UNIVERSITY of CALIFORNIA PRESS | Collabra

Schmalz, X., et al. (2017). German and English bodies: No evidence for cross-linguistic differences in preferred orthographic grain size. *Collabra: Psychology,* 3(1): 5, DOI: https://doi.org/10.1525/collabra.72

ORIGINAL RESEARCH REPORT

German and English Bodies: No Evidence for Cross-Linguistic Differences in Preferred Orthographic Grain Size

Xenia Schmalz[*,†], Serje Robidoux[*], Anne Castles[*], Max Coltheart[*] and Eva Marinus[*]

Previous studies have found that words and nonwords with many body neighbours (i.e., words with the same orthographic body, e.g., *cat, brat, at*) are read faster than items with fewer body neighbours. This body-N effect has been explored in the context of cross-linguistic differences in reading where it has been reported that the size of the effect differs as a function of orthographic depth: readers of English, a deep orthography, show stronger facilitation than readers of German, a shallow orthography. Such findings support the psycholinguistic grain size theory, which proposes that readers of English rely on large orthographic units to reduce ambiguity of print-to-speech correspondences in their orthography. Here we re-examine the evidence for this pattern and find that there is no reliable evidence for such a cross-linguistic difference. Re-analysis of a key study (Ziegler et al., 2001), analysis of data from the English Lexicon Project (Balota et al., 2007), and a large-scale analysis of nine new experiments all support this conclusion. Using Bayesian analysis techniques, we find little evidence of the body-N effect in most tasks and conditions. Where we do find evidence for a body-N effect (lexical decision for nonwords), we find evidence against an interaction with language.

Keywords: Psycholinguistic grain size theory; failure to replicate; body-rime correspondences; sublexical processing; Bayes Factor

1. Theories of reading across languages

While the majority of research on reading has traditionally come from English-speaking countries (Share, 2008), a small body of important research has moved beyond this anglocentricity, and towards theories and models that can be generalised to orthographies other than English. A question that has attracted a great deal of attention is the way in which *orthographic depth* affects reading processes (Katz & Frost, 1992; Schmalz, Marinus, Coltheart, & Castles, 2015; Ziegler & Goswami, 2005). Orthographic depth, broadly speaking, can be defined as the degree of ambiguity in the relationship between print and speech, which varies across languages. In shallow orthographies, such as Finnish, the relationship between each grapheme and phoneme is simple and predictable, whereas in deep orthographies, such as English, knowledge of complex conversion rules and whole words is needed to achieve high accuracy in reading aloud.

The major problem for children learning to read in a deep orthography is deriving the pronunciation of unfamiliar words, because the sublexical information of deep orthographies is, by definition, incomplete, inconsistent, and/or complex (Katz & Frost, 1992). The psycholinguistic grain size theory (Ziegler & Goswami, 2005) proposes one possible solution to this problem for the reader: the ambiguity associated with sublexical information can be reduced by relying on larger sublexical units and print-to-speech correspondences. In the case of English, linguistic analyses have shown that reliance on bodies, which consist of the vowel and coda of a monosyllabic word, reduces the unpredictability of vowel pronunciation (Peereman & Content, 1998; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995). For example, the word "talk" cannot be read aloud correctly using grapheme-phoneme correspondences (which would predict the pronunciation /tælk/), but can be read aloud correctly based on the body-rime correspondence "-alk" → /o:k/, as in "walk" and "stalk". As a result, the psycholinguistic grain size theory proposes that readers of deep orthographies such as English develop routine reliance on larger units. In contrast, readers of shallow orthographies can rely on smaller units, such as letters or graphemes, and still achieve high reading accuracy.

This main claim of the psycholinguistic grain size theory is intuitively very appealing. It has been a highly influential theory for explaining the results of cross-linguistic studies, with over 1000 citations of the Ziegler and Goswami (2005) review paper (Google Scholar; see Goodwin, August, & Calderon, 2015; Rau, Moll, Snowling, & Landerl, 2015, for some recent examples). The theory depends

* Department of Cognitive Science, ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, AU

† Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università degli Studi di Padova, IT

Corresponding author: Xenia Schmalz (xenia.schmalz@gmail.com)

critically on the assumption that there is an interaction between orthographic depth and reliance on large sublexical units, as we will discuss below. Although several studies have provided evidence for this claim (discussed in Section 1.2. below), it is important in psychological science to make sure that experimental findings are replicable (Earp & Trafimow, 2015; Ioannidis, 2005). Specifically, the current paper was motivated by several failures in our lab to find evidence for differential reliance on bodies in cross-linguistic comparisons of English and German readers (see analyses and results of Section 3).

Our main aim in this paper is to determine to what extent the existing evidence for the psycholinguistic grain size theory is compatible with the view that there is no cross-linguistic difference in the reliance on bodies, over the alternative hypothesis of a real difference. If the existing and new evidence do not support the main prediction of the psycholinguistic grain size theory, one needs to reconsider whether there are any alternative predictions that could be used to support the psycholinguistic grain size theory, or whether other theories of reading across languages have stronger explanatory power given the available data (see Section 4.3).

1.1. What counts as evidence for the psycholinguistic grain size theory?

Before evaluating the existing evidence for the psycholinguistic grain size theory, it is important to consider what kind of evidence can directly support it. Its most explicit prediction is that the deeper the orthography of a language, the more its readers should rely on sublexical units that are larger than letters or graphemes. Here, we define a sublexical orthographic unit as one that is not directly linked to lexical or semantic information (i.e., whole words and morphemes do not count as sublexical units). As a reflection of the importance of this prediction, all five studies reporting evidence for the psycholinguistic grain size theory include a manipulation to measure the reliance on bodies (discussed in detail in Section 1.2; Goswami, Gombert, & de Barrera, 1998; Goswami, Porpodas, & Wheelwright, 1997; Goswami, Ziegler, Dalton, & Schneider, 2003; Ziegler, Perry, Jacobs, & Braun, 2001; Ziegler, Perry, Ma-Wyatt, Ladner, & Schulte-Körne, 2003).

In addition to the sublexical-unit-size manipulations, Ziegler and colleagues interpret stronger length effects in German than English as support for the psycholinguistic grain size theory (Ziegler et al., 2001; Ziegler et al., 2003). Length effects reflect the finding that words or nonwords with more letters are processed more slowly than words or nonwords with fewer letters (New, Ferrand, Pallier, & Brysbaert, 2006; Weekes, 1997). Such effects are proposed to be a marker of sublexical decoding using small units, as the number of letters should matter if the system relies on a letter-by-letter processing strategy (Weekes, 1997). Given that the psycholinguistic grain size theory predicts that readers of shallow orthographies rely to a lesser extent on large units (such as bodies) and to a greater extent on small units (such as letters or graphemes), an increased length effect in a shallow orthography is consistent with the psycholinguistic grain size theory. However, this prediction is shared with another theory of reading across languages, namely the *orthographic depth hypothesis* (Katz & Frost, 1992). According to this hypothesis, the nature of the sublexical correspondences in deep orthographies, by definition, impedes the process of sublexical decoding. This leads to relatively greater reliance on lexical processes in deep compared to shallow orthographies. Consequently, readers of shallow orthographies should exhibit relatively stronger reliance on sublexical processing, which would also manifest as stronger length effects in shallow than deep orthographies.

To test the prediction that there is stronger reliance on lexical than sublexical processes for deeper compared to shallow orthographies, one can use the frequency effect as a marker of lexical processing (Frost, 1994; Frost, Katz, & Bentin, 1987; Schmalz, Beyersmann, Cavalli, & Marinus, 2016). Words with a high frequency are typically reported to be read faster than words with a low frequency. This is proposed to reflect a lower activation threshold for lexical entries of high-frequency words (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). Stronger reliance on lexical processing in deep compared to shallow orthographies, again, is a shared prediction of the orthographic depth hypothesis and the psycholinguistic grain size theory. According to the former, this should be driven by the slow-down of the sublexical route in deep orthographies. According to the latter, this would reflect the general notion that readers of deep orthographies rely on larger units (with whole words listed at the top of Ziegler and Goswami's proposed hierarchy; see their Figure 1).

In sum, stronger length effects for shallow than deep orthographies and stronger frequency effects for deep than shallow orthographies are consistent with the psycholinguistic grain size theory. However, by themselves these two marker effects cannot provide specific evidence for this theory, because the predictions are shared with the orthographic depth hypothesis. The only evidence that specifically supports the psycholinguistic grain size theory is the existence of cross-linguistic differences in the reliance on sublexical units of different sizes. Therefore, in the current paper we focus on this prediction.

In addition, studies providing evidence for the psycholinguistic grain size theory should exclude the possibility that correlated variables account for any cross-linguistic differences in reading (Cutler, 1981; Marinus, Nation, & de Jong, 2015). Generally speaking, potential confounds in psycholinguistic research can be associated either with language-level or participant-level factors. This issue is especially pertinent to cross-linguistic research, because languages tend to differ from each other on many aspects, and therefore it is often unclear to what language-level difference a cross-linguistic difference should be attributed (Schmalz et al., 2015). We discuss potential confounds that could provide alternative explanations for previous observations of differential reliance on bodies across languages in Section 1.2.

Observed cross-linguistic differences should also be considered in relation to potential participant-level confounds. In the case of reading, there are systematic differences as a function of orthographic depth in the

instruction methods that are used to teach children to read (Landerl, 2000; Wimmer & Goswami, 1994): In deep orthographies, such as English, whole-word instruction methods tend to be more popular, because of the assumption that teaching print-to-speech correspondences does not help with reading if these are unreliable. This is relevant to the psycholinguistic grain size theory: a previous study has shown that adults who had received whole-word reading instruction at school relied to a greater extent on bodies in a nonword reading aloud task than adults who had received phonics instruction (Thompson, Connelly, Fletcher-Flinn, & Hodson, 2009).

1.2. The current evidence for the psycholinguistic gain size theory

Taking into account the issues discussed above, we provide an overview of the existing studies on the psycholinguistic grain size theory, and consider whether these show evidence of cross-linguistic differences that can be unequivocally attributed to orthographic depth. The behavioural evidence supporting the view that readers of deep orthographies rely to a greater extent on large sublexical clusters than readers of shallow orthographies comes from two different psycholinguistic manipulations. In the first set of studies, participants are asked to read aloud nonwords with either an existing or a non-existing body (e.g., "dake", a body neighbour of "cake", or "daik", which has a unique body). In the second set of studies, reading aloud latencies are compared for both words and nonwords with many versus few body neighbours. In both manipulations, the idea is that bodies that occur frequently have stronger psychological salience than bodies that occur rarely or do not occur at all. Therefore, if readers routinely rely on bodies, they should show facilitation associated with body existence or frequency. If readers instead routinely rely on letters or graphemes, they should show less or no facilitation associated with body existence or frequency.

1.2.1. Body-existence studies

The psycholinguistic grain size theory predicts stronger facilitation associated with body-existence (i.e., faster response latencies for "dake" than "daik") in deep compared to shallow orthographies, because readers of deep orthographies should rely on bodies to a greater extent (Goswami et al., 1998). Three studies have been conducted to explicitly address this hypothesis, which compare reading of nonwords with existing versus non-existing orthographic clusters across languages.

