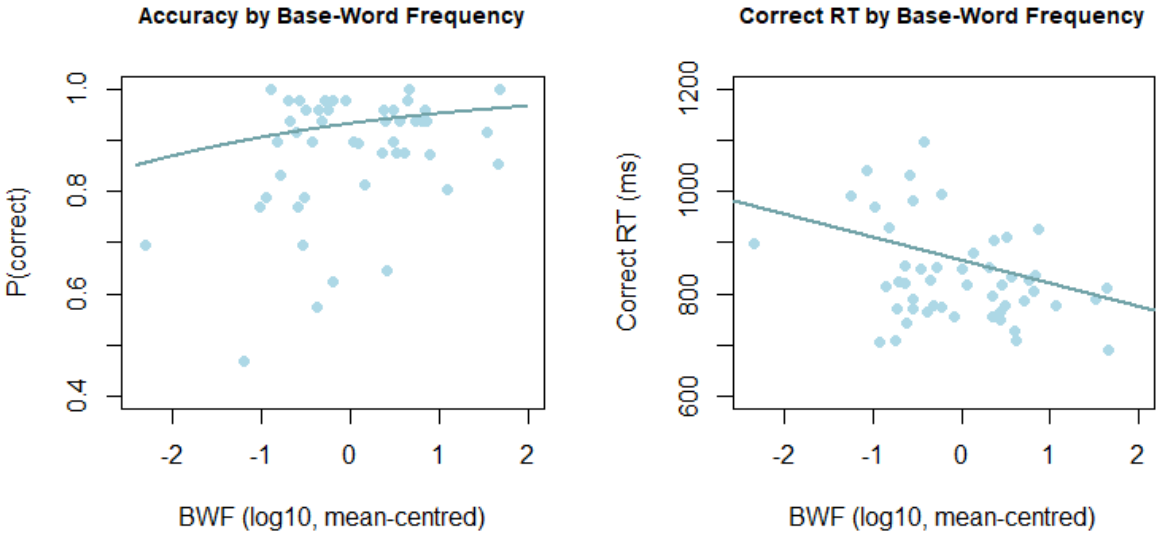

Figure 2. Response accuracy (left) and correct RT (right) for *pseudohomophone* stimuli (PH condition) as a function of associated *base-word frequency* (BWF) per million in the British National Corpus, represented on a mean-centred $\log_{10}$ scale. For illustration, observed data were aggregated into item means (blue dots, N = 50). GLMM fits were determined using trial-level data (also accounting for by-participant variation, which is not visible in the plots). The regression line on the left stems from a binary logistic GLMM predicting probability of correct responses from BWF and the two control predictors (*LBF* and *ON*, not shown); the curvature of the line results from converting model-estimated log odds into probabilities. The regression line on the right stems from a Gamma GLMM predicting correct RT from BWF and the two control predictors (*LBF* and *ON*, not shown).

Discussion

Apart from confirming that unpronounceable (UP) stimuli are considerably easier to identify as non-words than PH or PN stimuli (in line with earlier studies), Experiment 1 found a clear trend in the direction *opposite* to the classical pseudohomophone effect: PH stimuli were correctly classified *faster* than PN stimuli, and participants were numerically (though not significantly) more accurate in identifying PH than PN stimuli as non-words (**Figure 1**). This is remarkable considering that we not only selected our PH stimuli from

previous ‘classical’ demonstrations of the pseudohomophone effect, but also implemented conditions that should have promoted the occurrence of such an effect (warm-up session at the start of the experiment, cf. Underwood et al., 1988).

Our results go beyond previous findings whereby the pseudohomophone effect disappeared when PH and PN stimuli were closely matched in terms of orthographic features (e.g., Martin, 1981; Taft, 1982; Kunert & Scheepers, 2014). Indeed, the present data suggest that not only the occurrence, but even the *direction* of the pseudohomophone effect is subject to modulation in a lexical decision task. Specifically, when (i) orthographic features are carefully matched across the entire stimulus set *and* (ii) filler words are from a lower lexical frequency band than the PH stimuli’s base words, the pseudohomophone effect is reversed.

There was evidence that the present setup created conditions whereby the pseudohomophones were often perceived as incorrectly spelled *words*. In particular, the complementary results in **Figure 2** revealed that PH stimuli were classified as non-words *faster* (and marginally more accurately) when the lexical frequencies of their base words increased. This result resonates with findings from the naming literature (e,g., McCann & Besner, 1987; Taft & Russel, 1992) and with Ziegler et al. (2001)’s findings on lexical decision in German. It suggests that PH non-words were likely to activate phonologically related real-word representations during lexical decision, but crucially, in a way that *benefited* their correct identification as non-words rather than causing interference.

We hypothesize that the present findings reflect a two-stage process. In line with Coltheart et al. (2001), we assume that the first stage consists of *phonological decoding* of unfamiliar word candidates (the default, relatively slow processing route for non-word strings, according to the DRC model), combined with a search through the mental lexicon for associated word meanings. The second stage consists of an additional *orthographic check*

before the final lexical decision (see also, e.g., Ziegler et al., 2001; Tiffin-Richards & Schroeder, 2018). The first stage (phonological decoding and lexical search) is likely to be cut short when readers are presented with UP stimuli like NRUG, because such strings do not lend themselves to meaningful phonological decoding. This implies relatively fast and highly accurate 'non-word' decisions for UP stimuli. With pronounceable (PH and PN) stimuli on the other hand, phonological decoding and search through the mental lexicon continues, which may explain the increased RTs relative to the UP condition. The crucial difference between PH and PN stimuli is that while the latter (e.g. FLIS) do not possess any matching phonological representations in the mental lexicon, the phonological forms of the former (e.g., WAWK) are likely to activate the meanings of their base words (*walk* in this example). This means that the second stage (*orthographic verification*) can be initiated within a reasonable amount of time in the PH trials (leading to correct 'non-word' decisions), whereas in PN trials, search through the mental lexicon continues until a potential time-out threshold is reached. This could explain the longer RTs and slightly lower response accuracies for PN compared to PH stimuli.

More generally, perceiving pseudohomophones as misspelled words (via phonological access to their base-word meanings) could have given PH stimuli a small but measurable processing advantage over PN stimuli, since the lack of any phonological matches for the latter prolongs lexical search and creates more uncertainty about their lexical status. That is, PN stimuli were more likely to be perceived as *potential* words that are unfamiliar.

Importantly, under the conditions implemented here, a non-word's orthographic familiarity is unlikely to be assessed before phonological search through the mental lexicon has been completed – either via early interruption (UP condition), via finding a match (PH condition), or via cancellation after a time-out period (PN condition). In fact, without the

assumed precedence of phonological over orthographic processing, it would be difficult to account for the present data. Recall that bigram frequencies and numbers of orthographic real-word neighbours were controlled for (both by design and analytically), which renders cross-condition differences in word-like orthographic appearance an implausible explanation of our findings.

The mechanisms outlined above allow for specific predictions with regard to the *time-course* of processing PH, UP, and PN stimuli. In particular, we expected that PH stimuli are more likely to be confused with real words 'early on' compared to UP or PN stimuli. PN stimuli, by contrast, should exhibit longer periods of uncertainty about their lexical status than PH or UP stimuli. To test these predictions, the following experiment employed a continuous lexical decision task based on eye-tracking.

Experiment 2

This experiment used the same 2AFC lexical decision eye-tracking technique as in Kunert and Scheepers (2014). Participants saw pairs of letter strings (one being a word, the other being a non-word) and were instructed to *keep looking at the word and ignore the non-word* for a period of 3.5 seconds while their eye movements were recorded. Word and non-word stimuli were identical to those in Experiment 1. As explained in Kunert and Scheepeers (2014), the 2AFC eye-tracking paradigm offers advantages over classical lexical decision techniques in that it allows for the modelling of lexical decision processes as a continuous function of time rather than relying on a single response per trial. As such, it is not only virtually immune to speed-accuracy trade-offs in responding, but also enables a more fine-grained modelling of the time course of lexical decision processes.

Method

Participants. A new sample of 48 native English speakers took part in exchange for course credits. Participants' age ranged from 18 to 36 years, and 73% were female. All participants had normal or corrected-to-normal vision and reported having no visual or reading-related impairments. Ethical approval for this study was obtained from the College of Science and Engineering Ethics Committee at the University of Glasgow. All participants provided written informed consent.

Design and Materials. For Experiment 2, the PH, UP and PN stimuli, as well as the word stimuli, were identical to those in Experiment 1. Using a Latin square rotation scheme, we compiled six counterbalanced lists of word/non-word pairs.