The first is a cross-linguistic study of English, French, and Spanish (Goswami et al., 1998). Here, the authors found stronger body-existence effects in children as an increasing function of orthographic depth, both in accuracy and reading aloud latencies. However, for the accuracy analyses they did not take into account cross-linguistic differences in overall reading accuracy. Children learning to read in deep orthographies lag in their reading ability behind children learning to read in shallower orthographies (e.g., Frith, Wimmer, & Landerl, 1998; Seymour, Aro, & Erskine, 2003). This can create the illusion of interactions that are due to smaller absolute effects driven by lower average error rates or RTs in shallow orthographies (Faust, Balota, Spieler, & Ferraro, 1999). For example, in the accuracy analyses in Experiment 3 of Goswami et al. (1998), for monosyllabic words, there was no body-N effect on accuracy for Spanish children (with average accuracy rates of 95% across all age groups and conditions), while French and English children showed a significantly bigger body-N effect (for French, averaged across age groups, the accuracy rates were 78.0% for existing-body items and 64.9% for non-existing body items, and for English, 51.4% for existing-body items, and 31.9% for non-existing-body items). In fact, such overadditivity (apparent interaction due to lower overall accuracy or reaction times in one group compared to the other) could provide an alternative explanation for all accuracy analyses reported by Goswami et al. (1998).

The potential for false overadditivity is acknowledged in Goswami et al's analysis of the latency data: In both Experiments 1 and 3, they performed follow-up latency analyses including only a subset of children who were matched, across languages, on their overall reading speed, and failed to find evidence for a cross-linguistic difference in the size of the body-existence effect: in Experiment 1, they report a two-tailed p-value of 0.08, and in Experiment 3, $p = 0.12$. Thus, it cannot be concluded from this study that there are cross-linguistic differences that are attributable to orthographic depth rather than overall differences in reading accuracy and speed.

In a similar study using a Greek/English comparison, Goswami et al. (1997) found support for stronger reliance on the rhymes of bi- and trisyllabic nonwords in English compared to Greek children. Here, follow-up analyses of matched subgroups were reported for all critical comparisons, and the cross-linguistic difference persisted even when the children were matched across languages in their overall accuracy and speed. However, the Greek nonwords with non-existing bodies in this experiment had near-identical orthographic patterns across all items. Thus, even if the children were not familiar with these larger grain-sizes from their knowledge of the Greek orthography, these units would have become familiar to them after a few trials of the experiment. This was not the case for English, as there was no repetition of orthographic clusters within the English item set. Thus, while the Greek participants may have learned the non-existing rhymes due to the repetition, the English-speaking participants did not have this opportunity. This confound might lead to the interaction with stronger apparent reliance on rhymes in English than Greek, which is not related to orthographic depth. Therefore, it cannot be concluded from this study that the cross-linguistic differences are attributable to orthographic depth rather than the item characteristics.

The third cross-linguistic study on the body-existence effect was conducted by Goswami et al. (2003). The focus of this study was the interaction between language and blocking (i.e., whether nonwords with existing or non-existing bodies are read differently depending on whether they are presented in separate or mixed blocks). The authors addressed this question by performing a 2 (language) × 3 (age group) × 2 (blocked versus mixed presentation) × 2 (body existence) × 3 (number of syllables) ANOVA

on the accuracy rates. They report and analyse only accuracy data. For the current purposes, the critical effect is the two-way interaction between language and body-existence. This two-way interaction was not significant. However, there were significant three-way interactions that included this contrast: (1) a body-existence by language by blocking interaction, and (2) a body-existence by language by age interaction. The authors did not perform a post-hoc analysis to test under which circumstances the body-existence by language interaction emerged (because it was not relevant to their aims), but an inspection of the condition means suggests that the latter interaction may well have been driven mainly by German older children approaching ceiling accuracy for both types of nonwords. The three-way interaction between body-existence, language, and blocking reflects a stronger blocking effect for English than German children, but only for nonwords with existing bodies.

These results are difficult to interpret, because the critical two-way interaction was not significant, and the three-way interactions were not predicted *a priori*. A five-way ANOVA tests at least 15 contrasts. Due to this multiple comparison problem, there is an increased chance that a statistically significant interaction reflects a false positive (Cramer et al., 2015), especially if it was not predicted *a priori*. The study of Goswami et al. (2003) also suffered from a lack of power with only 9–13 participants per cell (as language, age, and blocking condition were between-participant factors). Underpowered studies that report significant results that were not predicted *a priori* are more likely to be false positives than the conventional 5%-rate (Button et al., 2013; Christley, 2010; Ioannidis, 2005; Royall, 1986). Thus, the results of this study, like the other two studies discussed above, do not provide convincing evidence for cross-linguistic difference in the reliance on bodies that could be attributed to orthographic depth.

1.2.2. Studies on the body-N effect

The second set of studies that report support for the psycholinguistic grain size theory manipulated body-N. Body-N for a given letter string is defined as the number of words that have the same body. The word *jazz* and the corresponding nonword *blazz* have a body-N value of 1, because *jazz* is the only word with this body; the word *blue* and the nonword *crue* have a body-N count of eight, with body neighbours such *true*, *cue*, and *clue*. In single-word reading aloud, words and nonwords with many body neighbours are generally processed faster than words with fewer body neighbours (Ziegler et al., 2001; Ziegler et al., 2003). In lexical decision, a high body-N count has been shown to facilitate the processing of words, while no effect has been found for nonwords (Ziegler & Perry, 1998). From the point of view of the psycholinguistic grain size theory, the body-N effect reflects sublexical processing of larger-than-grapheme units. According to an alternative view, the facilitatory body-N effect in reading aloud and in lexical decision for words could also reflect facilitation through the lexical activation of body neighbours (Forster & Taft, 1994; Goswami, 1993). In this case, lexical decisions to nonwords may be impaired by a high body-N count, because lexical activation will erroneously bias the reader towards a "yes"-response.

In two cross-linguistic studies, Ziegler and colleagues compared reading aloud latencies for words and nonwords which had either many or few body neighbours, in English and German adults (Ziegler et al., 2001) and children (Ziegler et al., 2003). As predicted by the psycholinguistic grain size theory, English participants showed a stronger body-N facilitation effect than German readers (and German participants showed a stronger length effect than English readers).

However, the items used in both studies contained a confound that undermines the conclusion that the results support the psycholinguistic grain size theory: namely, the body-N manipulation was significantly stronger for English than German. For the German items, the mean body-N counts were 8.89 (SD = 3.82) and 3.82 (SD = 1.79) for the high and low body-N items, respectively; for English, the corresponding values are 12.55 (SD = 4.41) and 3.33 (SD = 1.79). As a result, the strength of the manipulation was larger in English than German. We performed linear model analysis to assess whether the body-N manipulation differed significantly as a function of language. We used body-N condition and language (both contrast-coded as –0.5 and 0.5) as the independent variables and body-N, as a continuous measure, as the dependent variable. This analysis showed a main effect of language, with an overall higher body-N count in English than German, $t = -7.3$, $p < 0.0001$, a main effect of body-N condition, $t = 21.4$, $p < 0.0001$, and crucially, a body-N condition by language interaction, reflecting the stronger manipulation for English than German, $t = 4.5$, $p < 0.0001$. This stronger manipulation for English than German provides a possible alternative explanation for the stronger body-N effect in English than German.

In the Ziegler et al. (2001, 2003) studies, there was also a strong correlation between body-N and orthographic N ($r = 0.44$, $p < 0.0001$). Consequently, as was the case for body-N, the orthographic N manipulation was significantly stronger for English than for German. We confirmed this in a linear model analysis, as above, with language and body-N condition and their interaction as independent variable and orthographic N as a dependent variable. Language and body-N condition interacted, $t = 2.2$, $p = 0.0266$. Orthographic N is the number of words that can be created from a letter string by substituting a single letter (Coltheart, Davelaar, Jonasson, & Besner, 1977). Orthographic N has been shown to affect reading latencies (see Andrews, 1997, for a review), and the size of this effect differs across orthographies (Marinus et al., 2015). Body-N and orthographic N are conceptually different but highly correlated, therefore an item set failing to de-correlate these two concepts needs to consider the possibility that a body-N effect, instead, reflects an effect of orthographic N.[1] Orthographic N reflects the degree of interference or facilitation of similar words in the orthographic lexicon, rather than the psychological salience of a specific orthographic unit (Andrews, 1989, 1992; Coltheart et al., 1977).

As lexical processing has been proposed to be more important for English than shallower orthographies (Katz & Frost, 1992), a larger 'body-N' effect for English than German in the item set of Ziegler and colleagues (Ziegler et al., 2001; Ziegler et al., 2003) could also be a larger orthographic N effect, reflecting stronger reliance on lexical processing in English compared to German readers.

In Ziegler et al.'s (2003) study of developing readers, the confound with orthographic N is not addressed. However, in their study with adult readers, Ziegler et al. (2001) performed a follow-up analysis which included orthographic N as a covariate, but they did not report having tested for the presence of an interaction between language and body-N. Instead, they tested the body-N effect separately for each language, finding a significant effect for English but not German. However, this pattern of findings does not constitute evidence for an interaction (Gelman & Stern, 2006). The possibility remains, therefore, that there are no cross-linguistic differences in the size of the body-N effect once orthographic N is controlled for.

2. Evaluating the empirical evidence

Given the questions arising as a result of our analysis of the original studies, our aim in this section is to evaluate all available evidence on the body-N effect across languages varying in orthographic depth, using a combination of analytic approaches to examine the question thoroughly. First, we reanalyse Ziegler et al.'s (2001) data, as well as the data for their English words from the English Lexicon Project (ELP) database (Balota et al., 2007). We use linear models, which allows us to treat body-N as a continuous rather than a dichotomised variable (as demanded by within-participants ANOVA). This allows us to take into account the cross-linguistic difference in the strength of the manipulation, and also increases the power of the statistical analyses.

In a re-analysis of the original data of Ziegler et al. (2001), we aim to assess the evidence for an interaction between body-N and language while using body-N as a continuous variable (thus removing the confound of a stronger manipulation for English). If we continue to find evidence for this interaction, this would suggest that there is a cross-linguistic difference that might be attributed to orthographic depth. The trial-level data of the original study by Ziegler et al. (2001) has been lost (J. Ziegler, personal communication, 2 September, 2014), therefore we relied on the item-level data (i.e., data which has been averaged across subjects for each item) reported by Perry and Ziegler (2002). In addition, we aim to assess whether the effect is generalisable to other participants. To this end, we retrieved the trial-level data (i.e., RT data which have not been averaged across items or participants) from the English Lexicon Project (Balota et al., 2007).

2.1. Re-analysis of the body-N effect in Ziegler et al. (2001)

2.1.1. Multiple Regression

Using the item-level data provided in Perry and Ziegler (2002), we conducted a multiple regression analysis with body-N as a continuous variable. A model including body-N, lexicality, language, and their interactions as predictors, and RT as the dependent variable, showed a significant effect of lexicality, $t = -5.96$, $p < 0.0001$, with faster responses for words than nonwords. All other ps were greater than 0.1. Note that the effect of body-N was not significant, $t = -1.53$, $p = 0.13$, nor was the body-N by language interaction, $t = 1.34$, $p = 0.18$.

2.1.2 Bayes Factors

We performed the same analyses with an alternative inference method, namely Bayes Factors (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Bayes Factors quantify the degree to which the observed data are compatible with a pre-specified alternative hypothesis or the null hypothesis of no effect. Thus, a Bayes Factor can also provide evidence for a null effect, which is theoretically impossible with conventional frequentist testing (Dienes, 2014; Rouder et al., 2009). For all of the Bayesian analyses reported throughout the paper, we used the R package BayesFactor Version 0.9.12-2 and its default settings to calculate Bayes Factor values (BFs; Morey & Rouder, 2014). The Bayes Factor provides a continuous measure, with decreasing values below 1 providing increasingly stronger evidence for the null, and increasing values above 1 providing increasingly stronger evidence for the alternative hypothesis. For easier interpretability, we use a set of guidelines, as recommended by Rouder et al. (2009): Bayes Factor values between 3 and 1/3 provide equivocal evidence for or against the alternative hypothesis, respectively; Bayes Factor values greater than 3 (or smaller than 1/3) provide some evidence for (or against) the alternative hypothesis, and values greater than 10 (or smaller than 1/10) provide strong evidence.

The first model comparison tests for any influence of body-N, by comparing a 'full' model to a 'base' model, which excludes the main effect of body-N and any interactions with this term. Using Ziegler et al.'s (2001) item-level data, we compared a full model, identical to the LME model reported in the previous section, to the 'base' model. We obtained very strong evidence against the full model, BF = 0.02 (±4.14%).

In the second comparison, we tested for a main effect of body-N. Here, we compared a main-effects model to a main-effects model that excluded the main effect of body-N. This provided equivocal evidence against the presence of the body-N effect, BF = 0.44 (±3.73%). In a third comparison, we assessed the evidence for a body-N by language interaction. We compared the full model to an identical model which excluded the interaction of body-N and language, while retaining their main effects. For a body-N by language interaction, BF = 0.73 (±8.92%), providing equivocal evidence against it. Thus, in Bayesian terms, Ziegler et al.'s (2001) study cannot distinguish between the presence or absence of a body-N by language interaction, or whether there is substantial evidence for a main effect of body-N, although the first comparison suggests that a model excluding the main effect and interactions of body-N fares better than the full model.

2.1.3. Ziegler et al.'s (2001) items in the English Lexicon Project

An alternative question is whether a body-N effect can be found using the same items as Ziegler et al. (2001) but a different set of participants. Firstly, this will allow us to assess to what extent the findings of Ziegler et al. (2001) are generalisable across samples. Secondly, using the English Lexicon Project (Balota et al., 2007) allows us to perform the analyses on a trial-level item set, using linear mixed effect (LME) models (Baayen, Davidson, & Bates, 2008). LMEs are commonly used in psycholinguistic research, as they can simultaneously fit both participant- and item-level variance in the random effects structure. They also provide more power by using the information available from every trial rather than averaged data across participants or items, as is typically done in ANOVA approaches. The analyses were done in R (Version 3.1.1., R Core Team, 2013), using the packages lme4 1.1-12 and lmerTest 2.0-32.

We retrieved the trial-level data for Ziegler et al.'s English words from the English Lexicon Project (Balota et al., 2007). This corpus contains reading aloud latencies for 79 of the original 80 words. The trial-level dataset contained 2309 correct RTs, with an average of 29 observations per item. After removing data points with RTs < 300 ms or > 1800 ms (which resulted in an approximately normal distribution of the data, as shown by a qq-plot), we were left with a total of 2268 data observations. These data were fit with an LME model including random intercepts for both participants and items, as well as a fixed effect of body-N as a continuous predictor. The dependent variable was inverse RTs. This analysis did not show a significant body-N effect, $t = -0.43$, $p = 0.67$. For the Bayes Factor analysis, we compared the LME model against one which was identical except that it excluded the main effect of body-N. This comparison provided evidence against the presence of a body-N effect BF = 0.15 (±3.49%).

This set of analyses does not directly answer the question of whether there is a cross-linguistic difference in the size of the body-N effect. However, we found evidence against a main effect of body-N in English, using Ziegler's item set and data from the English Lexicon Project. As the interaction reported by Ziegler et al. (2001) was driven by a stronger body-N effect in English than German, the absence of a body-N effect in English in incompatible with the prediction of a smaller effect in German.

3. Large-scale analysis

The analyses above, at the very least, suggest that the body-N effect is not very stable. However, we did not find evidence against a body-N by language interaction. Given the results so far, it could be argued that a body-N effect exists in the population, but is very small. If the effect is small, the probability of reliably detecting this main effect – let alone an interaction involving this effect – in a typically-sized psycholinguistic experiment is also small (Button et al., 2013; Cohen, 1962; Vadillo, Konstantinidis, & Shanks, 2016). This issue can be addressed by conducting a high-powered study, or alternatively, by combining the data from multiple studies, if their design is sufficiently similar to allow this (Schmidt, 1992, 1996). We took the latter approach, as we have accumulated numerous experiments on the body-N effect both in English and in German (described in a later section and in the Appendix). The experiments were conducted with various *a priori* aims, which are described, along with their individual analyses and results, in the Supplementary Materials, downloadable from https://osf.io/myfk3/.

The main advantage of combining data from numerous studies is increased statistical power, compared to smaller-scale studies. In frequentist terms, this maximises our chance of finding significant body-N effects and body-N by language interactions. In Bayesian terms, larger studies tend to have stronger evidential value, which is quantified by more extreme BF values (i.e., large numbers when an effect is present or small numbers when the effect is absent, whereas for studies with low evidential value, the BF values hover around 1). More extreme BF values increase the confidence in the results.

As we had access to all trial-level data, we were able to perform LME analyses, which further increased our statistical power. Large-scale analyses allow us to statistically control for a number of covariates (Kliegl, Grabner, Rolfs, & Engbert, 2004; Yap, Balota, Sibley, & Ratcliff, 2012). While strong inter-correlation between the independent variables is a problem even in a large-scale analysis, the data presented here are drawn from individual studies, where the items were matched across body-N conditions on variables such as orthographic N, length, and frequency. This reduces the problem of multicollinearity compared to a large-scale analysis of an unselected or exhaustive set of items (Protopapas & Kapnoula, 2013).

In addition to the data collected in our lab, we attempted to obtain published or unpublished trial-level data on the body-N effect from other labs. This allowed us to add a study by M. Taft (unpublished; personal communication, 3 September, 2014). Two studies on the body-N effect in adults by Ziegler and colleagues could not be included, because there was no available trial-level data. These studies were a lexical decision experiment by Ziegler and Perry (1998; J. Ziegler, personal communication, 23 January, 2013), and the study described above by Ziegler et al. (2001). We were also unable to address the question of whether a body-N effect, and its interaction with language, might be more convincing in children. There is, to date, only one study of body-N effects in children (Ziegler et al., 2003), and we could not obtain either trial- or item-level data for this study. The studies by Goswami and colleagues also do not report data that could be used in a re-analysis of the body-existence effect across languages (Goswami et al., 1998; Goswami et al., 1997; Goswami et al., 2003).

3.1. Method

3.1.1. Studies included in the analyses

We analysed all skilled adult reader studies with available trial-level data, which used either single word reading aloud or lexical decision, and which manipulated the number of body-neighbours for words and/or nonwords.

These included eight experiments from our lab and one by M. Taft, and both lexical decision and reading aloud tasks in English and German (see Appendix).

The experimental procedures of all studies were typical of psycholinguistic research. In the studies from our lab, each item was presented to the participant using the software DMDX, for 5 seconds or until a response occurred (in the case of reading aloud, this was measured by when a voice key was triggered; in lexical decision, by a button press). All reading aloud responses were scored for accuracy offline with the program CheckVocal (Protopapas, 2007), which allows the researcher to adjust vocal response onsets, thus reducing the bias associated with voice key triggering for different first phonemes. The study by Taft was a lexical decision study, where each item was presented in random order until a response occurred. In the entire dataset, fourteen trials with RTs < 300 ms were discarded, as these are likely to reflect premature accidental button presses or voice-key triggers. Note that this trimming decision – and all other decisions about data analysis – were taken before any of the analyses were conducted.

All English participants were native speakers of English, recruited through Macquarie University in Sydney, Australia (or University of New South Wales, for Taft's study), and the German participants were native speakers of German, recruited through Potsdam University in Germany. A spreadsheet with the full (i.e., untrimmed) trial-level data (i.e., the data of Taft and our data) as well as the R scripts used in the large-scale study can accessed via https://osf.io/myfk3/.

The overall item characteristics across all studies that were included in the analyses (averages, SDs and correlations with body-N) are described in **Table 1**. The body-N counts are based on the same corpus analysis as those of Ziegler et al., (2001) to increase the comparability across languages (Ziegler, Stone, & Jacobs, 1997, for English, and Ziegler, personal communication, 25 January, 2012 for German). The frequency and orthographic N values are taken from WordGen (Duyck, Desmet, Verbeke, & Brysbaert, 2004), which is an interface for cross-linguistic research based on the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995). Regularity was defined as compliance to grapheme-phoneme correspondence rules, as implemented in the German and English versions of the DRC (Coltheart et al., 2001; Ziegler, Perry, & Coltheart, 2000).

Before reporting the results of the large-scale analysis, we outline the characteristics and basic results of the individual studies, and thereby show the degree to which the body-N effect is stable across experiments. Note that we are restricted in the types of conclusions that we can draw from a series of smaller studies: although the sample sizes of each study are similar to those of a typical psycholinguistic study, it is possible that they do not have enough statistical power to consistently detect a true small effect of Body-N and systematic differences across languages (Button et al., 2013; Meehl, 1990; Schmidt, 1992).

Basic descriptions and outcomes of the studies that were included in the meta-analysis are summarised in the Appendix and in the supplementary material. We re-analysed the trial-level data of all experiments with LMEs. The *t* and *p* values (calculated by the R packages "lme4" and "lmTest", respectively; Bates et al., 2015; Hothorn et al., 2015) provided in the "Results" column are based on LMEs, using body-N as a centred continuous predictor for inverse RTs (–1000/RT), with items and participants as random effects. The body-N slope was allowed to vary across participants. Inverse RTs were used to bring the distribution closer to normal.

To summarise the basic results of the item sets, we inspected the BFs as a function of task, lexicality, and language. These are presented in **Table 2**.

Table 2 shows that for most types of items, the BF provides either equivocal evidence, or stronger evidence for the absence of the main effect of body-N than for its presence. This does not support the notion of a real effect of body-N. An exception is the lexical decision task for German nonwords, which provides evidence for an *inhibitory* effect of body-N, albeit based on only one study.

3.2. Analyses and results

To further explore the pattern of results, we combined all studies described in the previous section to obtain greater power in the assessment of stability of the effects. Note that all analysis scripts (for R), as well as the untrimmed data, are available in the supplemental materials.

For three reasons, we did not analyse accuracy data. Firstly, overall accuracy rates were very high, mean = 95.25%, and the ceiling effects reduce our chances of finding meaningful effects. Secondly, the original studies on body-N effects in adults focussed on RTs: Ziegler et al. (2001) did not analyse the accuracy data, and Ziegler and Perry (1998) report only a weak body-N effect in accuracy in the words condition, which was significant by participants but not by items. Thirdly, the BayesFactor package, at this stage, has not implemented the possibility to test logistic models against each other (in trial-level accuracy data, the outcome variable is binary).

For RTs, we performed four groups of analyses: for nonwords in reading aloud, nonwords in lexical decision, words in reading aloud and words in lexical decision. We analysed these conditions separately, because different cognitive mechanisms may underlie response latency variance in each of the four conditions (Coltheart et al., 2001). This should be reflected in different patterns of the body-N effect. For example, we expect a facilitatory body-N effect for reading aloud words and nonwords, and for lexical decision for words, as stronger activation of a body unit may enhance lexical activation and/or the sublexical assembly process (Ziegler & Perry, 1998; Ziegler et al., 2001). For lexical decision for nonwords, however, we might expect an inhibitory body-N effect: if a high body-N nonword elicits more lexical activation compared to a low body-N nonword, it will be harder to reject in the lexical decision task (Coltheart et al., 1977). Each analysis included both the English and German items, which enabled us to assess any interactions between body-N and language, as this is relevant to the psycholinguistic grain size theory (Ziegler & Goswami, 2005).