The original non-words came in triplets of stimuli that were matched in terms of length, bigram frequency, and number of orthographic neighbours (e.g., PH: WAWK, UP: NRUG, PN: FLIS,). Moreover, each member of a triplet was closely matched to a real word (e.g., QUAY, GAWK, and FEUD, respectively). Thus, we first created three lists such that, across lists, each member of a non-word triplet was paired with each of the three real word counterparts. Further, to counterbalance the positioning of the word versus non-word per pair (with the word appearing either above or below the centre of the screen, and vice versa for the non-word), we had to generate another three lists. Each of the resulting six lists therefore contained 150 word/non-word pairs, such that (i) each of the 150 non-word stimuli (50 per condition) was paired with three different words across lists (the same word/non-word pairings therefore occurred in two of the six lists) and (ii) positioning of words versus non-words (above or below screen centre) was fully counterbalanced across conditions, both within and across lists. That is, half of the word/non-word pairings per condition per list had

the word appearing above and the other half below the centre of the screen, and this was counterbalanced across lists.

Each participant was randomly assigned to one of the six presentation lists. We therefore ended up with eight participants per list. Each list further included 10 word/non-word pairings as warm-up trials at the beginning. These were derived from the materials of the training session in Experiment 1 and were identical across lists. Half of the warm-up pairings had a PH stimulus above and an actual word below the screen's centre, and vice versa for the other half.

Procedure. Participants were tested using an SR-Research EyeLink II head-mounted eye-tracker using a sampling rate of 500 Hz. Participants were seated ca. 80 cm from a 21 inch CRT display running at 120 Hz refresh rate in 1024 × 768 pixel resolution. Prior to testing, the participants' dominant eye was determined using a simple parallax test. Viewing was binocular, but only the participant's dominant eye was tracked. The eye-tracker was calibrated using the standard 9-point calibration and validation routines of the SR-Research ExperimentBuilder software. Spatial calibration error stayed below 0.5 degrees of visual angle (software default). This happened at the start of each session, after the initial block of 10 warming-up trials, and then after each block of 50 experimental trials. The experiment ended with a "thank you" message after all 150 experimental trials were completed. The experimenter could interrupt the trial sequence at any point for recalibration (e.g., if they noticed a decline in measurement accuracy), but this was required very rarely. Before recalibration, participants were given the opportunity to take a short break if needed.

The 10 warming-up trials at the beginning appeared in a fixed random order. The ordering of the 150 experimental trials thereafter was randomly determined for each participant. Each trial started with the presentation of a black fixation dot against a light grey

background (RGB 225, 225, 225) which appeared in the centre of the screen. When the participant fixated it, the experimenter pressed the space bar of the Operator PC to trigger an automatic drift correction, upon which the fixation dot was replaced with a pair of letter strings for 3.5 seconds. The letter strings were printed in black 28pt Courier all-cap font on the same light grey background as before. Their geometric centres were located ca. 1.5 degrees of visual angle above respectively below the previous fixation dot, so that participants had to perform a vertical eye-movement away from the screen's centre to inspect either of them more closely. One of the letter strings was a word and the other one a non-word. Half of the trials per participant showed the word above and the non-word below the centre, and vice versa for the other half of trials (see Design and Materials). The participants' task was to *keep looking at the word and ignore the non-word*. After 3.5 seconds, a 500 ms blank screen appeared before the next trial was initiated.

No button responses were required, and a typical session lasted around 15-20 minutes.

Analysis

Data Pre-processing. To model probabilities of looks to the word vs. non-word over time, we first defined three areas of interest (AOIs): *word*, *non-word*, and *background*. The pixel coordinates of the longest letter strings in our materials were extended by roughly 0.5 degrees of visual angle above and below and to the left and right of its perimeters. Any fixations that were within those extended pixel coordinates were scored as looks to the *word* or *non-word*, respectively (depending on whether the word was above or below the centre of the screen), and all remaining fixations were scored as looks to the *background*. Eye-blink events (plus 50 ms before and after the they were detected by the eye-tracker) were scored as missing values on any of these AOIs. These scorings were down-sampled to 40Hz resolution (25 ms per time-slice) over the entire trial period of 3.5 seconds. We further discarded ca.

10% of trials where participants were already looking at the *word* or *non-word* AOI at the start of the trial, which was likely due to faulty drift-correction. For analysis, the data were normalised such that probabilities of looks to the *word*, *non-word*, and *background* always added up to 1 in each time-slice.

Figure 3 shows *cumulative-average fixation probabilities* over time for each AOI. In line with Kunert & Scheepers (2014), we used cumulative averages over time for statistical modelling since they produced smoother curves than the corresponding raw averages (albeit with somewhat delayed timing characteristics).[1] The general pattern was very similar in the raw averages and can be described as follows. During the first 125 ms, virtually all fixations were on the *background*, mainly because this AOI also included the position of the central fixation dot at the start of the trial. Thereafter, probabilities of fixations to the *word* versus *non-word* continuously increased at about equal rates (i.e., showing no discrimination between *word* and *non-word*) up until a point before probabilities of looks to the *non-word* approached their overall maximum. From then onwards (after around half a second), probabilities of looks to the *non-word* continuously dropped in favour of looks to the *word*, as is visible in a clear separation between the red and the green curves in the figure.

[1] The cumulative average (a.k.a. *rolling average*) of a time series can be illustrated as follows. Assume that over four consecutive time slices, a given AOI was either fixated (score of 1) or not (score of 0) as in the following example: $0 \rightarrow 1 \rightarrow 1 \rightarrow 0$. The cumulative average per time slice is calculated as $\frac{0}{1} \rightarrow \frac{0+1}{2} \rightarrow \frac{0+1+1}{3} \rightarrow \frac{0+1+1+0}{4}$ (i.e., $0 \rightarrow 0.5 \rightarrow 0.667 \rightarrow 0.5$). Note that the function is not monotonically increasing since the averaging denominator increases with each additional time slice.

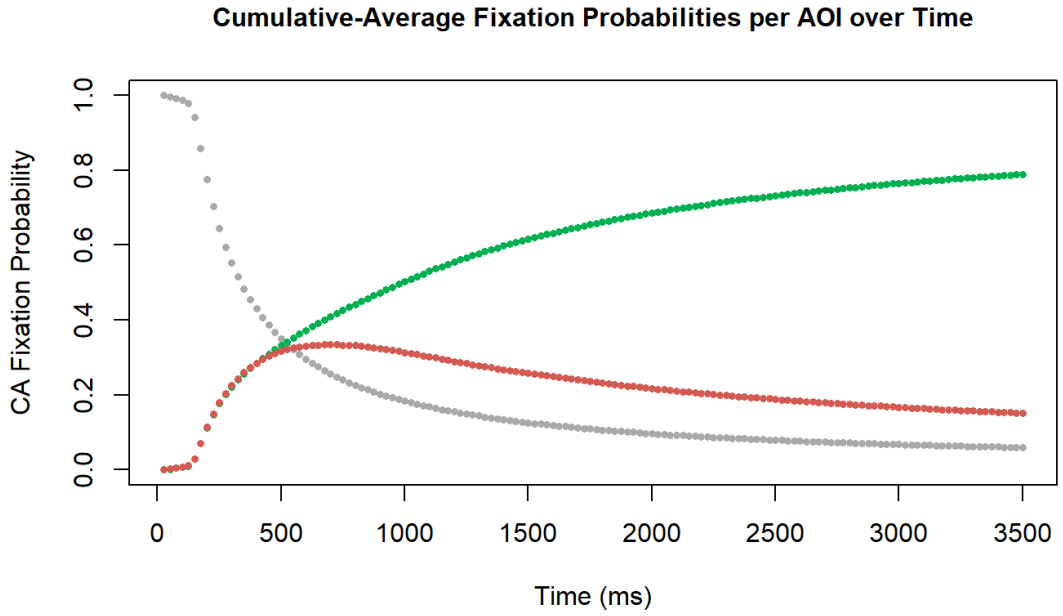

Figure 3. Cumulative-average fixation probabilities over time (40 Hz resolution), broken down by AOI: *word* in green, *non-word* in red, and *background* in grey.