	English						German					
	Words (N = 283)			Nonwords (N = 305)			Words (N = 158)			Nonwords (N = 169)		
	Range	Mean (SD)	Correlation	Range	Mean (SD)	Correlation	Range	Mean (SD)	Correlation	Range	Mean (SD)	Correlation
Body-N	1–23	8.65 (5.96)	N/A	0–20	7.91 (5.48)	N/A	1–19	6.92 (5.41)	N/A	0–19	6.58 (5.21)	N/A
Number of letters	3–7	4.47 (0.88)	$r = 0.01$, $p = 0.87$	3–6	4.61 (0.78)	$r = -0.02$, $p = 0.78$	3–6	4.38 (0.93)	$r = 0.10$, $p = 0.22$	3–7	4.40 (0.91)	$r = 0.16$, $p < 0.05$
Orthograph. N	0–25	5.43 (4.54)	$r = 0.29$, $p < 0.0001$	0–24	4.41 (4.17)	$r = 0.17$, $p < 0.01$	0–15	3.72 (3.02)	$r = 0.16$, $p < 0.05$	0–15	3.59 (2.79)	$r = 0.14$, $p = 0.07$
Onset complexity	0–3	1.57 (0.61)	$r = 0.34$, $p < 0.0001$	1–3	1.71 (0.53)	$r = 0.31$, $p < 0.0001$	0–4	1.39 (0.75)	$r = 0.20$, $p < 0.05$	1–4	1.47 (0.66)	$r = 0.40$, $p < 0.0001$
Consistency ratio	0.25–1	0.95 (0.12)	$r = 0.01$, $p = 0.92$	0.5–1	0.97 (0.08)	$r = 0.02$, $p = 0.76$	0.08–1	0.97 (0.12)	$r = -0.02$, $p = 0.79$	0.5–1	0.98 (0.08)	$r = 0.02$, $p = 0.81$
% Regular words	N/A	94.40%	N/A	N/A	N/A	N/A	N/A	96.05%	N/A	N/A	N/A	N/A
Log frequency	0.00–3.92	1.59 (0.67)	$r = -0.02$, $p = 0.68$	N/A	N/A	N/A	0.00–4.40	1.49 (0.72)	$r = 0.12$, $p = 0.13$	N/A	N/A	N/A

Table 1: Descriptive statistics across all items included in the analyses, and their correlation with body-N.

	English				German			
	Reading aloud		Lexical decision		Reading aloud		Lexical decision	
	Words	Non-words	Words	Non-words	Words	Non-words	Words	Non-words
BF > 3	0	0	0	0	0	0	0	1
BF ≈ 1	1	2	1	1	1	0	0	0
BF < 1/3	1	1	2	2	1	2	1	0

Table 2: Bayes factor values for the existing studies: Numbers of studies providing evidence for body-N effect, against body-N effect, or equivocal evidence. BF > 3 = number of studies providing support for a body-N main effect, BF ≈ 1 is equivocal evidence for and against the main effect of body-N, and BF < 1/3 is the number of studies providing evidence against an influence of body-N.

3.2.1. Measuring body-N: Types versus tokens

The published studies of the body-N effect used type body-N (the number of words with the same body) to quantify the effect (Ziegler & Perry, 1998; Ziegler et al., 2001; Ziegler et al., 2003). In the literature on word consistency effects, some evidence suggests that reliance on large units is instead driven by token frequency (Jared, McRae, & Seidenberg, 1990), which can be quantified in the context of the body-N effect as the summed frequency of all body neighbours. Practically, type and token counts are difficult to dissociate unless the item sets are created with the aim of de-correlating these variables, because they are correlated. For our combined item set, the correlation was $r(915) = 0.43$, $p < 0.0001$.

At the beginning of each set of analyses, we compared models including type versus token body-N as predictors. Our aim here was to isolate the more reliable predictor rather than adjudicating between the two measures. In every model comparison, type body-N provided a better fit to the data than token body-N according to the Akaike Information Criterion (AIC). For this reason, and also because previous research has relied on type body-N counts to quantify the body-N effect, we used type body-N counts for all subsequent analyses.

3.2.2. Body-N effect for nonwords in reading aloud

The analyses were conducted on inverse RTs as the dependent variable.[2] The predictors were body-N, language (German, English - contrast coded as 1 and –1, respectively, in the LME analyses, in order to obtain estimates of the main effects of language as a deviation from the grand mean), orthographic N, and onset complexity (the number of consonants in the onset).[3] The continuous predictors were centred by subtracting their mean from each value, so as to obtain LME parameter estimates for average values rather than extreme values of zero. We also included random intercepts for participant, item, and study as in all models. We did not include previous trial RT, previous trial accuracy, or trial order as predictors, because these were not available for all experiments.[4]

3.2.2.1. LME model analysis

As a first pass, we compared models containing (a) only the main effects of body-N, orthographic N, and language (no interactions), as well as onset complexity as a covariate, (b) adding all two-way interactions not including onset complexity, and (c) adding the three-way interaction between body-N, orthographic N, and language. We found that the model containing two-way interactions performed significantly better than the model with no interactions, $\chi^2(3) = 18.41$, $p < 0.001$, while there was no additional benefit of adding the three-way interaction, $\chi^2(1) < 1$.

The LME results for the two-way interaction model are summarised in **Table 3**. In the model including the main effects and all two-way interactions between body-N, orthographic N and language, as well as the main effect of onset complexity, we found a facilitatory main effect of orthographic N, and two-way interactions which we describe in more detail below: namely, an interaction between body-N and language, and an interaction between language and orthographic N. The interaction between body-N and orthographic N approached significance. The main effect of body-N did not approach significance.

In a set of follow-up contrasts, we explored the patterns of interactions in the results. Specifically, since we are interested in the body-N by language interaction, we sought to examine the effects of body-N for each language. **Table 3** shows that the body-N by language interaction is driven by a numerically inhibitory body-N effect for German, and numerically facilitatory effect in English. Using appropriate contrasts, we found that the body-N effect is not significant in German, $\beta = 0.003$, $t = 1.48$, $p = 0.14$, while it was significantly facilitatory in English, $\beta = -0.004$, $t = -2.58$, $p = 0.01$. This means that, in English, assuming a linear model, with the increase of one body-neighbour the response rate increases by 0.004 nonwords per second, at the values of predictor variables specified by the model.

3.2.2.2. Bayes Factor analysis

To mirror the LME analyses, we started with a comparison of a main-effects model (including language, orthographic N and body-N, as well as onset complexity as a covariate and items, participants, and study as random effects) to one which also included all two-way interactions. This provided evidence against the main-effects only model, BF = 0.12 (±1.46%). We further compared this two-way interactions model to a model including the three-way interaction, and again found evidence for the two-way interaction model, BF = 5.20 (±1.46%). We therefore adopted the two-way interaction model as a baseline for further model com-

	Estimate	Std. Error	t value	p value
Intercept	*–1.741*	*0.036*	*–48.421*	*<0.00001*
Main effects				
Body-N	–0.001	0.001	–0.401	0.689
Language	–0.008	0.029	–0.293	0.770
Orthographic N	–0.013	0.002	–6.425	<0.00001
Onset complexity	–0.019	0.014	–1.412	0.159
2-way interactions				
Body-N × language	0.003	0.001	3.194	0.001
Body-N × orthographic N	<0.001	<0.001	1.702	0.089
Language × orthographic N	–0.004	0.002	–2.593	0.010

Table 3: Results of the LME analysis for reading aloud nonwords, including body-N, Language, Orthographic N, and their two-way interactions, and the main effect of onset complexity.

parisons. To establish the importance of the main effect of body-N, we compared the two-way interactions model to one excluding both the main effect of body-N, and any interactions associated with it. Here, BF = 0.42 (±3.65%), thus providing equivocal evidence against any influence of body-N.

Even though the BF analysis does not favour a model which includes both the effects and interactions of body-N, it is not clear that we can conclude that there is either a main effect or interactions of body-N. It is possible, for example, that the main effect of body-N improves the model fit, but including the interactions decreases it and thereby counteracts a meaningful main effect. We therefore followed up with further model comparisons to establish the importance of the relevant effect and interactions.

To assess the importance of the main effect of body-N, we compared the "base" model (language and orthographic N and their interaction, plus main effect of onset complexity) to one which also included the main effect of body-N. For the model including the main effect of body-N, BF = 0.19 (±1.81%), suggesting that as a main effect, body-N is unlikely to have any influence on reading aloud nonword latencies.

As this does not rule out the possibility of a body-N by language interaction, which was significant in the LME analysis, we compared the model which included body-N (same as the body-N model for the previous analysis) against one which also included the interaction between body-N and language. Here, we found support for the model which included the interaction: BF = 4.94 (±2.01%).

Akin to the LME model, the Bayes Factor shows some evidence for a body-N by language interaction and evidence against a main effect of body-N. The LME analyses showed that the body-N was significantly facilitatory for English, but numerically inhibitory for German. As we did not control for multiple comparisons, it is unclear to what extent the significant body-N effect for English reflects a stable pattern: multiple comparisons increase the type-I error rate and thus compromise the frequentist properties of *p*-values (Cramer et al., 2015; Simmons, Nelson, & Simonsohn, 2011). In contrast to LMEs, the Bayesian approach is immune to multiple comparison problems (Dienes, 2011). Using the English data only, we compared a model with a main effect of body-N and orthographic N, to one which included only the main effect of orthographic N. This comparison showed evidence against the main effect of body-N in English only, BF = 0.19 (±1%).

3.2.2.3. Summary

The original aims were to establish whether there is a main effect of body-N, and whether body-N interacts with language. Both in the LME and BF analyses, we found no evidence for the presence of a main effect of body-N. The interaction between language and body-N emerges consistently in both analyses, likely due to the inhibitory direction of the slope of body-N in German and facilitatory direction of the slope in English. The slope in German was not supported by either analysis suggesting that the inhibitory trend that drives the interactions is spurious. Although the facilitation effect in English is significant in the LME analysis, the BF provides evidence against it.

3.2.3. Body-N effects for nonwords in lexical decision

We performed an equivalent set of analyses for the nonwords in the lexical decision task. The dependent variable and independent variables were identical to the previous set of analyses.

3.2.3.1. LME model analysis

We found no advantage of any model including interactions over one containing main effects only based on measures of model fit, both $\chi^2 < 4$ and $p > 0.2$. The main-effects only model for type body-N showed an inhibitory effect of body-N, $t = 2.86$, $p < 0.005$, and an inhibitory effect of orthographic N, $t = 3.23$, $p < 0.005$. All other $p > 0.4$. Note that the slope of the body-N effect was numerically steeper for German, $\beta = 0.005$, than for English, $\beta = 0.002$, suggesting numerically stronger inhibition in German than English.

3.2.3.2. Bayes Factor analysis

We compared the main effects only model to one which included two-way interactions, and to one which included three-way interactions. In both cases, the evidence was strongly in favour for the main-effects only model, BF > 100, which was adopted for further comparisons.

A BF comparison of the full main-effects model compared to one which excluded the main effect of body-N provided support for the body-N effect in nonword lexical decision latencies: BF = 4.75 (±2.66%). A comparison of the model which included an interaction between language and body-N as well as the main effects against a main-effects-only model provided evidence against the interaction, BF = 0.19 (±2.50%).

3.2.3.3. Summary

A relatively simple model that included no interactions was supported in this set of analyses. We found a stable inhibitory body-N effect for nonwords in lexical decision, in addition to an inhibitory effect of orthographic N. There was evidence against an interaction with language.

Given that we found a main effect of body-N, but no interaction with language, it is worth noting that inverse reaction time transformations, as used in the analyses above, have been criticised for masking interactions (Balota, Aschenbrenner, & Yap, 2013).[5] Thus, it could be argued that we did not find an interaction, because we used inverse instead of raw RTs. Therefore, we re-did the analyses using raw RT as a dependent variable, and found no improvement of fit for an interactive model compared to a main-effects only model in LME, $p > 0.5$, and evidence against a model containing the interaction between body-N and language using Bayesian techniques, BF = 0.11 (±1.9).

3.2.4. Body-N effects for words in reading aloud

In the third set of trial-level analyses, we explored body-N in the reading aloud task for words. The dependent and independent variables were identical to those for nonwords, but frequency was included as an additional predictor. An interaction of body-N and frequency is theoretically important: If bodies are processed as sublexical units, we should find a smaller effect for high-frequency words, because the rapid lexical activation associated with the processing of high-frequency words would mask the sublexical effect (Coltheart et al., 2001).

LME model analyses. Initially, we compared four models with increasing complexity: a main effects model including body-N, language, orthographic N and word frequency; a model adding the two-way interactions; a three-way interactions model; and the full model with the four-way interaction. The most complex model (including the four-way interaction between body-N, frequency, orthographic N and language) was favoured over the three-way interaction model, χ^2 (1) = 4.22, $p < 0.05$. The results of the full LME model are summarised in **Table 4**.

From the results presented in **Table 4**, there was a facilitatory main effect of frequency, and a facilitatory main effect of onset complexity. The two-way interaction between language and frequency was significant due to a stronger frequency effect for English than for German. There was a significant interaction between body-N and orthographic N. Importantly, the critical interaction between body-N and language did not approach significance, $\beta = 0.0001$, $t = 0.11$, $p = 0.91$. The four-way interaction was significant, $\beta = -0.001$, $t = 2.03$ $p = 0.04$. As this could suggest that the critical body-N by language interaction emerges only for a subset of words, we followed up with four contrasts, where we estimated the model parameters for different values of orthographic N and frequency, namely 1 SD above and below the mean. We estimated the body-N by language interaction for (1) high-frequency, high orthographic N words, (2) high-frequency, low orthographic N words, (3) low-frequency, high orthographic N words, and (4) low-frequency, low orthographic N words. For high-frequency, high orthographic N words, neither the body-N effect nor its interaction with language approached significance, both $p > 0.3$. For high frequency, low orthographic N words, the interaction between body-N and language was significant, $\beta = 0.0069$, $t = 3.265$, $p = 0.0012$, while the main effect of body-N was not, $p > 0.4$. For low-frequency, high orthographic N words, neither the main effect nor the interaction approached significance, $p > 0.6$, and for low-frequency, low orthographic N words, the effect of body-N was significant, though in the opposite to the expected direction, with longer RTs for high body-N items, $\beta = 0.0068$, $t = 2.758$, $p = 0.0062$. The interaction with language was not significant, $p > 0.1$.

For low orthographic N and high frequency words, we followed up on the body-N by language interaction by estimating the body-N effect separately for English and German. In English, the effect of body-N was not significant, $\beta = -0.0049$, $t = -1.514$, $p = 0.1310$. For German, there was a significant inhibitory effect of body-N, $\beta = 0.0088$, $t = 2.856$, $p = 0.0045$.

3.2.4.2. Bayes Factor analyses

In contrast to the LME analyses, the BF analysis very strongly favoured the main effects model over any of the more complex models, all BFs > 9000 for the main-effect model. Therefore, the model used in the following BF analyses included only the main effects of body-N, orthographic N, frequency, and language, as well as onset complexity as a covariate and study, item, and participant as random factors.

To establish whether body-N had an effect on reading aloud latencies, we compared a main effects model which excluded the body-N effect to one which included it. Here, we obtained equivocal evidence against the presence of a main effect of body-N, BF = 0.42 (±1.78%).

As for the nonword lexical decision analyses, we compared a main effect model which also included body-N by language interaction, to the main-effects only model. We obtained evidence against the model which includes the body-N by language interaction, BF = 0.23 (±2.76%).

	Estimate	Std. Error	t value	p value
Intercept	*–1.904*	*0.032*	*–59.046*	*<0.001*
Main effects				
Body-N	0.002	0.001	1.266	0.206
Language	–0.005	0.027	–0.205	0.838
Orthographic N	–0.003	0.002	–1.336	0.182
Log frequency	–0.034	0.012	–2.884	0.004
Onset complexity	–0.061	0.012	–5.169	<0.001
2-way interactions				
Body-N × language	<0.001	0.002	0.109	0.913
Body-N × orthographic N	–0.001	<0.001	–2.213	0.028
Body-N × log frequency	–0.003	0.002	–1.693	0.091
Language × orthographic N	0.001	0.004	0.316	0.752
Language × log frequency	0.043	0.021	2.002	0.046
Orthographic N × log frequency	<0.001	0.004	0.039	0.969
Three-way interactions				
Body-N × language × orthographic N	<0.001	<0.001	–1.641	0.102
Body-N × language × log frequency	0.004	<0.001	1.910	0.057
Body-N × orthographic N × log frequency	<0.001	0.001	0.217	0.829
Language × orthographic N × log frequency	0.001	0.004	0.367	0.714
Four-way interaction				
Body-N × language × orthographic N × log frequency	–0.001	0.001	–2.027	0.043

Table 4: Results from the LME for reading aloud words.

3.2.4.3. Summary

For the reading aloud word data, the LME analyses, at face value, suggest a highly complex interactive pattern (though in the absence of a significant body-N effect or a two-way body-N by language interaction), while the Bayes Factors support a simple model which contains only the main effects of language, frequency, body-N and orthographic N. Follow-up contrasts of the LME four-way interaction showed that there was a significant body-N effect for the high-frequency, low orthographic N, German words, and for the low-frequency, low-orthographic N English words. Contrary to the predictions of the psycholinguistic grain size theory, both significant body-N effects were inhibitory rather than facilitatory. Furthermore, the two significant body-N by language interactions reflected opposite directionalities: for high-frequency words, the German body-N slope was steeper than the English slope, and for low-frequency words, the English slope was steeper than the one for German.

As neither of these results have been expected *a priori*, they are likely to reflect spurious interactions. We are testing for multiple contrasts (Cramer et al., 2015), and the *p*-value for the four-way interaction just exceeds the 0.05-threshold. If there is a true effect, and given a large sample sizes such as ours, *p*-values are likely to be substantially smaller than the conventional cut-off of 0.05 (Lakens & Evers, 2014; Simonsohn, Nelson, & Simmons, 2014). Furthermore, an inhibitory body-N effect was not predicted *a priori*, nor was the particular interactive pattern that we report. In addition to the results of the Bayes Factor analyses, these points suggest that it is unlikely that the four-way interaction reflects a real population pattern. Neither the LME nor the Bayes Factor analyses supported a body-N main effect nor a two-way interaction between body-N and language.

3.2.5. Body-N effects for words in lexical decision

The last set of analyses was performed on lexical decision latencies for words. The dependent and independent variables were identical to the reading aloud for words analyses (and identical to the nonword analyses, except for the inclusion of frequency and its interactions).

3.2.5.1. LME model analysis

A model comparison showed a significant advantage for a model including all three-way interactions over one including only the two-way interactions, χ^2 (4) = 14.20, p < 0.01, but no further improvement of a model including the four-way interaction, χ^2 (1) = 2.19, p > 0.1. The results of the body-N model including all three-way interactions are summarised in **Table 5**.

Table 5 shows a main facilitatory effect of frequency, but no effect of language or body-N. The two-way interaction between language and body-N does not reach

significance, nor do any of the three-way interactions involving it.

3.2.5.2. Bayes Factor analysis

To mirror the LME analyses, we again constructed a set of models to assess the stability of the interactions. The evidence against the two-way interaction compared to the main effects model was equivocal, BF = 0.79 (±2.68%), as was the evidence against the two-way compared to the three-way interaction model 0.38 (±2.9%). The main effect model, however, was supported over the three-way interaction model, BF = 3.38 (±1.61%) and over a four-way interaction model, BF = 7.43 (±1.56%). Therefore, the evidence suggests that a main-effects only model performs substantially better than the models including the three- and four-way interactions, and numerically better than the two-way interactions model.

Excluding all interactions, we compared a model including body-N to one excluding it. Here, we found evidence against the model which included body-N, BF = 0.11 (±3.36%). Furthermore, we examined the theoretically important interaction between body-N and language. Here, the evidence for the body-N by language interaction was 0.32 (±1.28%), suggesting that body-N does not interact with language.

3.2.5.3. Summary

As for the reading aloud word results, the analyses for lexical decisions of words seem to be characterised by higher-order interactions according to the LME analyses, although the BF analyses showed little support for any interactions. None of the analyses, however, showed any evidence for the presence of a body main effect, nor for an interaction with language.

As a caveat, it should be noted that a meta-analysis of various lexical decision studies for words may mask differences between studies. In the case of orthographic N, it has been shown that the presence or absence of an effect depends on the types of nonwords that are used as foils (Andrews, 1989). This may influence the participants' decision criteria, such that they rely on summed lexical activation of all neighbours, leading to a facilitatory neighbourhood effect, if the task is easy, and on full lexical access when the task is difficult, resulting in an inhibitory neighbourhood effect if access to the specific orthographic form is slowed down by inhibition from its neighbours. In the case of our studies, there does not seem to be variability in the size of the body-N effect for lexical decisions on words: as shown in **Table 2**, three out of four studies provide evidence against a body-N effect, and one provides only equivocal evidence.

4. General Discussion

The psycholinguistic grain size theory predicts facilitatory effects of body-N overall, with a stronger effect in English than German due to the former's greater orthographic depth (Ziegler & Goswami, 2005; Ziegler et al., 2001; Ziegler et al., 2003). Two of the key studies supporting the psycholinguistic grain size theory are based on the body-N effect, and report evidence that the effect is stronger in English than German (Ziegler et al., 2001; Ziegler et al., 2003). A closer inspection of these two studies identified issues both with the methodology and the statistical analyses. We therefore aimed to assess the strength of the evidence for the claim that the size of the body-N effect differs across orthographies.

First, we conducted a re-analysis of the original data reported by Ziegler et al. (Perry & Ziegler, 2002; Ziegler et al., 2001), and one using trial-level reading aloud data from the English Lexicon Project (Balota et al., 2007) for the same items. Then, we carried out a large-scale analysis of nine studies collected by two different labs. In all of the analyses reported here, we found little evidence for a main effect of body-N effect or a cross-linguistic difference between English and German readers. When there was evidence for such a main effect (in the lexical decision data for nonwords of the large-scale analyses), there was evidence against a body-N by language interaction. When there was evidence for an interaction (reading aloud nonwords in the large-scale analysis), this occurred in the absence of a main effect of body-N. In the analyses of word (reading aloud and lexical decision) data, there was no support either for a body-N effect, nor for an interaction between body-N and language. These results suggest that the body-N effect is not a reliable marker effect for larger unit processing. Moreover, the data do not support the main claim of the psycholinguistic grain size theory: that English readers rely routinely on bodies, while German readers tend to rely on smaller units.

4.1. Some notes on interpreting LME and Bayes Factor analyses

The use of Bayes Factors is relatively new in psychological research, but has several advantages over traditional frequentist approaches. For our purposes, supplementing the LME analyses with Bayes Factors allowed us to provide direct support *for* the null hypothesis (no body-N effects and/or no body-N by language interaction), relative to the corresponding alternative hypotheses.

In several of the large-scale analyses, the LME showed statistically significant interactions, while the Bayes Factor provided evidence against the same interactions. It is possible that some of the interactions, in the population, are so small that they are closer to the BF's H_0 than H_1, and that the BF therefore erroneously provides evidence for the null. Conversely, it is also likely that some of the significant *p*-values are false positives, especially given the large number of comparisons presented in the current analyses (Cramer et al., 2015). Significant *p*-values may also reflect a violation of normality, which is an (implicit) assumption of the null-hypothesis model.