Curve Modelling. In our analyses, we were primarily concerned with modelling cumulative-average proportions of looks to the *non-word* AOI as a function of time (red curve in **Figure 3**), which is conceptually analogous to modelling error rates in a lexical decision task. Through trial and error (employing the extensive library of non-linear peak functions in *TableCurve2D*®), we identified the *Equilibrium Peak* function as the best description of these data, both overall and across individual non-word conditions. **Figure 4** illustrates the function and its parameters. While being rather complex, this function offered sufficient flexibility to ensure optimal fits of our data.

Six-Parameter *Equilibrium Peak* Function

$$\hat{y} = A + B\left(1 - exp\left(-C(x-F)\right) - \frac{D}{D+E}\left(1 + \frac{C\,exp\left(-(D+E)(x-F)\right) - (D+E)\,exp\left(-C(x-F)\right)}{D+E-C}\right)\right)(x > F)$$

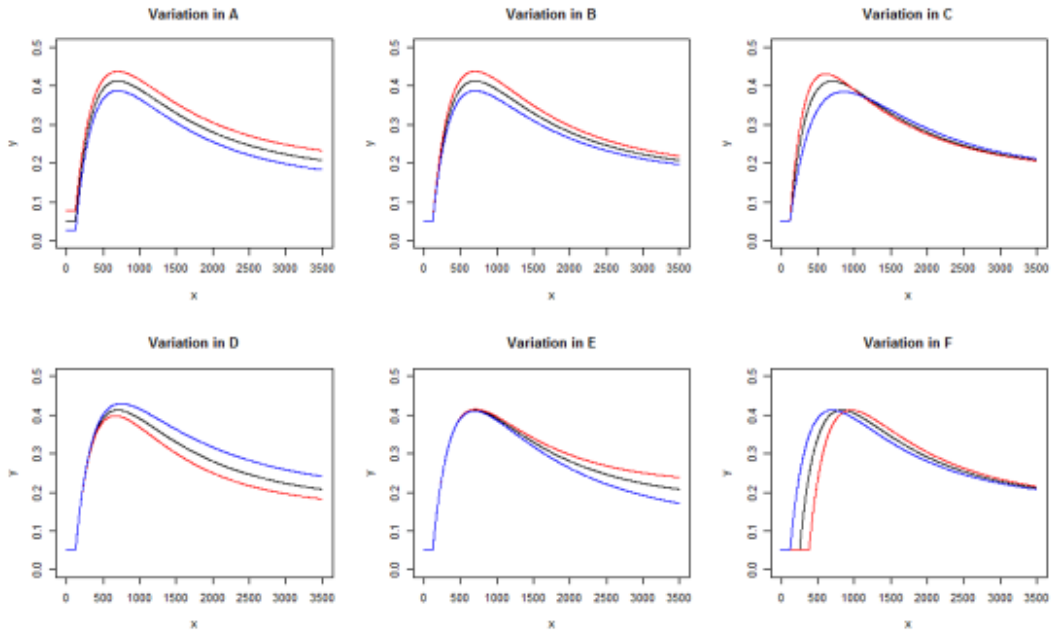

Figure 4. Six-parameter *Equilibrium Peak* function. The model equation is shown in the grey box at the top. The plots below illustrate how variation in each parameter (using arbitrarily 'low' [blue], 'medium' [black], and 'high' [red] settings) affects the shape of the curve while holding the remaining parameters constant. Parameter A indexes the y-axis intercept at $x = 0$. Parameter F captures the x-axis intercept ($\hat{y} > A$); parameters B, C, D, and E are responsible for different growth characteristics of $\hat{y}$ as a function of x before and after reaching the peak of the curve.

The next analytical steps involved (1) *model optimization* (identifying how much cross-condition parameter variation is truly necessary to accurately describe differences between the PH, UP, and PN conditions), and (2) *statistical inferencing* (deriving critical measures from the best model and testing differences between conditions for significance). Both of these steps relied on statistical bootstrapping, as explained in the next section.

Bootstrapping. We used resampling with replacement to simulate the sampling distribution of the cumulative-average probabilities of looks to PH, UP, and PN non-words over time, simultaneously considering participant- and item-related variability in those data. The approach was inspired by Brysbaert and Stevens (2018, p. 9), but tailored to the specifics of the present design.

Each resampling cycle proceeded as follows (see R code in online materials at https://osf.io/bd5c8). To preserve the counterbalancing across the six presentation lists, we randomly sampled (with replacement) eight participant IDs from each list, giving 8 × 6 = 48 participant ID samples. Next, since condition was between-items, we sampled 50 item IDs (again, with replacement) per each of the three non-word conditions, giving 50 × 3 = 150 item ID samples. Combining these samples via their Cartesian product (function `expand.grid()` in R) resulted in 48 × 150 = 7,200 random 'cases' in which some of the original participant-item combinations were over- or underrepresented, respectively (but importantly, without distorting the counterbalancing structure of the original design). For each of these random cases, we extracted the cumulative-average time series data from the original data set, and then calculated the (bootstrapped) mean cumulative-average time series per condition.

The above was repeated 10,000 times for the *model optimization* analysis and – to avoid 'double dipping' – another 10,000 times for *statistical inferencing*.

Model Optimization. Using the Levenberg-Marquardt algorithm (R package `minpack.lm`, Elzhov et al., 2023), we fitted the *Equilibrium Peak* function (**Figure 4**) to each of the 10,000 bootstrapped time series triplets (one time series per condition). We explored a range of different models that were nested within an 18-parameter *Full Model* (each condition was fitted to its own set of six parameters, 3A-3B-3C-3D-3E-3F) and a 6-parameter

Null Model (a single curve fitted through all three conditions, 1A-1B-1C-1D-1E-1F). Across resamples, the *Null Model* achieved a median R^2 of 0.9609 (variance explained by time only) and the *Full Modell* of 0.9995 (variance explained by time and condition); as is logically expected, the *Full Model* always improved on R^2 (min. improvement = 0.0017; max. improvement: 0.1727).

Using a nested backward model reduction approach (for details, see supplemental online materials at https://osf.io/m4zuk), we progressively simplified the *Full Model* to identify the most parsimonious model that performed 'virtually identically' to the *Full Model,* but with fewer parameters. Specifically, we required the selected model to have at least 99% distributional overlap with the *Full Model* in terms of R^2-improvement over the *Null Model* (see **Figure 5** for illustration), while using the smallest number of free-varying parameters possible.