4.2. Additional theoretical implications of body-N effects

The body-N effect has theoretical implications beyond the psycholinguistic grain size theory. Specifically, it is not clear whether the body-N effect reflects a lexical analogy strategy, where similar lexical entries facilitate word recognition (Forster & Taft, 1994; Goswami, 1993), or reliance on larger sublexical units (Coltheart, Curtis, Atkins, &

	Estimate	Std. error	t value	p value
Intercept	*–1.613*	*0.081*	*–19.989*	*0.001*
Main effects				
Body-N	<0.001	0.001	0.299	0.765
Language	0.032	0.078	0.406	0.723
Orthographic N	–0.001	0.003	–0.281	0.779
Log frequency	–0.099	0.011	–9.175	<0.001
Onset complexity	0.002	0.014	0.111	0.912
Two-way interactions				
Body-N × language	0.001	0.001	1.182	0.238
Body-N × orthographic N	–0.001	<0.001	–2.074	0.039
Body-N × log frequency	–0.004	0.002	–2.384	0.018
Language × orthographic N	0.003	0.003	0.959	0.338
Language × log frequency	0.033	0.011	3.020	0.003
Orthographic N × log frequency	–0.003	0.004	–0.646	0.518
Three-way interactions				
Body-N × language × orthographic N	–0.001	0.000	–1.889	0.060
Body-N × language × log frequency	0.002	0.002	1.054	0.293
Body-N × orthographic N × log frequency	<0.001	<0.001	0.818	0.414
Language × orthographic N × log frequency	–0.011	0.004	–2.806	0.005

Table 5: Results from the LME for lexical decision of words.

Haller, 1993; Patterson & Morton, 1985; Perry, Ziegler, & Zorzi, 2007). In the former case, we might expect overall stronger body-N effects for lexical decision than reading aloud data, because performance on the lexical decision task is assumed to rely to a greater extent on direct lexical access, while reading aloud is influenced to a greater extent by sublexical-phonological processes. For words in lexical decision, the body-N effect would be facilitatory, as the summed lexical activation of body neighbours would facilitate a correct "yes" response, while for nonwords the effect would be inhibitory, because the lexical activation from the body neighbours would push for a "yes" response even though the item is a nonword.

Conversely, if the effect of body-N reflects sublexical reliance on bodies, we would expect the strongest effect for reading aloud of nonwords. Reading aloud of nonwords must be achieved via a sublexical decoding mechanism, because lexical activation is not sufficient for a correct response – all other tasks, in theory, can be performed by relying solely on lexical activation (or the lack thereof, for lexical decision of nonwords).

In the large-scale analyses, the only condition which showed a stable effect of body-N was lexical decision for nonwords. Here, a higher body-N led to longer latencies, meaning that high body-N nonwords are more difficult to reject than low body-N nonwords. The body-N effect seems to exist in addition to an inhibitory orthographic N effect (which is relatively consistently reported in the existing literature on orthographic N; for a review, see Andrews, 1997). This suggests that bodies reflect some aspect of the lexical system: a high body-N nonword appears to cause lexical activation of its body neighbours, and this lexical activation makes it more difficult to determine that it is a nonword.

In the other conditions, there was no trace of a main effect of body-N. We are not implying that the results suggest that bodies have no psychological reality, as this would be inconsistent with a growing body of research using other paradigms showing reliance on bodies. As mentioned in the introduction, nonword reading studies that manipulate the existence versus non-existence of a body in real words (is *dake* easier to read than *daik*?) consistently show body effects (Andrews, Woollams, & Bond, 2005; Goswami et al., 2003; Rosson, 1985; Treiman, Goswami, & Bruck, 1990), as do nonword reading studies, where the use of bodies would predict a different pronunciation compared to grapheme-phoneme correspondences, such as *dalk*, which can be read to rhyme with "talk" (if body-rime correspondences are used), or "talc" (if grapheme-phoneme correspondences are used; Andrews & Scarratt, 1998; Brown & Deavers, 1999; Glushko, 1979; Schmalz et al., 2014). A further set of studies on the psychological reality of bodies show that words and nonwords with inconsistent bodies (e.g., "-eat", which can be pronounced as in "beat", "great" or "sweat") are read aloud more slowly than words with only one possible pronunciation (e.g., "-eet"; Andrews, 1982; Cortese & Simpson, 2000; Jared, 1997, 2002; Jared et al., 1990). In light of this other research, we believe the appropriate interpretation of the absence of the body-N effect in three out of four

conditions here is that the body-N effect is not a sensitive measure of reliance on bodies.

4.3. Evidence for cross-linguistic differences in the reliance on bodies

As described in the introduction, the key prediction that distinguishes the psycholinguistic grain size theory from the orthographic depth hypothesis is stronger reliance on orthographic units that are larger than graphemes and smaller than words in readers of deep compared to shallow orthographies. We did not find support for this prediction. However, there are alternative explanations for the absence of a body-N by language interaction. This warrants a thorough examination of all possibilities. A strong explanation should ideally account both for the current set of results, and also for previous studies that have been reported to support the psycholinguistic grain size theory. Future studies should aim to provide further evidence to distinguish between possible explanations.

There are three possibilities: First, that the psycholinguistic grain size theory is correct: there may be small, but theoretically meaningful, cross-linguistic differences in the reliance on bodies driven by orthographic depth. If these differences are sufficiently small, our current study would not be able to provide evidence for them, because it would have insufficient power (in frequentist terms) and because the prior for the alternative hypothesis used for the current analyses was set too high (in Bayesian terms). Second, there may be individual differences in the reliance on bodies, but these may not be driven by orthographic depth, but rather by cross-cultural differences in the type of reading instruction. Third, given the issues of the published studies which we identified in the introduction, and the results of our current analyses, it is possible that any cross-linguistic differences observed are noise around a true mean of zero.

4.3.1. Possibility 1: The psycholinguistic grain size theory is right

Our results do not support the psycholinguistic grain size theory, but they do not unequivocally disprove it. Especially given that we have not found the body-N effect to be a reliable marker effect, an alternative explanation is that the body-N manipulation is ill-suited for exploring cross-linguistic differences in the reliance on bodies. Furthermore, we report some slope differences which go in the direction expected by the psycholinguistic grain size theory: For reading aloud of nonwords, we find evidence for an interaction both in the LME and in the Bayes Factor analyses – although the main effect is not significant, the Bayes Factors gives evidence against a facilitatory effect of body-N in English, and the interaction is driven by an unexpected inhibitory trend in German. In an individual-study cross-linguistic comparison (reported in the supplementary materials), we find a significant body-N by language interaction in the item set listed as Experiments 4 and 7 in the Appendix – though, again, this occurs in the absence of a main effect, and the Bayes Factor provides weak evidence against the interaction, BF = 0.31 (±3.6%).

Note that the argument of a potential small effect can be almost always made against a study supporting an H_0. This is because most psychological theories, including the psycholinguistic grain size theory, make directional rather than quantitative predictions. Thus, if in the population, the benefit of each additional body neighbour is 1.1 ms for German and 1.2 ms for English, this would still support the psycholinguistic grain size theory, but one would need thousands of participants to provide evidence for such an alternative hypothesis.

Therefore, all existing data (including the published studies and our analyses) do not unequivocally rule out the possibility of a small cross-linguistic difference. To provide stronger evidence for or against the unique prediction made by the psycholinguistic grain size theory, future research could involve a large-scale confirmatory study. Such a study would be stronger if the reliance on bodies was tested using the body-existence effect, rather than the body-N effect: the body-existence effect seems to be reliable, as it has been reported by studies from various labs and with well-controlled stimuli (Andrews et al., 2005; Brown & Deavers, 1999; Rosson, 1985; Treiman et al., 1990). It would furthermore need to control for language-level, item-level, as well as participant-level confounds. Confirmatory analysis with a body-existence manipulation would address a drawback of the current study, namely that the body-N effect is not a reliable marker effect. Furthermore, the data in the current study were collected for various purposes, making the large-scale analyses exploratory. A future confirmatory study could plan, *a priori*, to collect sufficient data to confirm or disconfirm a smallest effect of interest, thus making sure that the study is adequately powered in frequentist terms, and that the prior in a Bayes Factor analysis is theoretically informed.

In summary, it is possible that the psycholinguistic grain size theory is acccurate, and that there are cross-linguistic differences in the reliance on bodies and other large sublexical units. However, the existing evidence to date – based on the current study, as well as previous experiments – does not provide a convincing evidence for it.

4.3.2. Possibility 2: Individual differences as a function of reading instruction

If a cross-linguistic difference exists, it does not automatically follow that this difference is attributable to orthographic depth. As outlined in the introduction, an alternative explanation is that reliance on bodies is driven by whole-word reading instruction rather than orthographic depth (Thompson et al., 2009). Reading instruction methods have been discussed by previous studies as a potential confound associated with the German/English comparison (Landerl, 2000; Wimmer & Goswami, 1994).

This would explain the trends in the right direction for the nonword reading aloud data, in the absence of convincing evidence for a main effect. If only some of the English participants show a body-N effect, this would lead to an overall facilitatory slope of body-N, but with an increase in variability that would make the overall results less clear-cut. The English-speaking participants were

recruited in Australia, where reading instruction methods are varied. Therefore it is likely that some of the participants had received whole-word reading instruction while others received phonics instruction.

Unfortunately, neither the published studies nor our own data allow us to address this hypothesis, as information about the participants' schooling was unavailable. In fact, such data are difficult to obtain for adult participants, unless a primary school curriculum is implemented on a national level, as adult participants rarely remember details about their reading instruction. Any follow-up work on the reliance on bodies should take this potential confound into account by collecting data in a country where reading instruction methods are standardised, or by recruiting children whose teachers have been interviewed to confirm that there are no cross-linguistic differences.

4.3.3. Possibility 3: There are no cross-linguistic differences in the body-N effect

Overall, we are confident in concluding that there is, to date, no reliable evidence for a cross-linguistic difference in the reliance on bodies as a function of orthographic depth. The existing published studies do not provide strong evidence, and our attempts to provide evidence for a body-N effect or an interaction with language have further weakened it. As discussed above, however, any stronger conclusions about the evidence for an absence of a cross-linguistic difference in the reliance on bodies need to be postponed until there is a confirmatory study addressing this issue.

4.4. Theoretical challenges for the psycholingustic grain size theory

While we have focussed here on explicit predictions of the psycholinguistic grain size theory, future theoretical and empirical work could use the broad framework of the psycholinguistic grain size theory in its current form to generate more explicit predictions about the exact factors and mechanisms that drive the reliance on various sublexical units and correspondences. From the existing literature, it is clear that there is a considerable degree of diversity in the type of linguistic units that underlie print-to-speech conversion across orthographies (Asfaha, Kurvers, & Kroon, 2009; Duncan et al., 2013; Morais, Alegria, & Content, 1987; Nag, 2007; Schmalz et al., 2014; Taft & Radeau, 1995). As the psycholinguistic grain size theory focuses mainly on orthographic depth – the degree to which small units are predictive of the correct pronunciation – it makes no direct predictions about other factors which may influence the reliance on particular sublexical units or correspondences. A description of language-level differences beyond orthographic depth, and how these could affect specific cognitive mechanisms would generate a wealth of testable predictions. Moving beyond orthographic depth would help to create a framework of reading and reading acquisition that is not limited to alphabetic orthographies (Schmalz et al., 2015; Share, 2014).

In addition to such language-level factors, future theoretical work on the psycholinguistic grain size theory could also clarify psychological factors and constraints that drive reliance on different types of sublexical units. The main claim is that statistical inconsistency causes reliance on larger units. However, most orthographies contain some level of inconsistency, and it is always necessarily the case that taking into account larger units – which maximises the informational value of the processed string – reduces ambiguity. Indeed, research has shown that readers of orthographies which are generally considered shallow rely on sublexical units that are larger than graphemes, namely bodies in German (Schmalz et al., 2014) and syllables in Spanish (Carreiras, Alvarez, & Devega, 1993). This is broadly consistent with the psycholinguistic grain size theory, in the sense that it is in line with the notion that readers of all orthographies reduce inconsistency by relying on larger units. However, it becomes unclear why one would expect cross-linguistic differences in "the size of the dominant spelling units, the number of different grain-size levels, and the reader's flexibility to switch between different levels" (p. 383, Ziegler et al., 2001), as it is advantageous for readers of any orthography to rely on larger units and to flexibly switch to smaller units when they are confronted with unfamiliar spelling patterns.