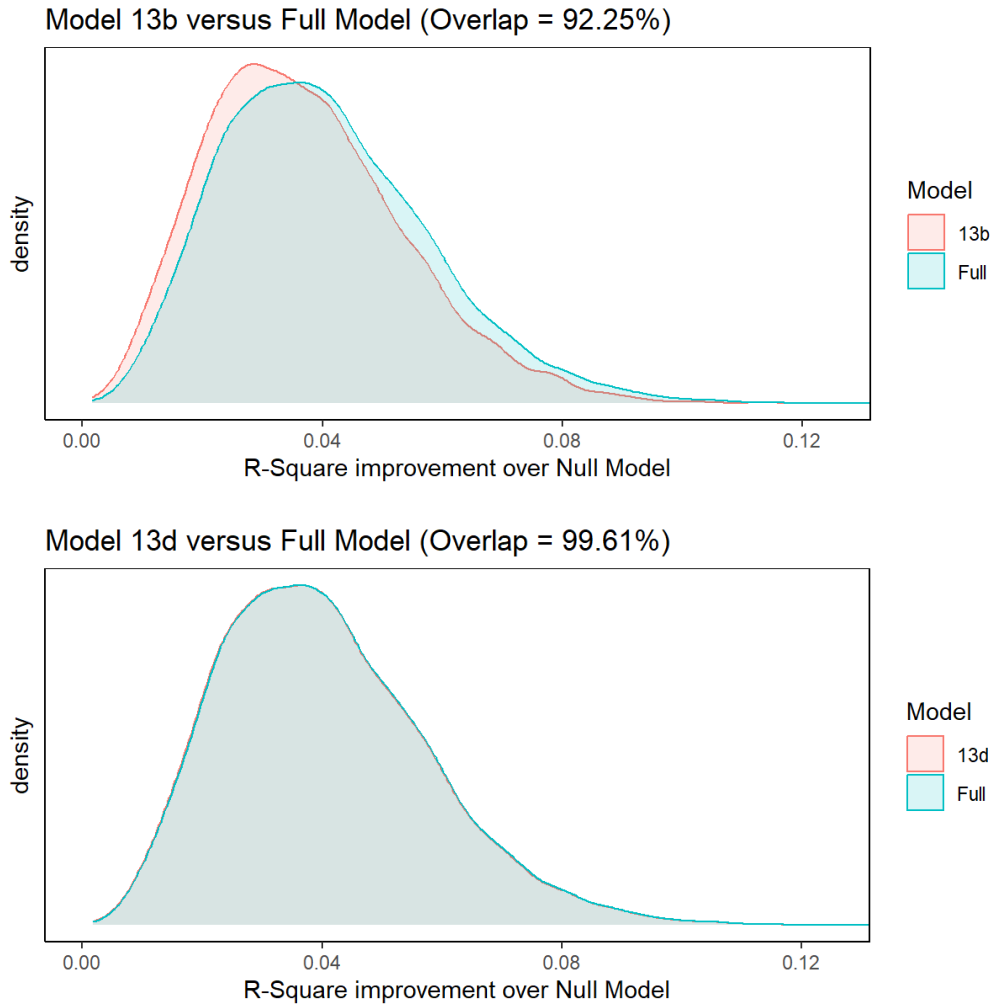

Figure 5. Example bootstrapped sampling distributions for R^2-improvement over the *Null Model* (x-axis values are positive because the *Null Model* always achieved inferior fits). Each plot shows the R^2-improvement distribution for a given model (densities in pink) compared to that of the 18-parameter *Full Model* (densities in turquoise); darker-shaded areas indicate distributional overlap. Of the two examples shown, model *13b* (top panel) performed notably worse than the *Full Model* (92.3% distributional overlap), whereas model *13d* (bottom panel) was almost indistinguishable from the *Full Model* (99.6% distributional overlap).

Of the 25 models tested, two fulfilled the above criteria, namely models *13d* and *13e*. They differed only in whether parameter D or parameter B was allowed to vary freely across all three conditions (potentially due to a trade-off between these two parameters in constraining the shape of the curve). In model *13d*, parameter B was

shared between the PH and PN condition and in model *13e*, it was parameter D that was shared between these two conditions. Apart from this, the two models were structurally identical. Both assumed the A and the F parameter to be shared across all three conditions – this is reassuring because these two parameters constrain the x-and y-axis intercepts at the very beginning of each trial (**Figure 4**), where conditions are not expected to differ. Moreover, both models assumed parameters C and E to vary freely across all three conditions. Model *13d* therefore comprised 13 parameters (1A-2B-3C-3D-3E-1F) and the same was true for model *13e* (1A-3B-3C-2D-3E-1F). Models with fewer parameters achieved notably poorer fits compared to the *Full Model*. Models *13d* and *13e* were therefore selected for statistical inferencing. Both came to virtually identical conclusions (see below).

Inferential Analyses. The selected models (*13d* and *13e*) were applied to cumulative-average time series data from 10,000 new resamples. Across these, both models achieved a median R^2 of 0.9993 (same as in the *model optimization* analysis, in fact). In line with Kunert and Scheepers (2014), we compared the three conditions (PH, UP, and PN) in terms of three composite measures that were derived from the relevant model fits: *Maximum Error Probability*, *Initial Uncertainty Time*, and *Home-In Time* (**Figure 6**).

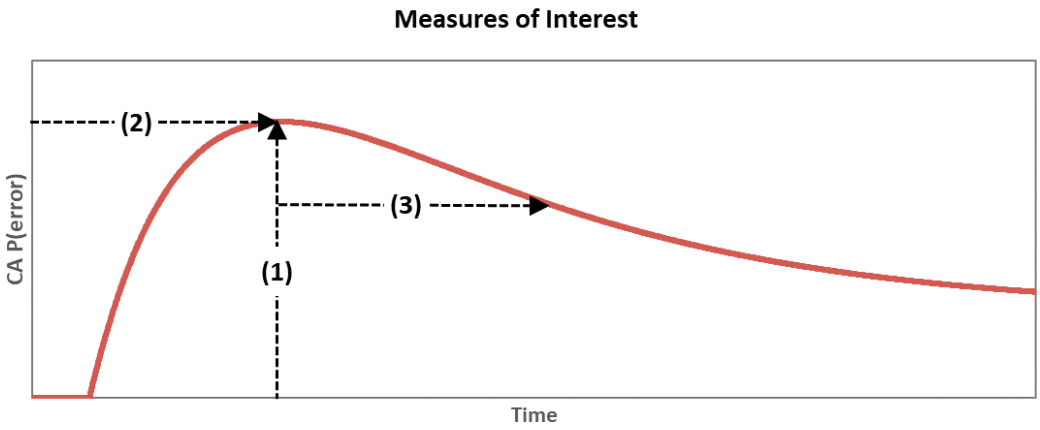

Figure 6. The three composite measures for comparison: (1) *Maximum Error Probability*, (2) *Initial Uncertainty Time*, and (3) *Home-In Time*.

Maximum Error Probability indexes the global peak in the cumulative-average probability of looks to the non-word. Higher values in this measure indicate that the non-word was more likely to be confused with the word in the given task. *Initial Uncertainty Time* measures the time from the onset of the trial until error probability reaches its maximum. As mentioned earlier, perceivers hardly discriminate between the word and the non-word during this time period (**Figure 3**), suggesting uncertainty about the non-word's lexical status. Finally, *Home-In Time* measures the time from having reached *Maximum Error Probability* to the point where error probability has dropped to 70% of the maximum in the right tail of the curve (interpretable as standardised time for homing-in on the actual word in the display).

Across resamples, the three conditions were compared to non-parametrically derive confidence intervals and *p*-values (bias-corrected and accelerated, cf. Efron & Tibshirani, 1994) for the differences in each measure.

Results

Table 4 shows parameter estimates and composite measures, as established by fitting each of the two selected models (*13d* and *13e*) to the actually observed data per condition. Values in brackets refer to bootstrapped standard errors. Notice in particular the difference between the PH and PN conditions in parameter C (see also **Figure 4**), indicating that initial error rate rose faster (and to a higher amplitude) in the PH than in the PN condition. **Figure 7** shows the corresponding plots with model *13d* being fitted to the originally observed data (plots for model *13e* looked nearly identical, and are therefore omitted), and **Table 5** lists two-tailed BCa *p*-values for each pairwise contrast across conditions, separately for each composite measure and model.

Table 4. Parameter estimates (*A* to *F*) and composite measures (*Max. Error Probability*, *Initial Uncertainty Time*, *Home-in Time*), broken down by selected model (*13d* on the left, *13e* on the right) and experimental condition (PH, UP, and PN). Shown are model estimates for the observed cumulative-average error probabilities per condition (see also Figure 7, top). Values in parentheses represent bootstrapped *SE*s per estimate, taking both by-participant and by-item variability into account. The scales for parameters *A* to *F* differ by order of magnitude, as indexed by the square-bracketed scaling factors in the leftmost column. For each parameter and model, the estimate from the mean observed data was always within one *SE* from the mean across the 10,000 bootstrap-models.

	***13d* (1A-2B-3C-3D-3E-1F)**			***13e* (1A-3B-3C-2D-3E-1F)**		
Estimate	PH	UP	PN	PH	UP	PN
A $[10^{-3} \times]$	4.987 (1.336)	4.987 (1.336)	4.987 (1.336)	4.987 (1.336)	4.987 (1.336)	4.987 (1.336)
B $[10^{-1} \times]$	4.378 (0.140)	4.181 (0.187)	4.378 (0.140)	4.396 (0.149)	4.181 (0.187)	4.368 (0.164)
C $[10^{-3} \times]$	5.171 (0.354)	4.673 (0.338)	4.272 (0.285)	5.133 (0.344)	4.673 (0.338)	4.288 (0.303)
D $[10^{-4} \times]$	4.995 (0.584)	5.759 (0.610)	5.099 (0.564)	5.057 (0.512)	5.759 (0.610)	5.057 (0.512)
E $[10^{-4} \times]$	1.827 (0.341)	1.678 (0.270)	2.116 (0.344)	1.867 (0.323)	1.678 (0.270)	2.086 (0.335)
F $[10^{2} \times]$	1.373 (0.242)	1.373 (0.242)	1.373 (0.242)	1.373 (0.242)	1.373 (0.242)	1.373 (0.242)
Max. Error Probability	0.348 (0.007)	0.319 (0.009)	0.336 (0.008)	0.349 (0.008)	0.319 (0.009)	0.336 (0.008)
Initial Uncert. Time (ms)	650 (23)	661 (21)	722 (29)	651 (23)	661 (21)	721 (30)
Home-in Time (ms)	1117 (107)	999 (79)	1207 (117)	1111 (100)	999 (79)	1210 (113)

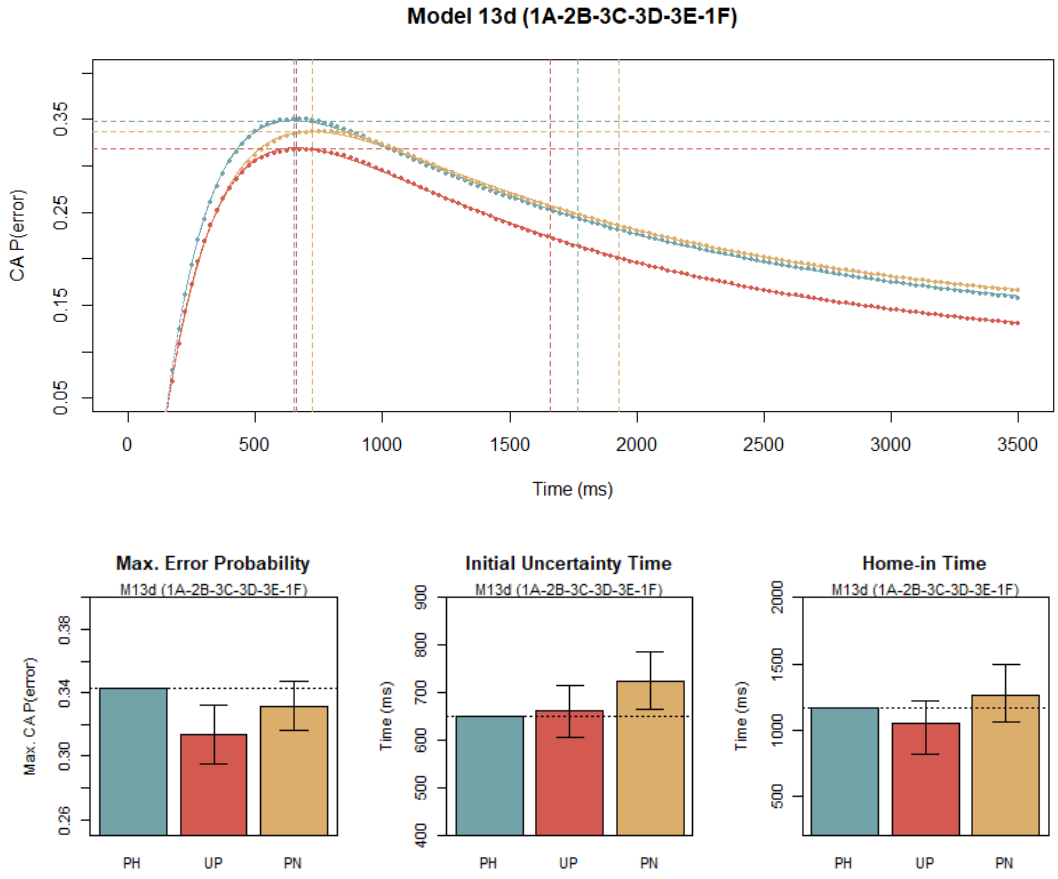

Figure 7. (Top panel) Observed cumulative-average probabilities of looks to the non-word over time by experimental condition (blue: PH, red: UP, amber: PN). Dots represent observed data points per time-slice, and the model fits (model *13d* only) are indicated by solid curves in the relevant colours. Horizontal dashed lines designate *Maximum Error Probabilities* per condition; vertical dashed lines indicate *Initial Uncertainty Time* and (*Initial Uncertainty Time* plus) *Home-In Time* per condition, respectively. **(Bottom panels)** *Maximum Error Probability* (left), *Initial Uncertainty Time* (middle), and *Home-In Time* (right) by levels of experimental condition (PH, UP, and PN). Error bars represent bootstrapped 95% CIs (two-tailed) for contrasts with the PH condition (comparison baseline, as indexed by horizontal dotted lines).

Table 5. Two-tailed BCa *p*-values for each cross-condition contrast (rows) and measure (columns), derived via applying the two selected models (*13d* and *13e*) to each of the 10,000 resamples; asterisks indicate significance at $p < .05$.

	Max. Error Prob.		*Init. Unc. Time*		*Home-In Time*	
	13d	13e	13d	13e	13d	13e
PH vs. UP	.002 *	.003 *	.720	.756	.177	.192
PH vs. PN	.131	.205	.016 *	.021 *	.357	.197
PN vs. UP	.079	.093	.041 *	.044 *	.032 *	.023 *

Discussion

Under the assumption of a two-stage process (*phonological decoding/search* followed by *orthographic verification*), we hypothesised that the lexical status of PN stimuli like FLIS can only be determined relatively late after a prolonged (yet unsuccessful and eventually timed-out) search for matching phonological entries in the mental lexicon. With UP stimuli like NRUG, by contrast, the initial phonological stage would be cut short because such stimuli cannot be phonologically decoded in a meaningful way. In accordance with the task instructions (*keep looking at the word and ignore the non-word*), UP stimuli should therefore be dismissed faster than PN stimuli. PH stimuli like WAWK should also be dismissed faster than PN stimuli, but this time because the former actually possess matching phonological representations in the mental lexicon, such that the phonological search stage can be completed relatively early (by finding an actual match) before the final stage (*orthographic verification*) is initiated.

Experiment 2 largely supported the above hypotheses. Specifically, we found that PN stimuli like FLIS were associated with significantly longer *Initial Uncertainty Times* than

UP stimuli like NRUG or PH stimuli like WAWK (**Figure 7**, **Table 4**, and **Table 5**). A comparable PN vs. UP difference also emerged in *Home-In Time* (again, indicating prolonged processing for PN stimuli), whereas contrasts involving the PH condition were not significant in this measure.

Regarding our second hypothesis, that PH stimuli should initially be more likely to be confused with real words than UP or PN stimuli, the relevant contrasts in **Table 5** (leftmost columns) only partially confirmed what we expected: while *Maximum Error Probability* was numerically highest in the PH condition, only the contrast with the UP condition was significant. However, we can see in **Figure 7** (top) that initial error probability rose faster, and to a higher amplitude, in the PH condition compared to the remaining two conditions. The *Maximum Error Probability* measure probably did not capture this because of its relative insensitivity to cross-condition differences *before* reaching the peak amplitude in error probability. When considering differences in parameter C specifically (**Table 4**), which is more sensitive to such pre-amplitude differences (**Figure 4**), it emerged that that the relevant estimates were indeed significantly higher for the PH condition than for the PN condition, with associated BCa *p*-values (two-tailed) of .008 and .019 in model *13d* and *13e*, respectively. Thus, while results regarding PH stimuli were not as clear-cut as expected, they were still consistent with our predictions.

The two experiments reported so far employed different methodologies (standard lexical decision in Experiment 1; 2AFC eye tracking in Experiment 2) and they both converged on showing a clear *reversal* of the classical pseudohomophone effect. In the discussion to Experiment 1, we conjectured that an important contributing factor to this reversal could be the *material context* in which the non-word stimuli were embedded. The following experiment put this conjecture to the test by using different word materials as fillers.

Experiment 3

The final experiment was a replication of Experiment 1. It employed the exact same lexical decision task, but instead of low-frequency filler words (matched with the PH, PN, and UP stimuli in terms of orthographic features), we now used words that fell into the same lexical frequency range as the PH stimuli's base words (i.e., the words that the PH stimuli were derived from).

Method

Participants

We recruited a new sample of 48 native English speakers aged 19 to 58 years (median = 23 years). About 64% of them were female. All participants were right-handed and had normal or corrected-to-normal vision. Ethical approval for this study was obtained from the BCU Faculty of Business, Law & Social Sciences Research Office. All participants gave written informed consent before taking part.

Materials, Procedure, and Analysis

Experiment 3 differed from Experiment 1 only in terms of the material context in which the non-word stimuli were embedded. We replaced the 150 words from Experiment 1 and 2 with new monosyllabic words that were sampled from roughly the same (i.e., relatively high) lexical frequency band as the base words of the PH stimuli. The new words had an average occurrence of 40.8 per million in the BNC (mean $\log_{10}$ frequency per million = 1.61; *SE* = 0.05). Recall that the word materials in Experiment 1 and 2 were of much lower lexical frequency (only three occurrences per million on average). As before, the new words were matched in numbers of characters with the non-word stimuli. Matching in terms of bigram frequency and/or number of orthographic neighbours was not feasible. The new words, and associated norms for length and lexical frequency, are available at https://osf.io/kqy9g.