In summary, there are two theoretical challenges for future work on the psycholinguistic grain size theory. First, predictions about factors beyond orthographic depth would help to provide a deeper understanding of the specific language-level variables that influence the reliance on various sublexical units, and how the cognitive mechanisms interact with language-level factors during reading acquisition. Second, a consideration of cognitive mechanisms that drive cross-linguistic differences may help to clarify how the specific statistical distributions of a given orthography may encourage readers to prefer units of a particular type.

4.5. Conclusion

In summary, the psycholinguistic grain size theory has proposed that readers of deep orthographies, such as English, rely on large units such as bodies to greater extent than readers of shallow orthographies, such as German. In the current paper, we show that there is no convincing evidence for this claim. We conclude that a confirmatory analysis is needed to provide stronger evidence for or against the psycholinguistic grain size theory and its prediction that orthographic depth affects the reliance on large sublexical units, after controlling for confounds such as reading instruction.

In the meantime, we propose that the routine reliance on small versus large sublexical units does not depend on the depth of the orthography. Instead, the existing evidence supports the Orthographic Depth Hypothesis, that lexical processing becomes relatively more important if the sublexical information is difficult to decipher (Katz & Frost, 1992). An important contribution of the psycholinguistic grain size theory lies in sparking interest in the use of different sublexical units across orthographies. Even in the absence of cross-linguistic differences in the reliance on orthographic bodies, future empirical and theoretical research could aim to establish the linguistic and

psychological factors that affect the choice of units across orthographies.

Data accessibility

The items used for the experiments described in the paper, the raw data, and the analysis script (R) can be found on the Open Science Framework platform: https://osf.io/myfk3/.

Acknowledgements

We thank Marcus Taft for providing his body-N data, and for helpful comments on an earlier version of this manuscript, Sachiko Kinoshita for insightful discussions about methodological and statistical issues, and David Balota and Melvin Yap for providing the trial-level ELP data. We are further grateful for feedback from Becky Treiman, Karin Landerl, and Conrad Perry, on an earlier version which was a part of XS's doctoral thesis, and to Jo Ziegler, for responding to our queries about the body-N data.

Competing Interests

The authors have no competing interests to declare.

Authors Contributions

XS, SR, EM, AC, and MC contributed to the conception and design of the paper. XS collected the data, XS and SR contributed to the analysis, and all authors contributed to the interpretation of the data and revision of the draft. All authors approved of the version submitted for publication.

Notes

[1] Note that this is also a problem for the studies on the body existence effect.

[2] It has been suggested that all models should also allow the by-participant slopes of body-N to vary as random factors, because a failure to do so may increase the Type-I error rate (Barr, Levy, Scheepers, & Tily, 2013). As all of our critical conclusions are based on null-effects, they are not compromised by the possibility of an increased Type-I error rate. In fact, recent simulations have shown that maximising the model structure may reduce statistical power (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2015). Variance components related to interaction effects are usually unreliable. Thus, excluding by-participant slopes may increase the chances to detect fixed effects, in the long run, relative to a maximal model.

[3] Onset complexity is mainly included to act as a covariate. Other than the item set of Ziegler et al., all studies that were included in the analyses manipulated body-N while keeping orthographic N constant, meaning that high body-N words tended to contain more complex onset clusters to reduce the orthographic N value. As this may act to suppress a body-N effect, we included the effect of onset complexity as a statistical control.

[4] Throughout the paper, we report analyses that include all items. There might be two reasons to include only items with low body-N counts: firstly, it could be argued that a body-N effect is evident only for low body-N items, if the psychological saliency of bodies operates in an all-or-none manner. Secondly, body-N counts are not linearly distributed, because there are more words with smaller body-N values. However, conducting the same analyses while excluding items with body-N > 5 did not change any of the critical results. We therefore report the full analyses, as they have higher statistical power.

[5] We thank an anonymous reviewer for pointing out this possibility.

References

Andrews, S. (1982). Phonological recoding: Is the regularity effect consistent? *Memory & Cognition, 10*(6), 565–575. DOI: https://doi.org/10.3758/BF03202439

Andrews, S. (1989). Frequency and neighborhood size effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 802–814. DOI: https://doi.org/10.1037/0278-7393.15.5.802

Andrews, S. (1992). Frequency and Neighbourhood Effects on Lexical Access: Lexical Similarity or Orthographic Redundancy? *Journal of Experimental Psychology: Learning, Memory & Cognition, 18*(2), 234–254. DOI: https://doi.org/10.1037/0278-7393.18.2.234

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review, 4*(4), 439–461. DOI: https://doi.org/10.3758/BF03214334

Andrews, S., & **Scarratt, D. R.** (1998). Rule and analogy mechanisms in reading nonwords: Hough Dou Peapel Rede Gnew Wirds? *Journal of Experimental Psychology-Human Perception and Performance, 24*(4), 1052–1086. DOI: https://doi.org/10.1037//0096-1523.24.4.1052

Andrews, S., Woollams, A., & **Bond, R.** (2005). Spelling-sound typicality only affects words with digraphs: Further qualifications to the generality of the regularity effect on word naming. *Journal of Memory and Language, 53*(4), 567–593. DOI: https://doi.org/10.1016/j.jml.2005.04.002

Asfaha, Y. M., Kurvers, J., & **Kroon, S.** (2009). Grain size in script and teaching: Literacy acquisition in Ge'ez and Latin. *Applied Psycholinguistics, 30*(04), 709. DOI: https://doi.org/10.1017/S0142716409990087

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

Baayen, R. H., Piepenbrock, R., & **Gulikers, L.** (1995). The CELEX Lexical Database. Release 2 (CD-ROM): Linguistic Data Consortium, University of Pennsylvania.

Balota, D. A., Aschenbrenner, A. J., & **Yap, M. J.** (2013). Additive Effects of Word Frequency and Stimulus Quality: The Influence of Trial History and Data Transformations. *Journal of Experimental Psychology-Learning Memory and Cognition, 39*(5), 1563–1571. DOI: https://doi.org/10.1037/a0032186

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The

English Lexicon Project. *Behavior Research Methods, 39*(3), 445–459. DOI: https://doi.org/10.3758/BF03193014

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

Bates, D. M., Maechler, M., Bolker, B., Walker, S., Bojesen Christensen, R., Singmann, H., . . . Grothendieck, G. (2015). Package 'lme4'. Retrieved from https://cran.r-project.org/web/packages/lme4/lme4.pdf

Brown, G., & **Deavers, R.** (1999). Units of Analysis in Nonword Reading: Evidence from Children and Adults. *Journal of Experimental Child Psychology, 73*, 208–242. DOI: https://doi.org/10.1006/jecp.1999.2502

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & **Munafo, M. R.** (2013). Confidence and precision increase with high statistical power. *Nature Reviews Neuroscience, 14*(8). DOI: https://doi.org/10.1038/nrn3475-c4

Carreiras, M., Alvarez, C. J., & **Devega, M.** (1993). Syllable frequency and visual word recognition in Spanish. *Journal of Memory and Language, 32*(6), 766–780. DOI: https://doi.org/10.1006/jmla.1993.1038

Christley, R. (2010). Power and error: increased risk of false positive results in underpowered studies. *Open Epidemiology Journal, 3*, 16–19. DOI: https://doi.org/10.2174/1874297101003010016

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*(3), 145–153. DOI: https://doi.org/10.1037/h0045186

Coltheart, M., Curtis, B., Atkins, P., & **Haller, M.** (1993). Models of Reading Aloud - Dual-Route and Parallel-Distributed-Processing Approaches. *Psychological Review, 100*(4), 589–608. DOI: https://doi.org/10.1037/0033-295X.100.4.589

Coltheart, M., Davelaar, E., Jonasson, T., & **Besner, D.** (1977). Access to the internal lexicon. In: Dornic, S. (Ed.), *Attention and Performance, VI*. Hillsdale, NJ: Erlbaum, pp. 535–555.

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

Cortese, M. J., & **Simpson, G. B.** (2000). Regularity effects in word naming: What are they? *Memory & Cognition, 28*(8), 1269–1276. DOI: https://doi.org/10.3758/BF03211827

Cramer, A. O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., . . . Wagenmakers, E.-J. (2015). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, 1–8.

Cutler, A. (1981). Making up materials is a confounded nuisance: or Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition, 10*(1–3), 65–70. DOI: https://doi.org/10.1016/0010-0277(81)90026-3

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*(3), 274–290. DOI: https://doi.org/10.1177/1745691611406920

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*. DOI: https://doi.org/10.3389/fpsyg.2014.00781

Duncan, L., Castro, S. L., Defior, S., Seymour, P. H. K., Baillie, S., Leybaert, J., . . . Serrano, F. (2013). Phonological development in relation to native language and literacy: Variations on a theme in six alphabetic orthographies. *Cognition, 127*(3), 398–419. DOI: https://doi.org/10.1016/j.cognition.2013.02.009

Duyck, W., Desmet, T., Verbeke, L. P. C., & **Brysbaert, M.** (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods Instruments & Computers, 36*(3), 488–499. DOI: https://doi.org/10.3758/BF03195595

Earp, B. D., & **Trafimow, D.** (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6*, 621. DOI: https://doi.org/10.3389/fpsyg.2015.00621

Faust, M. E., Balota, D. A., Spieler, D. H., & **Ferraro, F. R.** (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125*(6), 777–799. DOI: https://doi.org/10.1037//0033-2909.125.6.777

Forster, K. I., & **Taft, M.** (1994). Bodies, Antibodies, and Neighborhood-Density Effects in Masked Form Priming. *Journal of Experimental Psychology-Learning Memory and Cognition, 20*(4), 844–863. DOI: https://doi.org/10.1037/0278-7393.20.4.844

Frith, U., Wimmer, H., & **Landerl, K.** (1998). Differences in Phonological Recoding in German- and English-Speaking Children. *Scientific Studies of Reading, 2*(1), 31–54. DOI: https://doi.org/10.1207/s1532799xssr0201_2

Frost, R. (1994). Prelexical and Postlexical Strategies in Reading: Evidence from a Deep and Shallow Orthography. *Journal of Experimental Psychology: Learning, Memory & Cognition, 20*(1), 116–129. DOI: https://doi.org/10.1037/0278-7393.20.1.116

Frost, R., Katz, L., & **Bentin, S.** (1987). Strategies for Visual Word Recognition and Orthographic Depth: A Multilingual Comparison. *Journal of Experimental Psychology: Human Perception & Performance, 13*(1), 104–115. DOI: https://doi.org/10.1037/0096-1523.13.1.104

Gelman, A., & **Stern, H.** (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328–331. DOI: https://doi.org/10.1198/000313006X152649

Glushko, R. (1979). The Organization and Activation of Orthographic Knowledge in Reading Aloud. *Journal of Experimental Psychology-Human Perception and Performance, 5*(4), 674–691. DOI: https://doi.org/10.1037/0096-1523.5.4.674

Goodwin, A. P., August, D., & **Calderon, M.** (2015). Reading in Multiple Orthographies: Differences and

Similarities in Reading in Spanish and English for English Learners. *Language Learning, 65*(3), 596–630. DOI: https://doi.org/10.1111/lang.12127