In all other respects (non-word stimuli, experimental procedures, analysis), Experiment 3 was identical to Experiment 1.

Results

Excluding trials with RTs below 250 ms or above 2500 ms resulted in less than 1% data loss. The raw descriptive figures per condition are shown in **Table 6**. As would be expected, the high-frequency filler words in Experiment 3 were recognised ca. 10% more accurately and about 60 ms faster than the low-frequency filler words in Experiment 1. UP letter-strings were, again, most easily recognised as non-words. Most strikingly, the contrast between the PH and the PN condition now clearly pointed in the direction of the *classical* pseudohomophone effect.

Table 6. Observed probability of correct responses (Accuracy) and correct RT (in ms) across all valid trials per condition in Experiment 3 (PH = pseudohomophone, UP = unpronounceable non-word, PN = pronounceable non-word). The last row shows corresponding figures for the filler words used (these were of relatively high lexical frequency and were not considered in the inferential analyses). Means and *SE*s are not adjusted for covariate influences and do not take participant or item variability into account.

	Accuracy		Correct RT	
	Mean	SE	Mean	SE
PH	.929	.005	871	9.3
UP	.983	.003	639	5.3
PN	.952	.004	735	6.8
Filler Words	.971	.002	653	3.3

The inferential results for the non-word stimuli are shown in **Table 7**. **Figure 8** plots covariate-adjusted GLMM estimates per condition. Once again, UP stimuli were confirmed to have a very clear speed and accuracy advantage over both PH and PN stimuli. The PH vs. PN contrast in correct RT confirmed a *classical* pseudohomophone effect, with PH stimuli taking significantly longer to reject than PN stimuli. The corresponding contrast in response accuracy was not significant.

When the control predictors (*LBF* and *ON*) and their interactions with experimental condition were removed from the GLMMs, significance patterns stayed the same. The UP condition differed significantly from the PH baseline, both in terms of response accuracy (b = 1.463, SE = 0.323, z = 4.534, $p < 0.001$) and in terms of correct RT (b = –280.60, SE = 5.21, z = –53.879, $p < 0.001$). The PN condition did not differ reliably from the PH baseline in

response accuracy (b = 0.137, SE = 0.289, z = 0.474, p = .636), but very clearly so in correct RT, confirming a *classical* pseudohomophone effect (b = –168.58, SE = 4.78, z = –35.246, p < .001).

Table 7. Fixed effects estimates from the full GLMMs of *response accuracy* (left; corresponding *b*s and *SE*s are in log odds units), and *correct RT* (right; *b*s and *SE*s in millisecond units) in Experiment 3. The first three rows show estimated grand averages (intercepts), estimates for the effect of UP relative to PH, and estimates for the effect of PN relative to PH (the PH condition always served as a comparison baseline). The next two rows show estimated main effects of the two covariates (*LBF* and *ON*), and the last four rows show estimates for the *condition by covariate* interactions (again, relative to the PH baseline). Significant effect parameters (p < .05) are highlighted in bold.

	Accuracy				Correct RT			
Source	*b*	*SE*	*z*	*p*	*b*	*SE*	*z*	*p*
Intercept	4.099	0.180	22.700	< .001	813.88	14.28	56.983	< .001
UP	**1.155**	**0.359**	**3.219**	**.001**	**−281.14**	**21.29**	**−13.206**	**< .001**
PN	−0.167	0.316	−0.530	.596	**−169.42**	**20.13**	**−8.416**	**< .001**
LBF	**−0.307**	**0.132**	**−2.328**	**.020**	**47.22**	**6.27**	**7.533**	**< .001**
ON	**−0.327**	**0.113**	**−2.893**	**.004**	**13.45**	**5.82**	**2.311**	**.021**
UP × *LBF*	**0.935**	**0.346**	**2.699**	**.007**	**−45.83**	**14.93**	**−3.070**	**.002**
UP × *ON*	0.331	0.300	1.119	.263	−14.05	14.79	−0.950	.342
PN × *LBF*	**0.704**	**0.313**	**2.254**	**.024**	**−48.68**	**15.63**	**−3.114**	**.002**
PN × *ON*	0.459	0.240	1.917	.055	**−45.44**	**15.14**	**−3.000**	**.003**

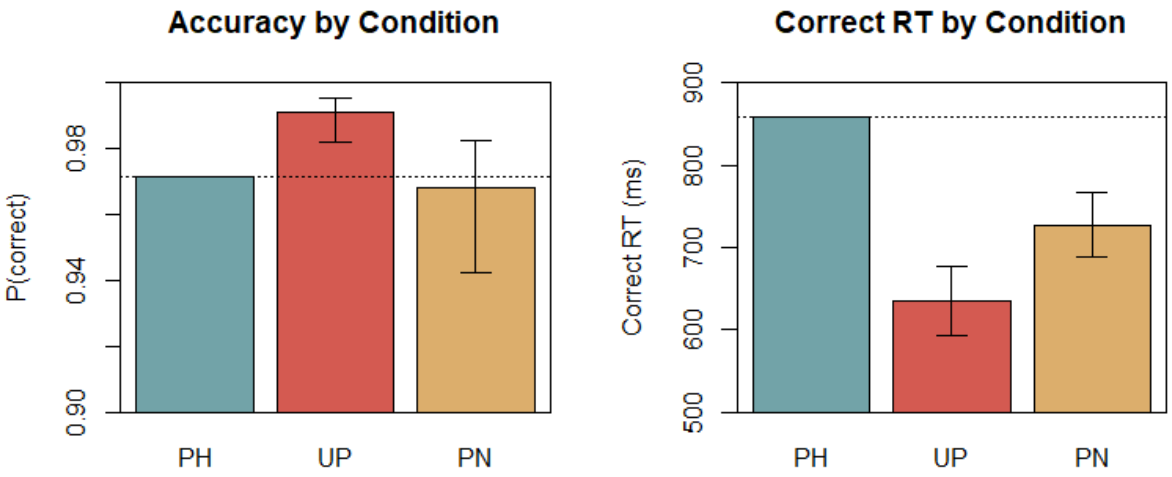

Figure 8. GLMM-estimated (covariate-adjusted) probabilities of correct responses (left) and correct RTs (right) as a function of experimental condition (PH = pseudohomophone, UP = unpronounceable non-word, PN = pronounceable non-word) in Experiment 3. Error bars represent 95% CIs for contrasts with the PH condition (comparison baseline, as indexed by horizontal dotted lines). The asymmetries in the CIs on the left result from converting log odds model predictions into probabilities.

In regards to control predictor influences (*LBF* and *ON*), **Table 7** confirms reliably negative main effects for response accuracy and reliably positive main effects for correct RT, just as in Experiment 1 (cf. **Table 3**). There was, however, a higher number of significant interaction terms compared to the earlier experiment.

Pseudohomophone processing as a function of base-word frequency. In analogy to the complementary analyses for Experiment 1, we also analyzed how the PH stimuli's base-word frequencies affected lexical decision performance in the PH condition. When the control predictors (*LBF* and *ON*) were excluded from analysis, the effect of base word frequency was marginally positive for response accuracy (b = 0.472, SE = 0.251, z = 1.876, p = .061) and marginally negative for correct RT (b = –32.48, SE = 19.14, z = –1.697, p = .090). However,

once influences of the covariates were accounted for, these trends changed direction and were clearly *not* significant (accuracy: $b = –0.085$, $SE = 0.256$, $z = –0.331$, $p = .741$, correct RT: $b = 15.58$, $SE = 19.98$, $z = 0.780$, $p = .435$). Base-word frequency therefore had no measurable impact beyond influences of the control predictors, which is in contrast to what we found in Experiment 1 (cf. **Figure 2**).

Discussion

Experiment 3 established two main findings that contrasted very sharply with the results from the previous two experiments. Firstly, it revealed that PH stimuli like WAWK were much harder to reject as words than PN stimuli like FLIS, consistent with a 'classical' pseudohomophone effect. Secondly, it failed to register any clear impact of the pseudohomophones' base-word frequencies on lexical decision performance in the PH condition. Experiment 3 differed from the previous two experiments in that the non-word stimuli (PH, UP, and PN condition) were tested within the context of high- rather than low frequency words as fillers. Material context obviously matters a lot in regard to the direction of the pseudohomophone effect and whether PH stimuli like WAWK are perceived as misspelled words or not.

General Discussion

We reported three visual lexical decision experiments comparing the processing of pseudohomophones like WAWK (PH condition) with that of unpronounceable non-words like NRUG (UP condition) and pronounceable non-words that are non-homophonic to real words, like FLIS (PN condition). The materials were carefully matched in terms of orthographic features (length, bigram frequency, and number of orthographic neighbours). In Experiment 1 and 2, this was true not only for the critical non-word stimuli, but also for the word materials

used. As a consequence of this matching, the real-word stimuli were from a generally much lower lexical frequency band than the PH stimuli's base words. Under these conditions, we found that the 'classical' pseudohomophone effect, whereby PH stimuli like WAWK are harder to recognise as non-words than PN stimuli like FLIS, was in fact not just neutralised (as in, e.g., Martin, 1981; Taft, 1982; Kunert & Scheepers, 2014) but indeed *reversed*. That is, PH stimuli were *easier* to recognize as non-words than PN stimuli. In Experiment 3, by contrast, we found clear evidence for a classical pseudohomophone effect when the non-word stimuli were tested against high- rather than low frequency words as fillers.

Experiment 1 employed a standard lexical decision task with a single button response per trial. Apart from reversing the 'classical' pseudohomophone effect, another key result from Experiment 1 was that lexical frequencies of the of the PH stimuli's base words (e.g., *walk* in the case of WAWK) were reliably negatively related to correct RTs in the PH condition (**Figure 2**). This suggests that the phonological forms of the PH stimuli activated corresponding word meanings in the mental lexicon, but in a way that *benefited* correct recognition of PH stimuli as non-words rather than causing interference. We hypothesized that PH stimuli were likely to be recognized as *misspelled words*, in the sense that their phonological forms activated related word meanings relatively quickly, and that correct 'non-word' decisions for these stimuli were actually based on an additional *orthographic verification* stage of processing (see also Tiffin-Richards & Schroeder, 2018). In contrast, PN stimuli like FLIS, which are pronounceable but do not possess any matching phonological (or indeed orthographic) forms in the mental lexicon, were likely to be perceived as *potential words that are unfamiliar*, causing longer (and eventually timed-out) searches through the mental lexicon and thus longer periods of indecision. Lastly, UP stimuli like NRUG, which cannot be phonologically decomposed in a meaningful way, were quickly recognized as non-words, causing the shortest RTs.

Based on the findings from Experiment 1, we derived two specific hypotheses with regards to the *time-course* of processing the relevant non-word stimuli. The first was that, compared to UP and PH stimuli, PN stimuli like FLIS should be associated with longer periods of uncertainty about their lexical status (whether the given string is a word or a non-word, respectively). The second was that PH stimuli like WAWK should initially be the most likely to be confused with actual words. These predictions were tested in Experiment 2, which employed the 2AFC eye-tracking paradigm originally proposed by Kunert and Scheepers (2014). Instead of collecting a single button response per trial, this experiment measured participants' eye-movement responses to word vs. non-word stimuli continuously over a relatively long period. Participants were presented with pairs of word vs. non-word letter strings for 3.5 seconds, and their task was to *keep looking at the word and ignore the non-word*. The materials were the same as in Experiment 1, except that they were arranged into word/non-word pairs. Through non-linear modelling of error probabilities (looks to the non-word rather than the word) as a continuous function of time, we found that our two predictions were confirmed: PN stimuli engendered the longest periods of uncertainty about their lexical status, as revealed in significant contrasts with the PH and the UP condition in *Initial Uncertainty Time*. In addition, PH stimuli were initially more likely to be confused with real words, both compared to UP stimuli (significant contrast in *Maximum Error Probability*) and compared to PN stimuli (significant contrast in parameter C of the *Equilibrium Peak* function used for modelling). The pattern of results in Experiment 2 therefore supported our interpretations of the data in Experiment 1.

Our final Experiment 3 replicated Experiment 1, but critically, it employed high frequency rather than low frequency words as fillers. This experiment established that the *classical* pseudohomophone effect likely only holds true when PH, UP, and PN stimuli are

tested within the context of relatively high frequency words as fillers. We also found no evidence for a PH base-word frequency effect in this experiment.

In the discussion of Experiment 1, we already highlighted that the present results are broadly in line with Coltheart et al.'s (2001) DRC model of word recognition. This theory postulates a fast and automatic *lexical* processing route for familiar letter strings such as LIFE, PUSH, MEET, etc., and a slower *phonological* processing route for less familiar letter strings like WAWK, NRUG, and FLIS (non-words used in all our experiments) or indeed QUAY, GAWK, and FEUD (real-word stimuli used in Experiment 1 and 2). We believe that this is key to understanding why we found a *reversal* of the classical psudohomophone effect in Experiment 1 and 2.

The low lexical frequencies of the filler words in Experiment 1 and 2 were likely to have made the phonological targets of the PH stimuli more salient (contributing to their recognition as *misspelled words*) than in *classical* demonstrations of the pseudohomophone effect, including our Experiment 3. Put differently, while classical findings suggested parallel consideration of phonological and orthographic constraints (leading to competition in case of PH stimuli, but no competition in case of PN stimuli), results from Experiment 1 and 2 of the present paper suggested a clear ordering of constraints (phonology before orthography)[2] such that PH stimuli exhibited a measurable processing advantage over PN stimuli in lexical decision.

Assuming adaptive flexibility in the cognitive processing architecture, we propose that the wider material context in which PH and PN stimuli are embedded (specifically,

[2] Alternatively, one could conceptualize this kind of adaptation as giving phonological constraints a higher *weight* relative to orthographical constraints. Indeed, parallel constraint satisfaction architectures (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998) are capable of emulating the behaviour of a serially ordered process in this way.

whether filler words are from a high or low lexical frequency band) acts as an important moderator of the relevant processing strategy, which in turn affects the direction of the pseudohomophone effect. This proposal is strongly supported by the contrasting findings between Experiment 1 and 2 on the one hand and Experiment 3 on the other.

Unfortunately, in most of the older studies, there is hardly any information about the lexical and orthographic features of the real-word fillers used. Some used a balanced set of high- and low-frequency words as fillers (e.g., Ziegler et al., 2001), while in most others, the real-word materials were considered theoretically irrelevant task necessities (e.g., Martin, 1981; Taft, 1982; Underwood et al., 1988; Van Orden et al., 1988; Seidenberg et al., 1996; Borowsky & Masson, 1999; Goswami et al., 2001, but see, e.g., Milota et al., 1997, who provided length and bigram frequency norms on the word materials used). With few exceptions (e.g., Ziegler et al., 2001; Kunert & Scheepers, 2014; Tiffin-Richards & Schroeder, 2018), there is also not much information about the lexical frequencies of the pseudohomophones' base words. This lack of information prohibits a detailed meta-analytical reexamination of previous work.

Our final point concerns the temporal dynamics of the purported context-sensitive adaptation processes. Our hunch is that it takes only a few word vs. non-word trials for participants to adjust their lexical decision strategy in accordance with the materials encountered so far. After mostly encountering low-frequency words (alongside non-words), they are likely to give phonological constraints precedence over orthographic ones (cf., Experiment 1 and 2); after mostly encountering high-frequency words, they are likely to adopt a more competitive lexical decision strategy (cf. Experiment 3). Unfortunately, our present data hardly allow us to explore this kind of adaptation in much detail, mainly because we used standard (i.e., rather unsophisticated) trial-randomization procedures that are not optimized for detecting learning over the course of an experiment. Future research could

make use of smarter randomization techniques (e.g., Liang & Barr, 2024) to investigate the process of adapting to the material context in greater temporal resolution.

Conclusion

In comparison to previous 'classical' demonstrations of the pseudohomophone effect, the present findings indicate an important role of the *material context* in which pseudohomophones like WAWK and comparison non-words like FLIS are embedded. Specifically, when real-word fillers are from a lower lexical frequency band than the pseudohomophones' base words (cf. Experiment 1 and 2), the likelihood of perceiving pseudohomophones as misspelled words increases, leading to a *reversal* of the classical pseudohomophone effect in relation to comparison non-words. In contrast, a classical pseudohomophone effect pattern emerges when using high-frequency words as fillers (cf. Experiment 3).

Our findings have both theoretical and methodological implications. On the theoretical side, they suggest that perceivers can flexibly adapt their lexical decision strategy (specifically, their reliance on phonological vs. orthographic constraints) in accordance with the overall material context. On the methodological side, our results highlight that lexical decision studies on non-word processing should also consider the lexical and orthographic properties of the real-word fillers used. Researchers should at minimum report the lexical frequencies of their real-word materials in the future.

References

Adelman, J. S. Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., Perry, J. R., Lupker, S. J., Forster, K. I., Cortese, M. J., Scaltritti, M., Aschenbrenner, A. J., Coane, J. H., White, L., Yap, M. J., Davis, C., Kim, J., & Davis, C. J. (2014). A behavioral database for masked form priming. *Behavior Research Methods*, 46, 1052-1067. https://doi.org/10.3758/s13428-013-0442-y

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445-459. https://doi.org/10.3758/bf03193014

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

Bates, D., Maechler, M. Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Berent, I., & Perfetti, C. A. (1995). A rose is a REEZ: The two-cycles model of phonology assembly in reading English. *Psychological Review*, 102(1), 146–184. https://doi.org/10.1037/0033-295X.102.1.146

Borowsky, R., & Masson, M. E. J. (1999). Frequency effects and lexical access: On the interpretation of null pseudohomophone base-word frequency effects. *Journal of Experimental Psychology: Human Perception & Performance*, 25, 270-275. 10.1037/0096 https://doi.org/-1523.25.1.270

Borowsky, R., Owen, W. & Masson, M. (2002). Diagnostics of phonological lexical processing: Pseudohomophone naming advantages, disadvantages, and base-word frequency effects. *Memory & Cognition*, 30(6), 969-984. https://doi.org/10.3758/bf03195781

Braun M., Hutzler F., Ziegler J. C., Dambacher M., Jacobs A. M. (2009). Pseudohomophone effects provide evidence of early lexicophonological processing in visual word recognition. *Human Brain Mapping*, 30, 1977-1989. https://doi.org/10.1002/hbm.20643

Briesemeister, B. B., Hofmann, M. J., Tamm, S., Kuchinke, L., Braun, M., & Jacobs, A. M. (2009). The pseudohomophone effect: evidence for an orthography–phonology-conflict. *Neuroscience Letters*, 455, 124-128. https://doi.org/10.1016/j.neulet.2009.03.010

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, 1(1), Article 9, 1-20. https://doi.org/10.5334/joc.10

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

Coltheart, M. (2005). The science of reading: A handbook. *Oxford: Blackwell*.

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. Chapman and Hall/CRC, UK. https://doi.org/10.1201/9780429246593

Elzhov T. V., Mullen, K. M., Spiess, A., & Bolker, B. (2023). `minpack.lm`: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus

Support for Bounds. R package version 1.2-4, https://CRAN.R-project.org/package=minpack.lm.

Forster, K.I., & Forster, J.C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35: 116. https://doi.org/10.3758/BF03195503

Frederiksen, J. R., & Kroll, J. F. (1976). Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance*, 2(3), 361-379. https://doi.org/10.1037/0096-1523.2.3.361

Frost, R. (1998). Toward a strong phonological theory of visual word recognition: True issues and false trails. *Psychological Bulletin*, 123(1), 71-99. https://doi.org/10.1037/0033-2909.123.1.71

Goswami, U., Ziegler, J. C., Dalton, L., & Schneider, W. (2001). Pseudohomophone effects and phonological recoding procedures in reading development in English and German. *Journal of Memory & Language*, 45, 648-664. https://doi.org/10.1006/jmla.2001.2790

Grainger, J., & Jacobs, A.M. (1996). Orthographic processing in visual word recognition: a multiple read-out model. *Psychological Review* 103(3), 518-565. https://doi.org/10.1037/0033-295X.103.3.518

Kunert, R., & Scheepers, C. (2014). Speed and accuracy of dyslexic versus typical word recognition: an eye-movement investigation. *Frontiers in Psychology*, 5: 1129. https://doi.org/10.3389/fpsyg.2014.01129

Liang, J., & Barr, D. J. (2024). Better power by design: Permuted-subblock randomization boosts power in repeated-measures experiments. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000717

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, 6: 1171. https://doi.org/10.3389/fpsyg.2015.01171

Lupker, S. & Pexman, P. (2010). Making things difficult in lexical decision: The impact of pseudohomophones and transposed-letter nonwords on frequency and semantic priming effects. *Journal of Experimental Psychology: Learning, Memory and Cognition* 36(5), 1267-1289. https://doi.org/10.1037/a0020125

Martin, R. C. (1981). The pseudohomophone effect: The role of visual similarity in non-word decisions. *Quarterly Journal of Experimental Psychology*, 34A, 395-409. https://doi.org/10.1080/14640748208400851

McCann, R. S., & Besner, D. (1987). Reading pseudohomophones: Implications for models of pronunciation assembly and the locus of word-frequency effects in naming. *Journal of Experimental Psychology: Human Perception and Performance*, 13(1), 14-24. https://doi.org/10.1037/0096-1523.13.1.14

McCann, R. S., Besner, D., & Davelaar, E. (1988). Word recognition and identification: Do word-frequency effects reflect lexical access? *Journal of Experimental Psychology: Human Perception and Performance*, 14, 693-706. https://doi.org/10.1037/0096-1523.14.4.693

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3), 283-312. https://doi.org/10.1006/jmla.1997.2543

Milota, V., Widau, A., Montie, R., McMickell, M. R., Juola, J., & Simpson, G.B (1997). Strategic reliance on phonological mediation in lexical access. *Memory & Cognition* 25(3), 333-344. https://doi.org/10.3758/bf03211289

Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General*, 118(1), 43–71. https://doi.org/10.1037/0096-3445.118.1.43

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. *URL* http://www.R-project.org/ .

Reynolds, M., & Besner, D. (2005). Basic processes in reading: A critical review of pseudohomophone effects in reading aloud and a new computational account. *Psychonomic Bulletin & Review,* 12(4), 622-646. https://doi.org/10.3758/BF03196752

Seidenberg, M.S. & McClelland, J.M. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review* 96, 523-568. https://doi.org/10.1037/0033-295X.96.4.523

Seidenberg, M. S., Petersen, A., MacDonald, M. C., & Plaut, D. C. (1996). Pseudohomophone effects and models of word recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 22, 48-62. https://doi.org/10.1037/0278-7393.22.1.48

Taft, M. (1982). An alternative to grapheme-phoneme conversion rules? *Memory & Cognition*, 10**,** 465-474. https://doi.org/10.3758/BF03197649

Taft, M., & Russell, B. (1992). Pseudohomophone naming and the word frequency effect. *Quarterly Journal of Experimental Psychology*, 45A, 51-71. https://doi.org/10.1080/14640749208401315

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Tiffin-Richards, S. P., & Schroeder, S. (2018). Verification of nonwords: The baseword frequency effect in children's pseudohomophone reading. *Psychonomic Bulletin & Review*, 25(6), 2289-2294. https://doi.org/10.3758/s13423-017-1424-3

Underwood, G., Roberts, M. & Thomason, H. (1988). Strategical invariance in lexical access: The reappearance of the pseudohomophone effect. *Canadian Journal of Psychology* 42(1), 24-34. https://doi.org/10.1037/h0084175

Van Orden, G.C., Johnston, J.C. & Hale, B.L. (1988) Word identification in reading proceeds from spelling to sound to meaning. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 14(3), 371-386. https://doi.org/10.1037//0278-7393.14.3.371

Van Orden, G. C , Pennington, B. F., & Stone, G. O. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review*, 97, 488-522. https://doi.org/10.1037/0033-295X.97.4.488

Van Orden, G. C. (1991). Phonological mediation is fundamental to reading. In D. Besner & G. Humphreys (Eds.), Basic processes in reading: Visual word recognition (pp. 77-103). Hillsdale, NJ: Erlbaum

Van Orden, G. C., Stone, G. O., Garlington, K. L., Markson, L. R., Pinnt, G. S., Simonfy, C. M., & Brichetto, T. (1992). "Assembled" phonology and reading: A case study in how theoretical perspective shapes empirical investigation. In R. Frost & L. Katz (Eds.), Orthography, phonology, morphology, and meaning (pp. 249-292). Amsterdam: North Holland. https://doi.org/10.1016/S0166-4115(08)62799-5

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5), 2020-2045. https://doi.org/10.1037/xge0000014

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning & Verbal Behavior*, 17(2), 143-154. https://doi.org/10.1016/S0022-5371(78)90110-X

Ziegler, J. C., Jacobs, A. M., & Klüppel, D. (2001). Pseudohomophone effects in lexical decision: Still a challenge for current word recognition models. *Journal of Experimental Psychology Human Perception & Performance,* 27, 547-559. https://doi.org/10.1037/0096-1523.27.3.547