Goswami, U. (1993). Toward an Interactive Analogy Model of Reading Development - Decoding Vowel Graphemes in Beginning Reading. *Journal of Experimental Child Psychology, 56*(3), 443–475. DOI: https://doi.org/10.1006/jecp.1993.1044

Goswami, U., Gombert, J., & **de Barrera, L.** (1998). Children's orthographic representations and linguistic transparency: Nonsense word reading in English, French, and Spanish. *Applied Psycholinguistics, 19*, 19–52. DOI: https://doi.org/10.1017/S0142716400010560

Goswami, U., Porpodas, C., & **Wheelwright, S.** (1997). Children's orthographic representations in English and Greek. *European Journal of Psychology of Education, 12*(3), 273–292. DOI: https://doi.org/10.1007/BF03172876

Goswami, U., Ziegler, J., Dalton, L., & **Schneider, W.** (2003). Nonword reading across orthographies: How flexible is the choice of reading units? *Applied Psycholinguistics, 24*, 235–247. DOI: https://doi.org/10.1017/S0142716403000134

Hothorn, T., Zeileis, A., Farebrother, R., Cummins, C., Millo, G., & **Mitchell, D.** (2015). Package 'lmtest'. Retrieved from https://cran.r-project.org/web/packages/lmtest/lmtest.pdf

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Plos Medicine, 2*(8), 696–701. DOI: https://doi.org/10.1371/journal.pmed.0020124

Jared, D. (1997). Spelling-Sound Consistency Affects the Naming of High-Frequency Words. *Journal of Memory and Language, 36*(4), 505–529. DOI: https://doi.org/10.1006/jmla.1997.2496

Jared, D. (2002). Spelling-Sound Consistency and Regularity Effects in Word Naming. *Journal of Memory and Language, 46*(4), 723–750. DOI: https://doi.org/10.1006/jmla.2001.2827

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

Katz, L., & **Frost, R.** (1992). The Reading Process is Different for Different Orthographies: The Orthographic Depth Hypothesis. In: Frost, R., & Katz, L. (Eds.), *Orthography, Phonology, Morphology, and Meaning*. Amsterdam: Elsevier Science Publishers, pp. 67–84. DOI: https://doi.org/10.1016/S0166-4115(08)62789-2

Kliegl, R., Grabner, E., Rolfs, M., & **Engbert, R.** (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*(1–2), 262–284. DOI: https://doi.org/10.1080/09541440340000213

Lakens, D., & **Evers, E. R.** (2014). Sailing from the seas of chaos into the corridor of stability practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*(3), 278–292. DOI: https://doi.org/10.1177/1745691614528520

Landerl, K. (2000). Influences of orthographic consistency and reading instruction on the development of nonword reading skills. *European Journal of Psychology of Education, 15*(3), 239–257. Retrieved from <Go to ISI>://000169003200001. DOI: https://doi.org/10.1007/BF03173177

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

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & **Bates, D.** (2015). Balancing Type I Error and Power in Linear Mixed Models. *arXiv preprint arXiv: 1511.01864*.

Meehl, P. E. (1990). Why Summaries of Research on Psychological Theories Are Often Uninterpretable. *Psychological Reports, 66*(1), 195–244. DOI: https://doi.org/10.2466/PR0.66.1.195-244

Morais, J., Alegria, J., & **Content, A.** (1987). The relationship between segmental analysis and alphabetic literacy: An interactive view. *Cahiers de Psychologie Cognitive. European Bulletin of Cognitive Psychology, 7*, 415–438.

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

Nag, S. (2007). Early reading in Kannada: the pace of acquisition of orthographic knowledge and phonemic awareness. *Journal of Research in Reading, 30*(1), 7–22. DOI: https://doi.org/10.1111/j.1467-9817.2006.00329.x

New, B., Ferrand, L., Pallier, C., & **Brysbaert, M.** (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review, 13*(1), 45–52. Retrieved from <Go to ISI>://000237304000005. DOI: https://doi.org/10.3758/BF03193811

Patterson, K., & **Morton, J.** (1985). From orthography to phonology: An attempt at an old interpretation. In: Patterson, K., Marshall, J., & Coltheart, M. (Eds.), *Surface Dyslexia*. Hillsdale, N.J: Lawrence Erlbaum Associates, pp. 335–359.

Peereman, R., & **Content, A.** (1998). Quantitative analyses of orthography to phonology mapping in English and French. Retrieved from http://homepages.vub.ac.be/acontent/OPMapping.html

Perry, C., & **Ziegler, J.** (2002). Cross-language computational investigation of the length effect in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 28*(4), 990–1001. DOI: https://doi.org/10.1037//0096-1523.28.4.990

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

Protopapas, A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods,*

39(4), 859–862. DOI: https://doi.org/10.3758/BF03192979

Protopapas, A., & **Kapnoula, E. C.** (2013). *Exploring word recognition with selected stimuli: The case for decorrelated parameters*. Paper presented at the 35th Annual Conference of the Cognitive Science Society, Berlin. https://www.researchgate.net/profile/Athanassios_Protopapas/publication/265509979_Exploring_word_recognition_with_selected_stimuli_The_case_for_decorrelated_parameters/links/5513f9f00cf2eda0df303797.pdf

Rau, A. K., Moll, K., Snowling, M. J., & **Landerl, K.** (2015). Effects of orthographic consistency on eye movement behavior: German and English children and adults process the same words differently. *Journal of Experimental Child Psychology, 130*, 92–105. DOI: https://doi.org/10.1016/j.jecp.2014.09.012

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

Rosson, M. B. (1985). The Interaction of Pronunciation Rules and Lexical Representations in Reading Aloud. *Memory & Cognition, 13*(1), 90–99. DOI: https://doi.org/10.3758/BF03198448

Rouder, J. N., Speckman, P. L., Sun, D. C., Morey, R. D., & **Iverson, G.** (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. DOI: https://doi.org/10.3758/PBR.16.2.225

Royall, R. M. (1986). The Effect of Sample-Size on the Meaning of Significance Tests. *American Statistician, 40*(4), 313–315. DOI: https://doi.org/10.2307/2684616

Schmalz, X., Beyersmann, L., Cavalli, E., & **Marinus, E.** (2016). Unpredictability and complexity of print-to-speech correspondences increase reliance on lexical processes: More evidence for the Orthographic Depth Hypothesis. *Journal of Cognitive Psychology, 28*(6), 658–672. DOI: https://doi.org/10.1080/20445911.2016.1182172

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

Schmalz, X., Marinus, E., Robidoux, S., Palethorpe, S., Castles, A., & **Coltheart, M.** (2014). Quantifying the reliance on different sublexical correspondences in German and English. *Journal of Cognitive Psychology, 26*(8), 831–852. DOI: https://doi.org/10.1080/20445911.2014.968161

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47*(10), 1173. DOI: https://doi.org/10.1037/0003-066X.47.10.1173

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*(2), 115–129. DOI: https://doi.org/10.1037//1082-989X.1.2.115

Seymour, P., Aro, M., & **Erskine, J.** (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143–174. DOI: https://doi.org/10.1348/000712603321661859

Share, D. (2008). On the Anglocentricities of Current Reading Research and Practice: The Perils of Overreliance on an "Outlier" Orthography. *Psychological Bulletin, 134*(4), 584–615. DOI: https://doi.org/10.1037/0033-2909.134.4.584

Share, D. (2014). Alphabetism in reading science. *Frontiers in Psychology*, 1–4. DOI: https://doi.org/10.3389/fpsyg.2014.00752

Simmons, J. P., Nelson, L. D., & **Simonsohn, U.** (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*. DOI: https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & **Simmons, J. P.** (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547. DOI: https://doi.org/10.1037/a0033242

Taft, M., & **Radeau, M.** (1995). The Influence of the Phonological Characteristics of a Language on the Functional Units of Reading - a Study in French. *Canadian Journal of Experimental Psychology-Revue Canadienne De Psychologie Experimentale, 49*(3), 330–348. DOI: https://doi.org/10.1037/1196-1961.49.3.330

Thompson, G. B., Connelly, V., Fletcher-Flinn, C. M., & **Hodson, S. J.** (2009). The nature of skilled adult reading varies with type of instruction in childhood. *Memory & Cognition, 37*(2), 223–234. DOI: https://doi.org/10.3758/MC.37.2.223

Treiman, R., Goswami, U., & **Bruck, M.** (1990). Not All Nonwords Are Alike - Implications for Reading Development and Theory. *Memory & Cognition, 18*(6), 559–567. DOI: https://doi.org/10.3758/BF03197098

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & **Richmond-Welty, E.** (1995). The Special Role of Rimes in the Description, Use, and Acquisition of English Orthography. *Journal of Experimental Psychology: General, 124*(2), 107–136. DOI: https://doi.org/10.1037/0096-3445.124.2.107

Vadillo, M. A., Konstantinidis, E., & **Shanks, D.** (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review, 23*(1): 87–102. DOI: https://doi.org/10.3758/s13423-015-0892-6

Weekes, B. (1997). Differential Effects of Number of Letters on Word and Nonword Naming Latency. *The Quarterly Journal of Experimental Psychology, 50A*(2), 439–456. DOI: https://doi.org/10.1080/713755710

Wimmer, H., & **Goswami, U.** (1994). The influence of orthographic consistency on reading development: word recognition in English and German children. *Cognition, 51*(1), 91–103. DOI: https://doi.org/10.1016/0010-0277(94)90010-8

Yap, M. J., Balota, D. A., Sibley, D. E., & **Ratcliff, R.** (2012). Individual Differences in Visual Word Recognition: Insights From the English Lexicon Project. *Journal of Experimental Psychology-Human Perception and Performance, 38*(1), 53–79. DOI: https://doi.org/10.1037/a0024177

Ziegler, J., & **Goswami, U.** (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological Bulletin, 131*(1), 3–29. DOI: https://doi.org/10.1037/0033-2909.131.1.3

Ziegler, J., & **Perry, C.** (1998). No more problems in Coltheart's neighbourhood: resolving neighbourhood conflicts in the lexical decision task. *Cognition, 68*, B53–B62. DOI: https://doi.org/10.1016/S0010-0277(98)00047-X

Ziegler, J., Perry, C., & **Coltheart, M.** (2000). The DRC model of visual word recognition and reading aloud: An extension to German. *European Journal of Cognitive Psychology, 12*(3), 413–430. DOI: https://doi.org/10.1080/09541440050114570

Ziegler, J., Perry, C., Jacobs, A. M., & **Braun, M.** (2001). Identical Words are Read Differently in Different Languages. *Psychological Science, 12*(5), 379–384. DOI: https://doi.org/10.1111/1467-9280.00370

Ziegler, J., Perry, C., Ma-Wyatt, A., Ladner, D., & **Schulte-Körne, G.** (2003). Developmental dyslexia in different languages: Language-specific or universal? *Journal of Experimental Child Psychology, 86*, 169–193. DOI: https://doi.org/10.1016/S0022-0965(03)00139-5

Ziegler, J., Stone, G. O., & **Jacobs, A. M.** (1997). What is the pronunciation for -ough and the spelling for /u/? A database for computing feedforward and feedback consistency in English. *Behavior Research Methods Instruments & Computers, 29*(4), 600–618. DOI: https://doi.org/10.3758/BF03210615

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.72.pr

How to cite this article: Schmalz, X., Robidoux, S., Castles, A., Coltheart, M., & Marinus, E. (2017). German and English bodies: No evidence for cross-linguistic differences in preferred orthographic grain size. *Collabra: Psychology,* 3(1): 5, DOI: https://doi.org/10.1525/collabra.72

Submitted: 24 October 2016 **Accepted:** 19 January 2017 **Published:** 07 March 2017

UNIVERSITY of CALIFORNIA PRESS Collabra

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS