AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology:
Human Perception and Performance

© 2020 American Psychological Association
ISSN: 0096-1523

2020, Vol. 46, No. 11, 1235–1251
http://dx.doi.org/10.1037/xhp0000853

Do Effects of Visual Contrast and Font Difficulty on Readers' Eye Movements Interact With Effects of Word Frequency or Predictability?

Adrian Staub
University of Massachusetts Amherst

The time a reader's eyes spend on a word is influenced by visual (e.g., contrast) as well as lexical (e.g., word frequency) and contextual (e.g., predictability) factors. Well-known visual word recognition models predict that visual and higher-level manipulations may have interactive effects on early eye movement measures, because of cascaded processing between levels. Previous eye movement studies provide conflicting evidence as to whether they do, possibly because of inconsistent manipulations or limited statistical power. In the present study, 2 highly powered experiments used sentences in which a target word's frequency and predictability were factorially manipulated. Experiment 1 also manipulated visual contrast, and Experiment 2 also manipulated font difficulty. Robust main effects of all manipulations were evident in both experiments. In Experiment 1, interactions between the effect of contrast and the effects of frequency and predictability were numerically small and statistically unreliable in both early (word skipping, first fixation duration) and later (gaze duration, go-past time) measures. In Experiment 2, frequency and predictability did demonstrate convincing interactions with font difficulty, but only in the later measures, possibly implicating a checking mechanism. We conclude that although the predicted interactions in early eye movement measures may exist, they are sufficiently weak that they are difficult to detect even in large eye movement experiments.

Public Significance Statement
This study investigates a basic, but unresolved, question about how words are recognized in the course of normal reading. Do properties of the text such as visual contrast and font influence word recognition itself? Or does word recognition happen only after we have fully recognized individual letters?

Keywords: eye movements, reading, word frequency, predictability

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

At least since Sternberg (1969), cognitive scientists have investigated the processing stages underlying performance in a task by assessing patterns of interaction, and noninteraction, among the effects of experimental manipulations on mean response time (RT). Interactive effects indicate that two factors do not influence entirely separate, serially ordered processing stages, while strictly additive effects are predicted by a staged model. While additivity in mean RT can also emerge from an architecture in which the factors influence a common stage (McClelland, 1979), it does so under restricted circumstances (Roberts & Sternberg, 1993). This logic has been deployed extensively in the visual word recognition literature. RTs from single-word tasks such as lexical decision and speeded pronunciation (or "naming") have provided much of the evidentiary base for models of visual word recognition (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Harm & Seidenberg, 2004). Experiments using these tasks have investigated interactions between effects of factors such as word frequency, semantic priming, and stimulus degradation (e.g., Becker & Killion, 1977; Yap, Balota, Tse, & Besner, 2008). Following Sternbergian logic, the patterns of interaction and noninteraction among these factors have informed theories of the system's architecture.

A largely separate body of research has investigated the visual, cognitive, and linguistic processes involved in reading connected text by tracking readers' eyes movements. In normal reading, the time the eyes spend on a word is reliably influenced both by the word's context-independent frequency (e.g., Rayner & Duffy, 1986; Staub, White, Drieghe, Hollway, & Rayner, 2010) and by the word's predictability in context (e.g., Ehrlich & Rayner, 1981; Staub, 2011a). The duration of the eyes' initial fixation is longer on a low-frequency word than on a high-frequency word, and it is longer on a word that is unpredictable given its preceding context

This article was published Online First July 30, 2020.

This work was supported by a grant from the National Science Foundation (BCS 1732008). Thanks to Chuck Clifton, John Kingston, Jon Burnsky, and Kuan-Jung Huang for helpful discussion. Serene Elbach, Subin Jung, Alex Green, Leo Stern, Eunji Choi, Megan Agro, Cora Vitale, Sabrina Elbach, Hiralben Patel, and Jisun Kim provided assistance with data collection. Materials, scripts, and data are available at: https://osf.io/bnavz/?view_only=41a452245f5347d6b519aa21eba407dc.

Correspondence concerning this article should be addressed to Adrian Staub, Department of Psychological and Brain Sciences, University of Massachusetts Amherst, 430 Tobin Hall, Amherst, MA 01003. E-mail: astaub@umass.edu

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

(even when perfectly plausible) than on a word that can be predicted from context, as assessed by the cloze task. The probability that a word is skipped, rather than directly fixated, is also influenced by both variables (e.g., Angele, Laishley, Rayner, & Liversedge, 2014; Drieghe, Rayner, & Pollatsek, 2005). These are among the central empirical results that are accounted for by computational models of eye movements in reading such as E-Z Reader (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle, Rayner, & Pollatsek, 2003; Reichle, Warren, & McConnell, 2009) and SWIFT (Engbert, Nuthmann, Richter, & Kliegl, 2005). Unsurprisingly, reading is also influenced by the quality of the visual stimulus, with readers making longer fixations on faint text than clear text (e.g., Reingold & Rayner, 2006; White & Staub, 2012).

The main empirical goal of the present study is to assess interactions in the eye movement record between frequency and predictability, on the one hand, and visual contrast on the other, as well as interactions between the former variables and the less-studied effect of font difficulty (Rayner, Reichle, Stroud, Williams, & Pollatsek, 2006). A strictly additive pattern, in which effects of contrast or font difficulty do not interact with effects of higher-level (frequency, predictability) variables, is consistent with an architecture in which the influence of visual manipulations is confined to an early "perceptual normalization" stage (Yap & Balota, 2007; Yap et al., 2008) that strictly precedes lexical processing itself. On the other hand, an interactive pattern, in which the frequency effect is larger with degraded text, is generally predicted by the architecture of well-known models of visual word recognition (Coltheart et al., 2001; McClelland & Rumelhart, 1981; Morton, 1969; Rumelhart & McClelland, 1982; see Balota, Yap, Cortese, & Watson, 2008, for discussion).

Abstracting away from some differences in implementation between these models, the prediction of an interaction effect is derived as follows. It is assumed that the frequency effect in tasks such as lexical decision and naming (i.e., the difference in RT between high-frequency and low-frequency words) arises because high-frequency words require less activation than low-frequency words to reach a threshold for selection. The size of this RT difference will depend on the rate at which activation is accumulated; when activation accumulates slowly, a given difference between words in how much activation is required to reach threshold will translate into a larger effect on RT than when activation accumulates quickly. The rate at which lexical units are activated depends on, among other things, the quality of the stimulus; lexical activation will increase more slowly when the stimulus is degraded than when the stimulus is intact, leading to the prediction that the observed frequency effect should be larger with degraded stimuli.

The key assumption in this prediction of an interaction is that processing is *cascaded*, rather than *thresholded*. When processing is cascaded, partial activation at one level has an immediate impact on units at a later level (Coltheart et al., 2001). Specifically, differences in the rate at which visual features are activated, depending on whether text is clear or degraded, will have downstream consequences for the rate at which lexical activation accumulates. In a threshold model, on the other hand, activation is passed from one level to the next only once processing at the former level is complete; regardless of the effect of stimulus quality on the rate at which visual features are identified, the output of this processing stage is passed all at once to the word level, once a "verdict" has been reached. A threshold model is, in effect, a strictly staged model, and would predict strictly additive effects of stimulus degradation and word frequency.

The evidence regarding interaction in single word tasks such as lexical decision and naming is quite mixed. In the naming task, word frequency and stimulus quality appear to interact in the predicted manner, with degraded text eliciting a larger frequency effect (e.g., O'Malley, Reynolds, & Besner, 2007; Yap & Balota, 2007). However, in lexical decision, the effects of word frequency and stimulus quality on lexical decision RT have generally been additive, rather than interactive (e.g., Becker & Killion, 1977; Yap & Balota, 2007). Even fairly recently, apparent findings of interaction (Masson & Kliegl, 2013) have been shown to be unreliable (Balota, Aschenbrenner, & Yap, 2013; Masson, Rabe, & Kliegl, 2017). Complicating matters further are findings that stimulus quality does interact with semantic priming in lexical decision, with degraded targets benefitting more from a preceding semantic prime (e.g., Balota et al., 2008; Borowsky & Besner, 1993), and that word frequency and priming also interact in lexical decision (e.g., Becker, 1979). One proposal to account for this patterns holds that stimulus quality influences an early stage, word frequency influences an entirely separate, later stage, and semantic priming influences both (Borowsky & Besner, 1993).

Models of eye movements in reading have engaged only indirectly with the question of whether early visual processing and lexical processing occur in separate, serially ordered stages. The architecture of the E-Z Reader model (Reichle et al., 2009) seems to imply that they do, as it assumes a prelexical visual processing stage (V) for each word, which strictly precedes lexical processing. It is at the V stage at which visual features and possibly letter identities are extracted, before the model's lexical processing stages (L1 and L2). Reingold and Rayner (2006) refer to a stimulus quality manipulation as influencing the L1 stage, but this is not consistent with the description elsewhere of the model's V and L1 stages; for example, Reichle et al. (2003) write that the L1 stage for a word must "wait until early visual encoding of that word has been completed" (p. 452). The SWIFT model (Engbert et al., 2005), on the other hand, might predict interactions, especially between predictability and lower-level manipulations, as a single "lexical processing rate" parameter is influenced both by visual factors, such as the eccentricity of a word relative to the point of fixation, and by cloze probability. It is important to note, however, that these issues have not been directly addressed by either model; we know of no simulations, with either model, of the effects of joint manipulations of lexical or contextual and visual factors.

The potential interaction between word frequency and stimulus quality has been addressed in several eye movement experiments, in English (Sheridan & Reingold, 2013; Warrington, McGowan, Paterson, & White, 2018, Experiment 1), German (Jainta, Nikolova, & Liversedge, 2017, Experiment 3), and Chinese (Liu, Li, & Han, 2015; Wang et al., 2018). The empirical picture is mixed. Jainta et al. (2017) found no significant interactions, and the two Chinese studies (Liu et al., 2015, and Wang et al., 2018) also did not find interactions emerging consistently across experiments. However, these were relatively small experiments, both in terms of the number of subjects and in terms of the number of observations per subject in each cell of the experimental design. For example, Jainta et al. (2017, Experiment 3) included only 20 subjects, each reading 10 sentences in each experimental condition; below, we discuss the issue of power in eye movement

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

studies targeting interaction effects. On the other hand, both Sheridan and Reingold (2013) and Warrington et al. (2018) reported significant interaction effects, whereby the effect of word frequency was larger with faint text than with clear text in the reading time measures of first fixation duration (the duration of the reader's very first eye fixation on a target word) and gaze duration (the sum of the durations of all eye fixations on the first encounter with a word). The Sheridan and Reingold (2013) study was large compared with many eye movement studies, with 72 subjects each reading 36 sentences in each condition, and in this study, a modest 8 ms interaction effect on first fixation duration (a 19 ms frequency effect with normal text, vs. a 27 ms effect with faint text) did reach significance. The interaction effect was stronger in the gaze duration measure, with the frequency effect increasing from 56 ms with normal text to 82 ms with faint text. Sheridan and Reingold (2013) interpreted this result as confirming Reingold and Rayner's (2006) assumption that word frequency and stimulus quality do influence at least one common stage. Warrington et al. (2018) obtained quite similar results.

However, there is reason to question the generalizability of the results from both Sheridan and Reingold (2013) and Warrington et al. (2018). Sheridan and Reingold (2013) manipulated the contrast of only the critical word, rather than the whole sentence. This effectively "highlights" the critical word in the faint text condition. White and Staub (2012) explicitly compared reading of normal text, text in which the entire sentence was faint, and text in which only a critical word was faint, and found that the effect of faint text on gaze duration was more than three times as large when only the critical word was faint. They also found that the effect of faint text was distributionally very different in the two faint conditions. Thus, a substantial part of the effect of presenting only a single faint word may be because of the anomalous presentation of the word, in relation to the surrounding text, rather than to stimulus degradation itself. To the extent that this manipulation interacts with word frequency, the source of this interaction is unclear.

Warrington et al. (2018), on the other hand, did manipulate the contrast of entire sentences. However, their experiment also examined the between-subjects variable of participant age, with 16 younger, 16 middle-aged, and 16 older adults. The frequency-by-stimulus quality interactions that they obtained are carried primarily by the older adult group. For example, the younger adults showed a 36 ms effect of frequency on gaze duration with clear text, and a 42 ms effect with faint text (i.e., a 6 ms interaction effect), while the older adults showed a very similar 35 ms effect with clear text, but a 64 ms effect with faint text (a 29 ms interaction effect). Though the three-way interaction of age, stimulus quality, and word frequency did not reach significance in most measures, the same trend is usually apparent. The analyses did not directly address the question of whether the frequency-by-stimulus quality interaction was statistically reliable in the younger adult group.

In summary, there are no existing eye movement studies that demonstrate a significant frequency-by-stimulus quality interaction when (a) the entire sentence's contrast is manipulated, rather than a single target word, and (b) the effect is not potentially modulated by an effect of age. The experiments that have manipulated the entire sentence's contrast and have examined only younger, college-age readers (Jainta et al., 2017, Experiment 3; Liu et al., 2015, Experiment 2), have not found significant interactions; indeed, Jainta et al. (2017, Experiment 3), which is the only published study with younger readers, a whole-sentence contrast manipulation, and an alphabetic script, reported a numerically *smaller* frequency effect with faint text than with clear text in the measures of first fixation duration and gaze duration. However, these have been relatively small studies. Thus, we regard it as an open question whether frequency and stimulus quality do interact, when the potentially confounding "highlighting" of a faint word is not at issue, and when younger readers are the target population.

The predictability-by-stimulus quality interaction has not been investigated at all, to our knowledge. This is a critical empirical gap. One reason for interest in this question arises from the fact that frequency and predictability themselves have additive, not interactive, effects on reading time (e.g., Kretzschmar, Schlesewsky, & Staub, 2015; Rayner, Ashby, Pollatsek, & Reichle, 2004; see Staub, 2015, for a review). The pattern in the skipping probability measure is less clear in some studies (e.g., Rayner et al., 2004), but in this case, too, there is little evidence for interaction when considering the literature as a whole (Staub, 2015). While an additive pattern is consistent with the possibility that frequency and predictability themselves influence separate stages, this inference would be more strongly supported by the finding that one of these factors, but not the other, interacts with stimulus quality. Such a dissociation would imply that the roles of frequency and predictability are functionally distinct, and would suggest an ordering of the two effects, with one influencing a relatively early stage that is also influenced by stimulus quality, and the other influencing only a later stage.

Are there theoretical reasons to suppose that either word frequency or predictability should specifically interact with the effect of stimulus quality? At least two lines of evidence suggest that predictability may be more likely to interact with stimulus quality. The first comes from joint manipulations of predictability and frequency with manipulations of *parafoveal preview*. Staub and Goddard (2019), following up on patterns in previous studies such as Balota, Pollatsek, and Rayner (1985) and Reingold, Reichle, Glaholt, and Sheridan (2012), found that when the boundary paradigm (Rayner, 1975) is used to deny the reader parafoveal preview of a target word until it is directly fixated, the predictability of the target word no longer has an influence on early eye movement measures, while the frequency effect persists. The lack of predictability effect with invalid preview has now been replicated in Chinese reading (Chang et al., 2020), as well as in an eye movement corpus study (Luke, 2018) using the moving window paradigm (McConkie & Rayner, 1975). Thus, predictability may have most or all of its effect while a word is being viewed parafoveally, before it is directly fixated, while frequency also influences processes that occur during foveal viewing. In other words, the effect of predictability may occur primarily during very early processing stages.

The second line of evidence comes from the patterns of interaction and noninteraction in the lexical-decision task that we have discussed above. As we have noted, it is well established that in the lexical-decision task, the effects of stimulus quality and semantic priming are interactive, with a larger priming effect for degraded targets (e.g., Becker, 1979), while the effects of stimulus quality and word frequency appear to be strictly additive (e.g., Becker & Killion, 1977). It is at least a plausible assumption that a semantic

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

prime in the lexical-decision task operates similarly to a constraining context in normal reading, by "preactivating" a word form or a broader set of semantic features (Staub, 2011a, 2015). On this assumption, the dissociation that is present in the lexical-decision task, where only priming, not frequency, interacts with stimulus quality, suggests that in reading, the effect of predictability but not frequency may interact with the effect of stimulus quality.

Thus, in Experiment 1 of the present study we explore the two-way interactions in the eye movement record among frequency, predictability, and stimulus quality, in a design with substantial power to detect these interactions. We expect to replicate the additive effects of frequency and predictability on the mean of the first fixation duration and gaze duration measures, as well as their additive effects on word skipping probability. The central questions are, first, whether frequency interacts with stimulus quality, an interaction that has been obtained in some studies (Sheridan & Reingold, 2013; Warrington et al., 2018), but notably not in experiments in which the contrast of the entire sentence is manipulated and subjects are young adults (e.g., Jainta et al., 2017), and second, whether predictability interacts with stimulus quality, an interaction that has not been previous investigated.

In Experiment 2, we assess interactions between frequency, predictability and a manipulation of font difficulty. Font difficulty manipulations have generally been assumed to target letter-level processing, as opposed to processing of visual features (e.g., Pelli, Burns, Farell, & Moore-Page, 2006; Sanocki & Dyson, 2012). Rayner et al. (2006) tested both the interaction between frequency and font difficulty and the interaction between predictability and font difficulty, using two separate sets of materials. They did not find any interaction in the former case, and only marginal interactions in the latter case. Again, however, this was a relatively small study, with 32 subjects (16 younger and 16 older readers), each reading 20 target words at each level of frequency and font difficulty, and only nine target words at each level of predictability and font difficulty. Slattery and Rayner (2010) also failed to find reliable interactions between font difficulty and word frequency, but again in small experiments. Thus, we again regard it as an open question whether there are interactions between effects of either frequency or predictability and font difficulty.

A dissociation between the two experiments, in which frequency, predictability, or both factors interact with the effect of font difficulty, but not visual contrast, would also be highly informative about relationships between visual, letter-level, and word-level processing. A classic demonstration of interaction between letter- and word-level processing comes from the "word superiority effect," whereby recognition of a briefly presented letter is improved when the letter is presented in a word (Reicher, 1969; Wheeler, 1970). This effect has led to the assumption not only of cascaded processing between the letter and word levels, but also of feedback from a word to its constituent letters (Coltheart et al., 2001; McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982). Notably, there is no feedback in these models from the letter or word level to the level of visual features; interaction between effects of feature-level manipulations (i.e., stimulus quality) and word-level manipulations (e.g., frequency) is predicted to emerge not because of feedback, but because of cascaded processing. In summary, these models would seem to predict stronger interaction between font difficulty and predictability or frequency than between stimulus quality and predictability or frequency, because only the former interactions would be strengthened by feedback between levels.

Before proceeding, we discuss three issues that are relevant to the design, analysis, and interpretation of the present experiments. First, we consider the question of which eye movement measures should be expected to show statistical interactions, assuming interactive processing. While the naming and lexical decision tasks deliver a single RT on each trial, multiple eye movement measures can be extracted for each word, and these reflect at least partially distinct processing stages. To provide a relatively complete picture of how the manipulated variables influence eye movements, we report a range of measures for the critical word, from skipping probability, which reflects processing of the word that takes place in the parafovea, before it is even fixated, through first fixation and gaze duration, to go-past time, which sums all fixations from when a word is initially fixated until the eyes move past it to the right, including any regressive rereading triggered by the critical word. But in which of these measures would statistical interactions be expected, if lexical processing is itself affected by stimulus quality or font difficulty?

The most straightforward prediction is that interactions should emerge in the earliest measures that show effects of the higher-level variables of frequency and predictability: word skipping and first fixation duration. If feature or letter-level activation feeds forward in a cascaded manner, then manipulating the difficulty of feature or letter processing should impact lexical activation from the very earliest moments. Thus, a pattern in which an interaction emerges in measures such as gaze duration and go-past time, but not in word skipping or first fixation duration, would be unexpected under an interactive framework. However, word skipping is a binary measure—on each trial, the critical word is skipped, or it is not—and inferences from statistical additivity or interaction in this measure to staged or interactive processing operations would depend on the precise details of a linking hypothesis between the continuous progress of lexical processing, in the parafovea, and the binary skipping decision. On the other hand, first fixation duration is a temporal measure whose duration is regarded by contemporary eye movement models regard as directly reflecting the duration of an early stage of lexical processing (Engbert et al., 2005; Reichle et al., 2003; Snell, van Leipsig, Grainger, & Meeter, 2018). In E-Z Reader, for example, a saccade program to end this fixation is initiated when the L1 stage of lexical processing completes. (The only exception to this is when the system rapidly refixates the same word in a more nearly optimal location.) Thus, we regard first fixation duration as providing the most direct test of the prediction that the effects of word frequency and predictability should be more pronounced with faint text or a difficult font.

The second issue is whether to analyze raw or transformed fixation duration measures. Because fixation duration distributions, like most other RT distributions in cognitive tasks, are right-skewed (Staub et al., 2010), the residuals from linear mixed effects models will not be normally distributed when raw fixation durations in individual trials are the unit of analysis. A power transform such as the log transform will often solve this problem, as the distribution of log fixation durations will usually be approximately normal. Thus, many researchers (e.g., Baayen, 2008) have recommended such transformations when using mixed-effects models for statistical analysis of RTs (though see Liceralde & Gordon, 2018, for an argument that the case for transformation

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

based on violations of normality is less clear than has been assumed).

However, the log transform can dramatically change patterns of interaction; a strictly additive pattern in the means of raw RTs can correspond to an underadditive interaction in log RTs, and a superadditive interaction in raw RTs can correspond to an additive pattern in log RTs (Balota et al., 2013; Lo & Andrews, 2015). The patterns of statistical additivity or interaction that are predicted by staged or interactive processing models, respectively, are patterns in the means of untransformed data. If, for example, stimulus quality affects only the duration of a visual processing stage that is complete before frequency begins to exert its influence on lexical processing, then reducing stimulus quality should increase processing time by exactly the same increment across levels of frequency; this is just to say that the two manipulations should have additive effects on raw RT. Such a staged model does *not* make the prediction that the effects of stimulus quality and frequency should be additive on the log scale, but rather that there should be some degree of underadditivity on the log scale; see Balota et al. (2013) for further discussion of this issue.

Thus, data transformation, in the present case, may be statistically appropriate (cf. Liceralde & Gordon, 2018), but it is interpretively problematic. To address this issue, we present linear mixed effects models of raw fixation duration measures, but also conduct the same analyses on log transformed data, and report any cases where the two sets of models differ in their patterns of significance. As we discuss below, we also constructed post hoc Bayesian mixed-effects models to further investigate the range of plausible values of the critical interaction effects.

Finally, the third issue relates to the power of eye movement studies that target interaction effects. How large a study, in terms of both number of subjects and number of observations per condition, is required for reasonable power to detect a plausibly sized interaction effect in first fixation duration? (See Brysbaert, 2019; Leon & Heo, 2009; Wahlsten, 1991, for broader discussion of sample size and power in the context of testing interaction effects in factorial designs.) For concreteness, assume that in the clear text condition, there is a true 20 ms effect of frequency on mean first fixation duration, which increases to 30 ms in the faint text condition. We can estimate the sample size required to observe this 10 ms interaction effect with power of .8 and $\alpha = .05$ by means of a one-sample *t* test against the null hypothesis that the mean of the interaction effect is 0 ms, if we have an estimate of the interaction effect's standard deviation, across subjects; that is, an estimate of the standard deviation of the difference of differences of condition means (((faint/LF) – (faint/HF)) – ((clear/LF) – (clear/HF))).[1]

The standard deviation of the interaction effect will depend strongly on the number of observations in each cell of the design (see Brysbaert & Stevens, 2018). Analysis of data from previous experiments in our laboratory suggests that when the number of observations per cell is 20 (before any data loss, exclusion of trials because of word skipping, etc.), the standard deviation of an interaction effect on first fixation duration is often near 40 ms. Adopting this 40 ms figure as a conservative estimate, 128 subjects would then be required to detect our assumed 10 ms interaction effect, with power of .8. This number decreases, however, as the number of trials per cell increases. With 40 observations in each cell the critical standard deviation tends to be at or below 30 ms, in our previous data, and using this figure reduces the required number of subjects to 73. The number of subjects required will, of course, increase if the true interaction effect is smaller than 10 ms.

Thus, in order for an individual eyetracking experiment to provide a meaningful test of an interaction effect in a measure such as first fixation duration, it should be substantially larger than has been the norm in the field, both in terms of the number of subjects and the number of items. We planned that in each of the present experiments, 80 subjects would each read 160 sentences, 40 in each of the four conditions that define the critical 2 × 2 interactions. We somewhat overshot this goal in Experiment 2, ultimately running 92 subjects.

This power analysis also suggests that it is probably the case that no eyetracking experiment with a factorial design has ever had sufficient power to detect a three-way interaction effect on first fixation duration, as this effect size would be expected to be even smaller than a two-way interaction effect. The experiments we present here, which are very large by eyetracking standards, still do not have sufficient power to carry out convincing tests of three-way interactions, so we will not attempt to interpret any such interactions, or lack thereof. See Button et al. (2013) for extensive discussion of why low power not only reduces the probability of detecting a real effect, but also compromises interpretation of rejections of the null.

In summary, in this study we present two well-powered eye movement experiments designed to test interactions between the effects of word frequency, predictability, and visual contrast (Experiment 1) and word frequency, predictability, and font difficulty (Experiment 2). Patterns of interaction and noninteraction among these variables have the potential to elucidate the degree of interactivity that obtains between visual, letter, and lexical processing during normal reading.

Experiment 1

Method

Subjects. All subjects were students at UMass Amherst who received course credit for their participation. All were speakers of English as a first language, and none reported any history of reading or language disorder. Eighty-four subjects were run, of whom four were excluded based on either poor performance on comprehension questions or excessive blinking or track loss (as defined below), leaving final $N = 80$ (66 women; age range from 18 to 27 years, median = 20).

Procedure. Movements of the right eye were recorded, sampling at 1000 Hz, using an EyeLink 1000 (SR Research, Toronto, ON, Canada) eyetracker, interfaced with a PC computer. Sentences were displayed on a CRT monitor 55 cm from subjects. The resolution of the eyetracker was less than one character. The experiment was implemented with the EyeTrack software, and initial stages of data analysis were carried out with Robodoc and EyeDry (http://blogs.umass.edu/eyelab/software/).

[1] This simple method estimates the power to detect the interaction effect in a "by-subjects" (F1) ANOVA. Power estimation for mixed effects models is usually accomplished by simulation, for example, Matuschek, Kliegl, Vasishth, Baayen, and Bates (2017). It is not clear whether a mixed-effects model should be assumed to be more or less powerful than ANOVA; this will depend strongly on the details of model specification, and of course on the details of the data.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Subjects were instructed to read for comprehension. A three-point calibration procedure was performed at the start of the experiment, at the midway point, and as needed between trials. Each sentence appeared on the screen, on a single line, when the subject fixated a box at the left edge of the monitor. Each session lasted approximately 40 min. Individual eye fixations less than 80 ms in duration and within one character of a previous or subsequent fixation were incorporated into this neighboring fixation. Individual trials were excluded if there was a blink or track loss during first pass reading of the critical word.

Two subjects were excluded based on poor performance on comprehension questions that followed half of the sentences (below 70%); all others achieved at least 78% correct. In addition, two subjects lost more than 20% of trials because of blink or track loss, and were excluded from subsequent analysis.

Materials. Each subject read 160 critical sentences that were adopted from Kretzschmar et al. (2015). Full item statistics and norming procedures are described therein. Each sentence contained a target word ranging in length from four to eight characters (almost always a noun, and an adjective in the few remaining cases) that was either high in frequency (HF; mean of 124 occurrences per million, based on Subtlex corpus); (Brysbaert & New, 2009) or low in frequency (LF; mean of two occurrences per million). The HF and LF words did not differ significantly in length (means of 5.37 and 5.63 characters, respectively). Each target word was presented in two sentences, once in a context that rendered it highly predictable (HP; mean cloze of .78; minimum of .4), and once in a context in which it was unpredictable (LP; mean cloze of .01; maximum of .12). The full set of materials is available on the Open Science Framework (OSF).[2] Blink or track loss on the critical word resulted in the exclusion of 4.0% of trials, leaving a total of 12,285 trials for analysis.

The manipulation of visual contrast was blocked; a subject read the first half of the sentences at one level of visual contrast, and the second half at the other level. The order of blocks was counterbalanced between subjects, and subjects were given a break between blocks. Each block contained half of the full design, that is, 20 sentences at each combination of frequency and predictability. Each of the 80 target words occurred once in each block. Within blocks, sentences were presented in an individually randomized order to each subject. Each block began with four practice trials. Half of the trials in each block were followed by two-choice comprehension questions, to which the subject responded by buttonpress. Questions appeared once the sentence was removed from the screen. In total, each subject read 20 sentences in each of the eight cells defined by the full factorial design (i.e., Frequency × Predictability × Contrast). As the analysis of each two-way interaction involves collapsing across one of the factors, 40 trials contributed to each cell, as described above.

The contrast levels were designed to approximate those used by White and Staub (2012). In the clear condition, black text was presented against a gray background (luminance 56.4 cd/m^2), while in the faint condition, slightly darker gray text (luminance 46.1 cd/m^2) was presented against the same background. Both conditions used 11 point Monaco font, a fixed-width font that is the standard in most experiments in our laboratory. Examples of the stimulus conditions are shown in Table 1. (Note that the faint text may appear barely legible in this table, though subjects found it to be legible in the experiment itself. As the data will indicate, subjects showed standard frequency and predictability effects, as well as performing well on comprehension questions).

Results

Five eye movement measures for the critical word were computed. These include three reading time measures: *first fixation duration*, which is the duration of the reader's first eye fixation on the target word; *gaze duration*, which is the sum of all eye fixation durations on the word on the reader's first pass, that is, before leaving the region to the left or right; and *go-past time*, which is the sum of all fixation durations beginning with the first on the word until the reader exits the word to the right, thereby including any regressive rereading of earlier material and any rereading of the critical region itself. In computing all three of these measures, a trial is not counted if the reader skipped the word on the first pass through the sentence rather than fixating it directly. We also report two binary dependent measures: *skipping proportion*, which is the proportion of trials on which the target word was skipped on first pass, rather than directly fixated; and *regression proportion*, which is the proportion of trials on which first-pass reading of the critical word ended with a regressive (i.e., leftward) saccade, rather than a forward saccade.

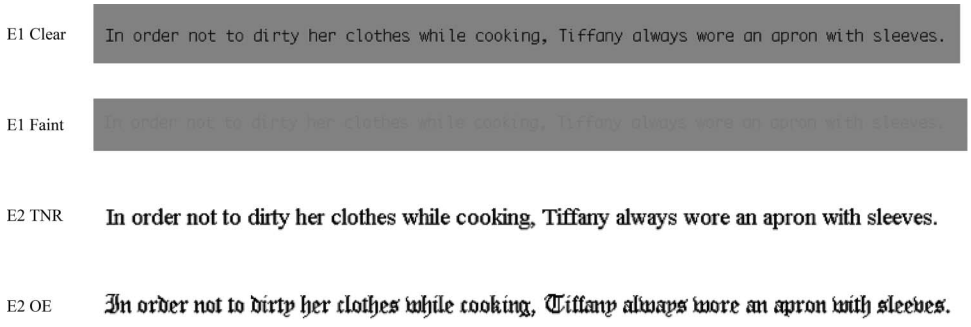
Condition means (and standard error, by subjects) are shown in the left panels of Figure 1. Fixed-effect estimates from our linear (for reading time measures) and logistic (for skipping and regressions) mixed-effects models are shown in the right panels of Figure 1, along with 95% confidence intervals on these estimates. All models included random intercepts for both subjects and items, and random slopes, by subject, for all three of the main effects. We also included random by-item slopes for predictability. The first fixation model also included random by-item slopes for stimulus quality, but other models would not converge if this additional set of random effects was included. Random by-item slopes for frequency were not appropriate; because the item was defined as the target word, frequency was a between-item manipulation. Fixed effects were coded by means of effect-coded contrasts, with the HF, HP, and clear text conditions coded as −0.5, and the LF, LP, and faint text conditions coded as 0.5. All *p* values for the reading time models are based on the Satterthwaite approximation to the denominator degrees of freedom, as implemented in the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017). Data and model code are available at our OSF link.

For expository simplicity, we discuss the three reading time measures first, then the skipping and regressions measures. For all three measures, there were sizable and significant ($p < .001$) main effects of all three manipulations. The effects of predictability and frequency were similar in size, with model estimates of about 19 ms in first fixation duration for both factors, and about 28 and 34

[2] Available at OSF website https://osf.io/bnavz/?view_only=41a452245f5347d6b519aa21eba407dc. For the present study, several sentences were slightly edited from their original form in Kretzschmar et al. (2015) to improve their naturalness. In most cases, only material following the target word was edited, which would not affect the cloze probability of the target word. In six of the 160 sentences (two high predictability, four low predictability) the sentence underwent a slight edit before the critical word. These changes were extremely minor, for example, changing *spots* to *parts* in the sentence *The couple visited all the interesting spots of the island during their visit.*

Table 1

Example Sentence (Target Word Is Apron; Low Frequency, High Predictability) in Each of the Visual Contrast Conditions of Experiment 1 (Variability in Monitor or Printer Settings May Result in Text in Faint Condition Rendering as Illegible in This Image) and Each of the Font Difficulty Conditions of Experiment 2 (TNR: Times New Roman; OE: Old English)



E1 Clear	In order not to dirty her clothes while cooking, Tiffany always wore an apron with sleeves.
E1 Faint	In order not to dirty her clothes while cooking, Tiffany always wore an apron with sleeves.
E2 TNR	In order not to dirty her clothes while cooking, Tiffany always wore an apron with sleeves.
E2 OE	In order not to dirty her clothes while cooking, Tiffany always wore an apron with sleeves.

ms in gaze duration for predictability and frequency, respectively. The model estimates of the effect of visual contrast were even larger, about 51 ms in first fixation duration and 62 ms in gaze duration. These stimulus quality effects are larger than those reported with a similar contrast manipulation by White and Staub (2012) and by Warrington et al. (2018).

As expected based on many previous null findings, the frequency-by-predictability interaction was not significant in any reading time measure. The main goal of the present experiment was to evaluate the predictability-by-stimulus quality and frequency-by-stimulus quality interactions. In first fixation duration, both of these interactions were in the expected direction—larger effects with faint text—but were numerically very small (4 and 6 ms, respectively), and nonsignificant. In gaze duration, both effects were again small (7 and 10 ms), and while the predictability-by-stimulus quality interaction did not reach significance, the frequency-by-stimulus quality interaction did ($p = .02$). In go-past time, the predictability-by-stimulus quality estimate was actually 1 ms in the opposite direction, that is, a 1 ms smaller go-past effect of predictability for faint text, and the frequency-by-stimulus quality effect was a nonsignificant 12 ms in the expected direction.

As discussed above, we also constructed identical models of log-transformed reading time measures. The patterns of significance and nonsignificance were identical to the models of raw reading times, with the exception that the frequency-by-stimulus quality interaction in the gaze duration measure did not near significance ($p = .80$) when using log transformed gaze durations. This is consistent with the general tendency of the log transform to suppress superadditive interactions.

As a final reading time analysis, in response to a reviewer's suggestion we performed a post hoc analysis of *single fixation duration*. Single fixation duration is identical to first fixation duration, but restricted to trials on which the reader made only a single fixation on the word before leaving it. Of the 9,584 trials on which the target word was fixated on first pass reading, 8,360 (87.5%) were single fixation trials. The pattern of significance mirrored first fixation duration, with three significant main effects and no significant interactions; all interaction $ps > .10$ in the analysis of raw reading times, and all interaction $ps > .22$ in the analysis of log-transformed times. Our model of raw reading times estimated the predictability-by-stimulus quality interaction effect to be 5 ms, and the frequency-by-stimulus quality interaction effect to be 3 ms.

As expected, a word was more likely to be skipped when it was predictable ($p < .001$) and when it was frequent ($p = .02$). These effects did not interact. Though the effect of stimulus quality on skipping was in the expected direction—fewer skips with faint text—this effect did not reach significance. There was a significant ($p = .02$) frequency-by-stimulus quality interaction, in the opposite direction from the frequency-by-stimulus quality interaction effect on gaze duration: The effect of frequency on skipping was slightly smaller, not larger, with faint text.

Finally, predictability significantly influenced the probability of regressing from the target word ($p < .001$). Neither frequency nor stimulus quality influenced regression probability, and there were no significant interaction effects.

To further investigate the range of true effect sizes that is consistent with the observed interaction effects in first fixation duration and gaze duration, we followed up these analyses by constructing Bayesian mixed-effects models using the *brms* package in R (Bürkner, 2017). These models had exactly the same fixed and random effect structure as the frequentist mixed-effect models we reported above. We ran versions that used the *brms* package's default uninformed (flat) priors, as well as versions that used informed priors on the main and interaction effects. For first fixation duration, we specified a Gaussian prior on each main effect with mean of 20 ms and standard deviation of 10 ms, and a Gaussian prior on each interaction effect with mean of 0 ms and standard deviation of 10 ms. For gaze duration, we specified a Gaussian prior on each main effect with mean of 30 ms and standard deviation of 15 ms, and a Gaussian prior on each interaction effect with mean of 0 ms and standard deviation of 15 ms.

All parameter estimates, in both first fixation and gaze duration models, were extremely similar to the parameter estimates from the frequentist models, and the estimates from the models with flat and informed priors were extremely similar to each other; in other words, the data are sufficiently constraining that the choice of priors has little effect on the posteriors. Here we discuss only the critical interaction effects from the model with informed priors. In

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

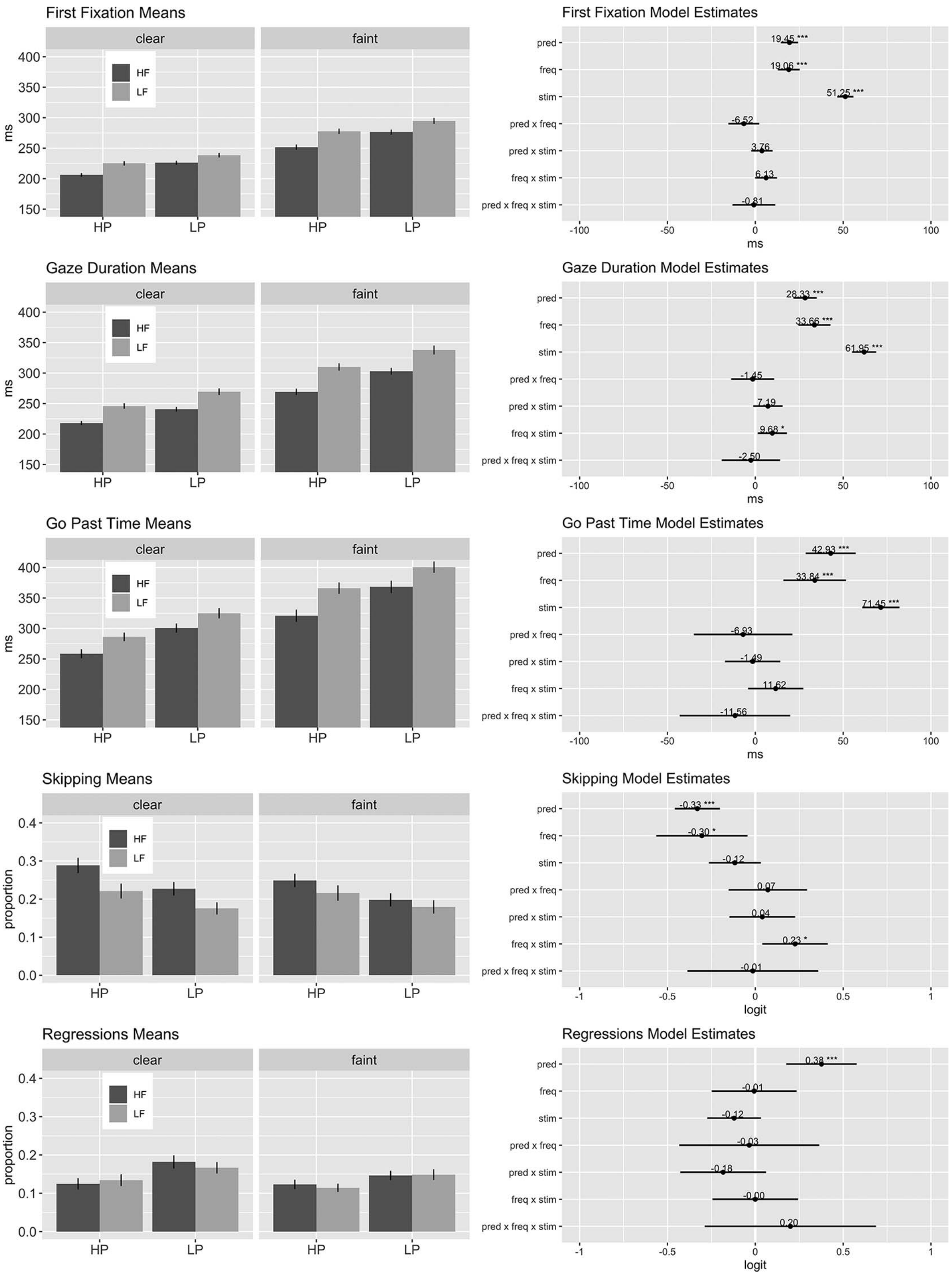


Figure 1. Experiment 1 data and model fits. Left: Means and standard errors, by subject, for five eye movement measures described in the text. Right: Parameter estimates from mixed-effects models described in the text, with 95% confidence intervals. HP = high predictability; LP = low predictability; HF = high frequency; LF = low frequency. * $p < .05$. *** $p < .001$.

the first fixation duration model, the estimate of the predictability-by-stimulus quality interaction was 3.48 ms, and the estimate of the frequency-by-stimulus quality interaction was 5.56 ms. In the gaze duration model, these estimates were 6.66 and 8.88 ms, respectively. The posterior distributions on these effects, from models with both sets of priors, are shown in Figure 2. Consistent with the results of the frequentist hypothesis tests, the only case in which the 95% highest density interval (HDI) does not include

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

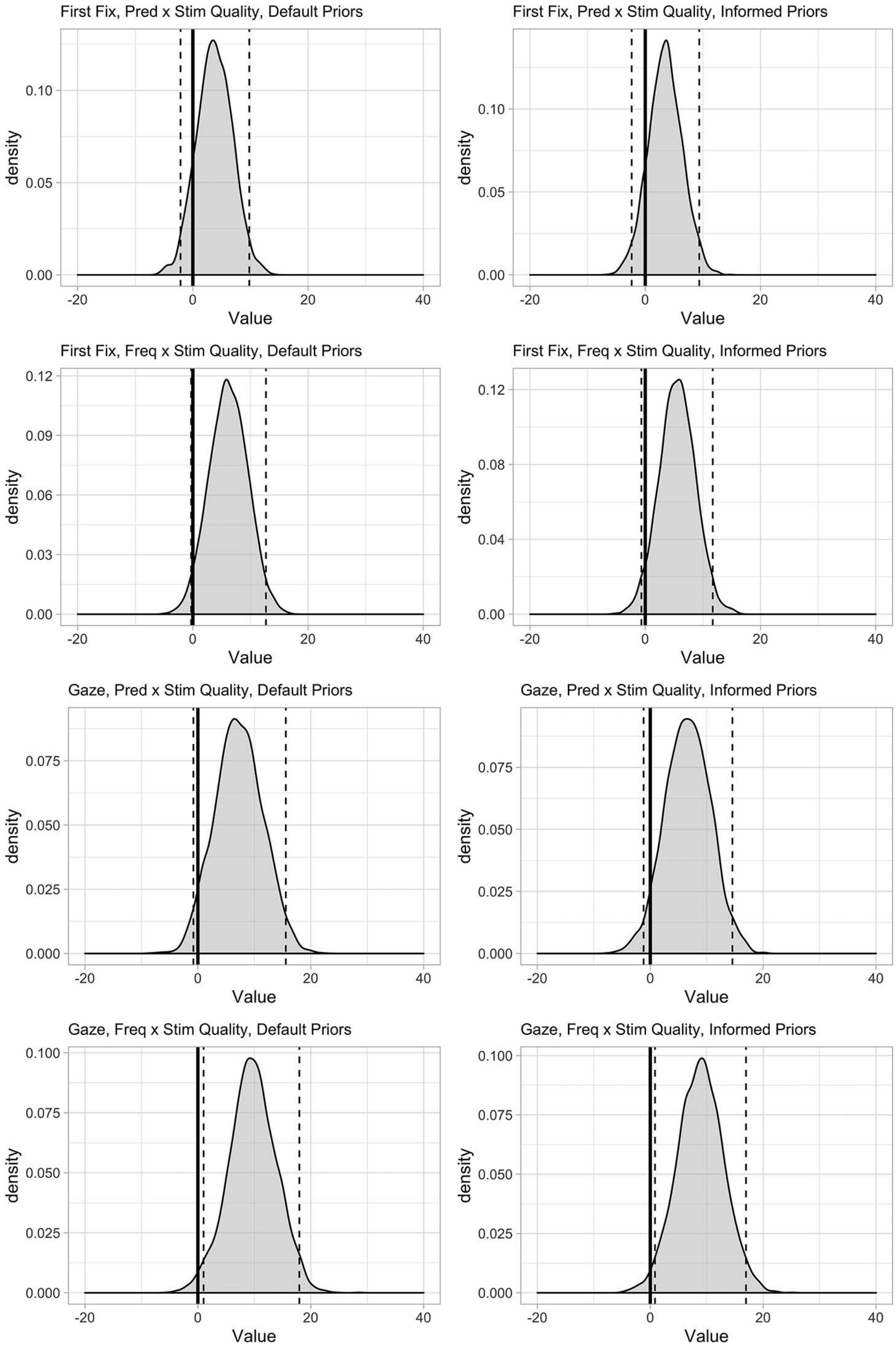


Figure 2. Experiment 1 posterior distributions for predictability-by-stimulus quality and frequency-by-stimulus quality interactions in first fixation duration and gaze duration, from Bayesian mixed-effects models with default priors (left) and informed priors (right). Dashed lines represent boundaries of 95% highest density interval (HDI).

0 is the frequency-by-stimulus quality interaction in gaze duration. Indeed, the extent of the HDI in each case is close to the extent of the frequentist 95% confidence intervals (CIs) we reported above.

Discussion

The present study replicated well-established findings: Frequency, predictability, and stimulus quality strongly influenced

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

reading times, and frequency and predictability also influenced the probability that the target word was skipped. Skipping was not significantly affected by stimulus quality; this is consistent with White and Staub's (2012) findings with a whole-sentence contrast manipulation, but not with Warrington et al. (2018), who did find an effect of contrast on word skipping. Consistent with several previous studies, predictability also influenced the probability of a regression from the target word (e.g., Staub, 2011a), but also consistent with many other studies, frequency did not (e.g., Abbott & Staub, 2015). As in previous studies, the effects of frequency and predictability were additive, not interactive, in all measures.

The experiment was designed to address the question of whether there are interactions between the effects of frequency and/or predictability and the effect of stimulus quality. There was little evidence for a predictability-by-stimulus quality interaction. While there were numerically small (4 and 7 ms) interaction effects in the predicted direction in the first fixation duration and gaze duration measures (i.e., a larger predictability effect with faint text), these effects did not reach significance. The go-past interaction effect was 1 ms in the opposite direction. There was also no evidence for an interaction effect in the skipping measure.

The conclusions regarding the frequency-by-stimulus quality interaction are somewhat less clear. These variables showed a nonsignificant 6 ms interaction effect in first fixation, a significant 10 ms interaction in gaze duration, and a nonsignificant 12 ms effect in go-past time. The significant gaze duration interaction did not approach significance when log-transformed gaze durations were used as the dependent measure. More important, given the multiple comparison issue that is present when testing an effect across a range of eye movement measures (von der Malsburg & Angele, 2017), an interaction appearing in gaze duration but not the other reading time measures might be regarded with skepticism. Indeed, the Bonferroni correction of the interaction *p* value (that was $p = .02$) recommended by von der Malsburg and Angele (2017) would render it nonsignificant. Further complicating the interpretation of this effect is the interaction that appeared in the skipping measure, in the opposite direction, with faint text reducing, rather than increasing, the effect of word frequency. This opposite interaction is problematic for any theoretical interpretation, as it is difficult to conceive of any account that would predict an interaction in one direction during parafoveal processing, giving rise to the skipping interaction, but in the other direction during foveal viewing, giving rise to the gaze duration interaction. It is worth noting that this skipping interaction is not attributable to a floor effect in the faint text conditions, as in all of these conditions the skipping rate was over 17%. However, it is also worth noting that the skipping interaction, like the gaze duration interaction, would not survive correction for multiple comparisons.

In addition to these complications, there is no statistical evidence that the predictability-by-stimulus quality interactions and the frequency-by-stimulus quality interactions are themselves different. Inspection of Figure 1 reveals that for all measures, the 95% CI on each effect includes the point estimate of the other effect; the Bayesian HDIs in Figure 2 illustrate the same point. While the frequency-by-stimulus quality interaction was significant in the gaze duration measure and the predictability-by-stimulus quality interaction was not, the estimates of these two interaction effects differ by less than 3 ms.

In summary, we conclude from the present experiment that frequency and predictability *may* weakly interact with visual contrast in measures such as first fixation duration and gaze duration, but if so, these interactions are so weak that they would be reliably detected only by an experiment that is even larger than this one. In the earliest measure that is sensitive to word frequency and predictability, skipping probability, there is no sign whatsoever of an interaction with visual contrast in the theoretically expected direction, that is, a larger effects of these variables with faint text; the observed interaction with frequency is in the opposite direction.

In Experiment 2, we test interactions between lexical and letter-level, as opposed to visual, processing. We replace the contrast manipulation with a manipulation of font difficulty. Font manipulations are generally regarded as affecting the ease of letter identification by varying the mapping between a visual stimulus and letter identity (e.g., Pelli et al., 2006; Sanocki & Dyson, 2012), as opposed to varying the difficulty with which visual features are themselves identified. In the terms of a model such as Coltheart et al. (2001), a visual contrast manipulation would be thought of as targeting the feature level, while a font difficulty manipulation would be thought of as targeting the letter level. Even if predictability and frequency do not interact with featural processing, they may interact with letter-level processing.

Experiment 2

Method

Subjects. Ninety-three subjects from the same pool as Experiment 1 participated; we originally intended this experiment to be identical in size to Experiment 1, but ran additional subjects that had already been scheduled when the intended number of subjects had been run. No one participated in both experiments. No subjects were excluded because of poor performance on comprehension questions, with all achieving at least 80% accuracy. One was excluded because of losing more than 20% of trials to track loss, leaving 92 subjects in the final analysis (70 women; age range from 18 to 22 years, median = 19).

Procedure. The procedure was identical to Experiment 1.

Materials. Subjects read the same 160 sentences as in Experiment 1. The only difference between the experiments was that the blocked manipulation of visual contrast was replaced with a blocked manipulation of font difficulty. Following Rayner et al. (2006), the variable-width Times New Roman (TNR) and Old English (OE) fonts were used as easy and difficult fonts, respectively. To match as closely as possible the spatial extent of the text, TNR and OE were presented in 14-point and 16-point size, respectively. Black text was presented against a white background in both font conditions. A total of 4.5% of trials was removed because of blink or track loss on the critical word, leaving 14,064 trials for analysis.

Results

Statistical analysis was carried out as in Experiment 1. All mixed-effects models included random intercepts for both subjects and items, random slopes, by subject, for all three of the main effects, and random by-item slopes for predictability. Condition means (and standard error, by subjects) are shown in the left panels of Figure 3. Fixed-effect estimates from our linear (for reading

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

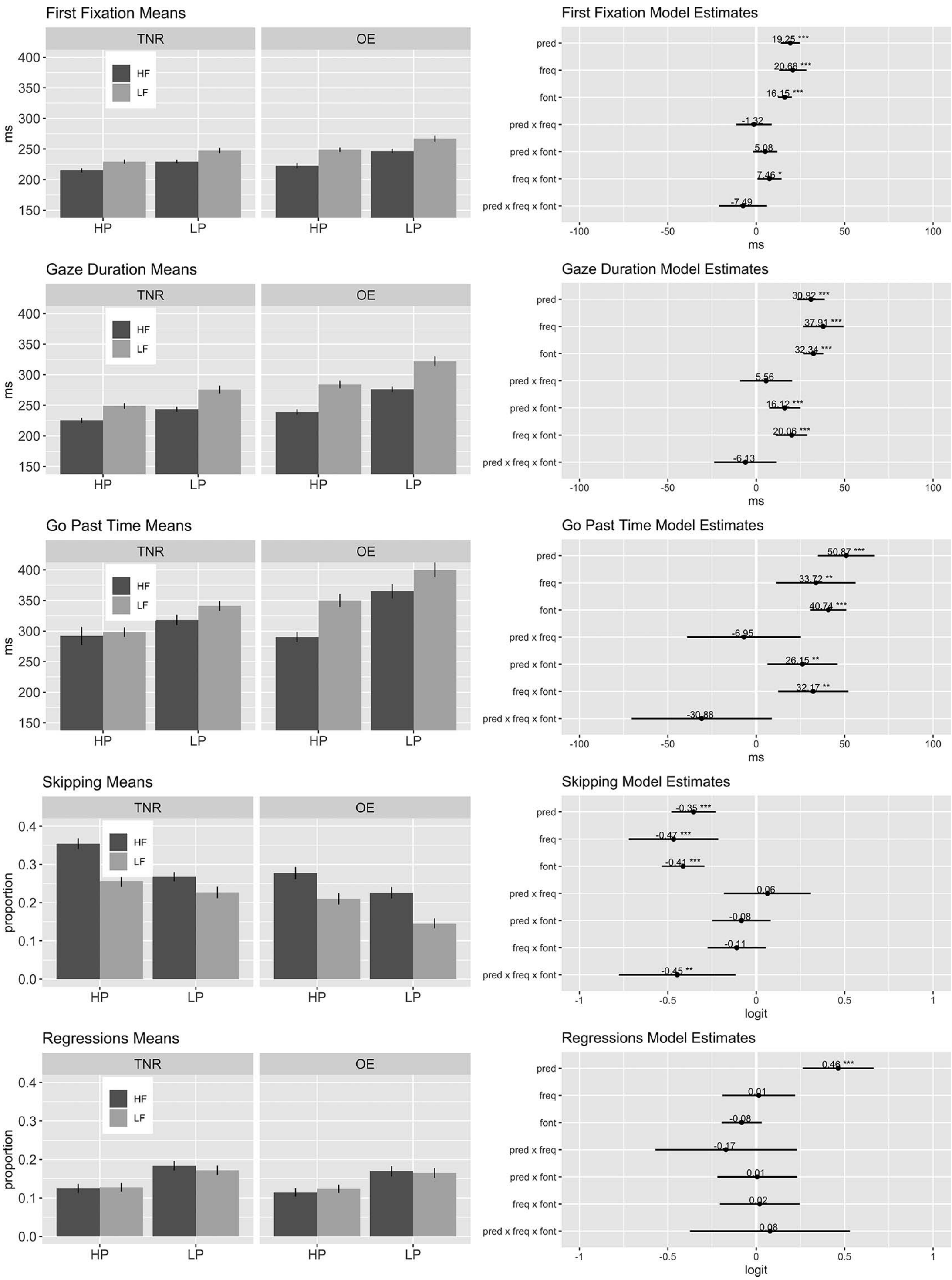


Figure 3. Experiment 2 data and model fits. Left: Means and standard errors, by subject, for five eye movement measures described in the text. Right: Parameter estimates from mixed-effects models described in the text, with 95% confidence intervals. HP = high predictability; LP = low predictability; HF = high frequency; LF = low frequency. * $p < .05$. ** $p < .01$. *** $p < .001$.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

time measures) and logistic (for skipping and regressions) mixed-effects models are shown in the right panels of Figure 3, along with 95% CIs on these estimates.

As in Experiment 1, there were significant ($p < .001$) effects of all three manipulations on all three reading time measures. The effects of predictability and frequency were similar in size to Experiment 1, while the effect of font difficulty was somewhat smaller than the effect of visual contrast in Experiment 1, though still sizable.

As in Experiment 1, predictability and frequency did not interact in any measure. Frequency and font difficulty showed a significant interaction in first fixation duration ($p < .05$), though predictability and font difficulty did not ($p = .14$). However, there were significant interactions between predictability and font difficulty, and between frequency and font difficulty, in both gaze duration ($p < .001$) and go-past time ($p < .01$). The model estimates are 16 and 22 ms for the gaze duration interactions of font difficulty with predictability and frequency, respectively, and the model estimates of the go-past interaction effects are 26 and 32 ms, respectively.

As in the analysis of Experiment 1, we also conducted the same analyses with log-transformed reading times. These models revealed identical patterns of significant effects to the models of raw reading times.

Again, we also conducted a post hoc analysis of single fixation duration, which included 9,280 (87.4%) of the 10,618 trials with a first pass fixation. In the analysis of raw single fixation duration, there were significant main effects of all three variables, but there were also 8 ms effects of both the frequency-by-font difficulty interaction and the predictability-by-font difficulty interaction, which reached statistical significance ($p = .03$ in both cases). However, in the analysis of log-transformed single fixation duration, the predictability-by-font difficulty interaction was no longer significant ($p = .10$), while the frequency-by-font difficulty interaction was $p = .05$.

The logistic regression model of skipping revealed three significant main effects, with more skips of predictable words, high frequency words, and words in Times New Roman font. No two-way interactions were significant. The three-way interaction was significant, but as noted above, we will not attempt to interpret this. Finally, as in Experiment 1, the model of regressions revealed only a predictability effect on regression probability ($p < .001$).

As for Experiment 1, we constructed Bayesian mixed-effects models of first fixation duration and gaze duration to further explore the critical interaction effects. Again, we constructed separate models with default (flat) priors and with informed priors, using the same prior specification as for Experiment 1. Posterior densities for the critical interaction effects are shown in Figure 4. The results from the two models are very similar, and are again extremely similar to the results from the frequentist analysis, with one exception: The frequency-by-font difficulty interaction in first fixation duration, which was significant in the frequentist analysis, has a 95% HDI from the default prior model that does not include 0, while 0 is well within the 95% HDI from the informed prior model.

Discussion

Experiment 2 again replicated the well-established effects of both predictability and frequency on reading times and word skipping probability, as well as the lack of interaction between these effects. As in Experiment 1, and consistent with the previous literature, predictability also influenced regression probability, while frequency did not.

The experiment also found a significant effect of font difficulty on reading times. This effect was substantial in size, but was smaller than the very large effect of stimulus quality that we observed in Experiment 1. We note that the 95% CIs on these effects in the two experiments (see Figures 1 and 2) are nonoverlapping for all three of our primary reading time measures. Unlike the effect of stimulus quality, however, font difficulty did influence the skipping rate; again, note the nonoverlapping CIs from the two experiments. Thus, we see a dissociation whereby stimulus quality had a more pronounced effect than font difficulty on reading times, while having a smaller (and indeed, nonsignificant) effect on skipping. We return to this dissociation in the General Discussion.

The main goal of this experiment was to assess the potential interactions between predictability and font difficulty, and between frequency and font difficulty. These interaction effects were not in evidence in the skipping probability measure. In first fixation duration, the predictability-by-font difficulty interaction was, like in Experiment 1, in the expected direction but very small in absolute terms, and nonsignificant. The interpretation of the frequency-by-font difficulty in first fixation duration is somewhat more complex, as this reached significance in the frequentist models, but the Bayesian model with informed priors was less clear. However, interactions between font difficulty and both predictability and frequency were clearly evident in both gaze duration and go-past time.

General Discussion

The results of these two experiments may be summarized as follows. Both experiments replicated the effects of frequency and predictability on reading times and on word skipping, which have been observed many times in previous studies. Both experiments replicated the lack of interaction between these effects, which is also expected based on previous results. In Experiment 1, there was also a very pronounced effect of the stimulus quality manipulation on reading times, and in Experiment 2, there was a slightly smaller, though still sizable, effect of the font difficulty manipulation on reading times. The font difficulty manipulation also influenced skipping probability.

With respect to the interaction effects that were the experiments' main focus, the results are more complex. In discussing these results, we proceed from the temporally earliest eye movement measures to the latest. Neither experiment demonstrated interaction effects on skipping, with the exception of a frequency-by-stimulus quality interaction effect of an unpredicted form (i.e., a smaller frequency effect on skipping with faint text) in Experiment 1. This was despite the presence of main effects of both frequency and predictability on skipping, in both experiments, and a main effect of font difficulty on skipping in Experiment 2.

In the beginning of the article, we suggested that first fixation duration is the measure on which interactive effects would be most clearly expected, on an interactive processing model. On this measure, the small numerical interaction effects, in the predicted direction, are remarkably similar in the two experiments: Predict-

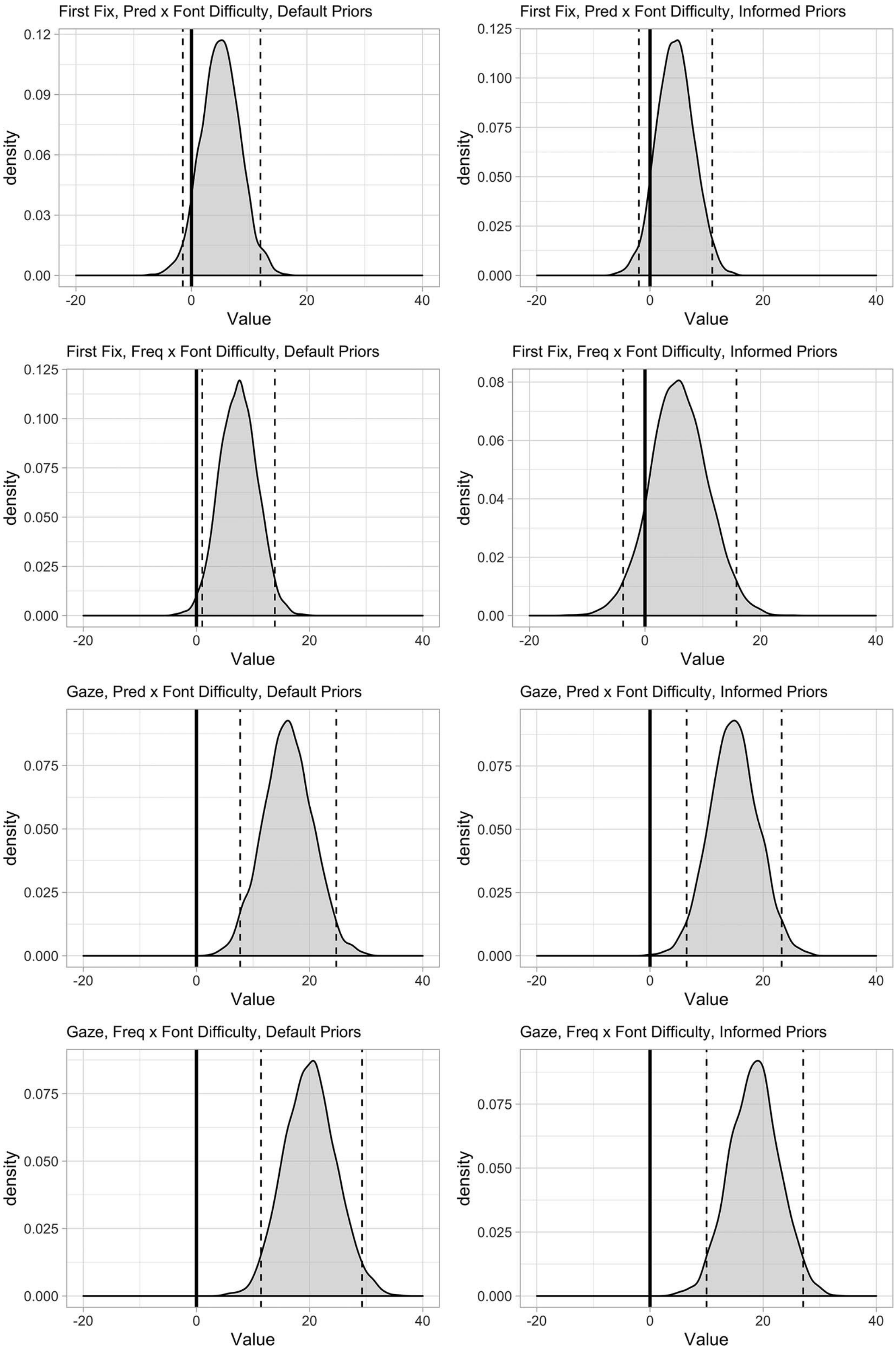


Figure 4. Experiment 2 posterior distributions for predictability-by-font difficulty and frequency-by-font difficulty interactions in first fixation duration and gaze duration, from Bayesian mixed-effects models with default priors (left) and informed priors (right). Dashed lines represent boundaries of 95% highest density interval (HDI).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

ability showed a 4 ms interaction effect with stimulus quality, and a 5 ms interaction effect with font difficulty, while frequency showed a 6 ms interaction effect with stimulus quality, and a 7 ms interaction effect with font difficulty. Only the last of these four effects reached statistical significance, and even in this case the Bayesian analysis was more equivocal. Still, the results of both experiments are consistent with the existence of very weak interaction effects on first fixation duration, in the range of 4–7 ms. Even in the Bayesian models with priors centered at 0, the peak of the posterior distributions for these interaction effects was in the vicinity of 5 ms. A real effect of this size would simply be too small to reliably reach statistical significance, even in experiments the size of the present ones. Recall that in our power analysis motivating the size of the present experiments we assumed a first-fixation duration interaction of 10 ms, calculating that 73 subjects would be needed to obtain power of .8 with 40 observations per subject in each cell of the interaction. Using the same assumptions, fully 285 subjects would be required to achieve power of .8 to detect a first-fixation interaction effect of 5 ms.

Turning to the gaze duration and go-past measures, the results of the two experiments were more distinct. With the exception of the frequency-by-stimulus quality interaction in gaze duration (which did not reach significance with log transformed gaze durations), the critical interactions were not significant in Experiment 1. However, in Experiment 2, the interactions of both frequency and predictability with font difficulty were larger, and reached significance in both gaze duration and go-past time.

The very small, and statistically unreliable, numerical interaction effects on first fixation duration in Experiment 2 (on the order of 5–7 ms) and on single fixation duration (8 ms) suggest that the larger gaze duration interaction effects (16 and 20 ms, for the predictability-by-stimulus quality and frequency-by-stimulus quality interaction, respectively) are because of differences between conditions in (a) the probability that a word is refixated, after the initial fixation, and (b) the duration of any subsequent fixations, on the occasions when a word is refixated. Post hoc analyses confirm that this is the case. First, focusing on the predictability-by-stimulus quality interaction, we see that with TNR font, a refixation is only slightly more frequent for a low-predictability word (10.6% of trials vs. 8.3%), while this difference is larger with OE font (18.0 vs. 12.7%). There is also an interactive pattern in the duration of fixations following the first, on those trials when the target did receive an additional fixation. With TNR font, this duration was 10 ms longer with a low-predictability word (206 vs. 196 ms), while with OE font, this difference was 32 ms (248 vs. 216 ms). Turning to the frequency-by-font difficulty interaction, we see similar patterns. With TNR font, refixations were more common with low-frequency words (11.9 vs. 6.8%) but this difference increased with OE font (19.1 vs. 11.4%). On trials when the target word was refixated in the TNR condition, this additional fixation time was 12 ms longer for low-frequency words (206 vs. 194 ms), but this difference increased to 41 ms in the OE conditions (250 vs. 209 ms).

What might explain the presence of interactions appearing primarily in later measures reflecting refixations? We offer the following speculative account. When reading in an unfamiliar font, readers may experience genuine uncertainty about which visual forms map to which letters. For example, to a reader unfamiliar with Old English font, the letter at the beginning of *wore* in the example sentence in Table 1 may look as much like an *m* as a *w*. Readers may resolve such uncertainty by making a within-word refixation to directly fixate the letter in question. We propose that such uncertainty is more likely to arise for infrequent or unpredictable words than for frequent or predictable ones, as in the former cases initial hypotheses about letter identity receive less support from either the lexicon or from context. An interaction effect might appear primarily in refixation-based measures if readers tend to use such refixations, when reading Old English font, to obtain more visual evidence to resolve uncertainty about letter identity. On this view, the observed interaction effects in gaze duration and go-past time would implicate differences across conditions in the need to deploy a late checking mechanism.

What are the theoretical implications of this pattern of results? Are the results consistent with a strictly staged model, such as Yap and Balota's (2007) proposal that stimulus quality influences an entirely prelexical 'perceptual normalization' stage, or E-Z Reader's (Reichle et al., 2009) assumption of a visual processing (V) stage that entirely precedes the lexical processing (L1 and L2) stages? The additive patterns in word skipping would appear to be consistent with such a model, but we are able to reconcile the first fixation duration results with such a model only if we discount the consistent, but nonsignificant, interaction effects in the range of 5 ms.

Is an interaction effect of this very modest size consistent with the predictions of interactive word recognition models such as Coltheart et al. (2001)? Deriving such predictions would require more explicit linking between word recognition models and models of eye movement control than currently exists, though see Li and Pollatsek (in press) for a very recent model of eye movement control in Chinese reading that contains a word recognition system following the interactive activation principles of McClelland and Rumelhart (1981). At present, it is not possible to say definitively whether interaction effects of this size on the first fixation measure are, in fact, weaker than would be predicted by such interactive models. For detailed discussion of whether the modeling framework of Coltheart et al. (2001), as modified in the CDP + model (Perry, Ziegler, & Zorzi, 2007) is capable of simulating additive effects of stimulus quality and word frequency under some circumstances, see Ziegler, Perry, and Zorzi (2009) and Besner and O'Malley (2009). What can be said with certainty is simply that the present data argue against any model that predicts more robust interaction effects in early eye movement measures; at a maximum, only very weak interactions are present, even when very robust effects of both visual and lexical factors are present. In this sense, the results provide an important constraint.

We briefly return to a different dissociation between the effects of the stimulus quality and font difficulty manipulations. The former manipulation clearly had stronger effects on fixation durations, while only the latter had a significant effect on word skipping. This dissociation should be regarded with some caution until it is replicated, but if it is reliable it is surprising. It is widely believed that word skipping results from a word being identified in the parafovea, while the previous word is still being fixated (e.g., Choi & Gordon, 2013). On this assumption, it is not at all obvious why stimulus quality, a variable that had an extremely large impact on first fixation duration, would show such a small, indeed nonsignificant, effect on skipping probability.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Finally, we remark on a number of other important issues in the eye movement literature that are addressed by the present data. First, these are the highest-powered experiments with factorial manipulations of frequency and predictability, and they fully confirm the lack of interaction between these factors. This lack of interaction remains an intriguing puzzle, in our view. The two factors also had essentially identical patterns of interaction, and noninteraction, with font difficulty and stimulus quality. Thus, the experiments provide no support for the conjecture (Engbert et al., 2005; Staub & Goddard, 2019), laid out in the beginning of the article, that predictability may specifically influence early visual or orthographic processing stages. This hypothesis would predict especially large predictability effects with faint text or a difficult font, which did not appear in the present study.

Second, the results of these high-powered experiments are not consistent with the idea that predictability has a stronger effect on word skipping than does frequency. This idea is embodied in the architecture of the E-Z Reader model (Reichle et al., 2003), which implements a "guessing" mechanism that results in highly predictable words sometimes being skipped without parafoveal lexical processing. This produces a pattern in which the predictability effect on skipping is predicted to be substantially stronger than the frequency effect, even when predictability does not have a stronger effect on fixation durations (e.g., Reichle & Drieghe, 2013). In both of the present very high-powered experiments, the two manipulations had almost identically sized effects on first fixation duration, and also had extremely similar-sized effects on skipping; there is no evidence at all for a dissociation between the two measures.

Third, in both of these experiments predictability had a significant, if modestly sized, effect on regression probability, while word frequency did not. These findings are broadly consistent with the previous literature (e.g., Abbott & Staub, 2015; Staub, 2011a), but this dissociation has not been well established. This dissociation is predicted if it is assumed that interword regressions arise primarily because of difficulties associated with integration of a word into its sentence context (e.g., Reichle et al., 2009). On this view, manipulations of syntactic processing difficulty or plausibility should have particularly strong effects on regression probability, and there is abundant evidence that they do (e.g., Staub, 2011b). However, this view would also naturally predict some effect of predictability, as an unpredictable word may be occasionally difficult to integrate. On the other hand, strictly lexical manipulations, such as manipulations of word frequency, should not influence regressions.

Conclusion

Two highly powered experiments assessed interactions in readers' eye movements between effects of word frequency and predictability on the one hand, and effects of visual and orthographic manipulations on the other. In neither experiment were there interactions in the word skipping measure, while there were numerically small, and generally nonsignificant, interactions in the first fixation duration measure. In the later measures of gaze duration and go-past time, there were clear interactions between font difficulty and both frequency and predictability, though not between visual contrast and these variables. While the results do not clearly arbitrate between the predictions of staged and interactive processing models, they do place important constraints on such models and on future experiments: Models should predict only very weak (if any) interactions of this sort in early eye movement measures, and no eyetracking experiments are likely to be powerful enough to detect true interaction effects of the size that this study reveals to be plausible.

References

Abbott, M. J., & Staub, A. (2015). The effect of plausibility on eye movements in reading: Testing E-Z Reader's null predictions. *Journal of Memory and Language, 85,* 76–87. http://dx.doi.org/10.1016/j.jml.2015.07.002

Angele, B., Laishley, A. E., Rayner, K., & Liversedge, S. P. (2014). The effect of high- and low-frequency previews and sentential fit on word skipping during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1181–1203. http://dx.doi.org/10.1037/a0036396

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511801686

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,* 1563–1571. http://dx.doi.org/10.1037/a0032186

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology, 17,* 364–390. http://dx.doi.org/10.1016/0010-0285(85)90013-1

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (2008). Beyond mean response latency: Response time distributional analyses of semantic priming. *Journal of Memory and Language, 59,* 495–523. http://dx.doi.org/10.1016/j.jml.2007.10.004

Becker, C. A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 252–259. http://dx.doi.org/10.1037/0096-1523.5.2.252

Becker, C. A., & Killion, T. H. (1977). Interaction of visual and cognitive effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 389–401. http://dx.doi.org/10.1037/0096-1523.3.3.389

Besner, D., & O'Malley, S. (2009). Additivity of factor effects in reading tasks is still a challenge for computational models: Reply to Ziegler, Perry, and Zorzi (2009). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 312–316. http://dx.doi.org/10.1037/a0014555

Borowsky, R., & Besner, D. (1993). Visual word recognition: A multistage activation model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 813–840. http://dx.doi.org/10.1037/0278-7393.19.4.813

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2,* 16. http://dx.doi.org/10.5334/joc.72

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977–990. http://dx.doi.org/10.3758/BRM.41.4.977

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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80,* 1–28. http://dx.doi.org/10.18637/jss.v080.i01

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376. http://dx.doi.org/10.1038/nrn3475

Chang, M., Zhang, K., Hao, L., Zhao, S., McGowan, V. A., Warrington, K. L., . . . Gunn, S. C. (2020). Word predictability depends on parafoveal preview validity in Chinese reading. *Visual Cognition, 28,* 33–40. http://dx.doi.org/10.1080/13506285.2020.1714825

Choi, W., & Gordon, P. C. (2013). Coordination of word recognition and oculomotor control during reading: The role of implicit lexical decisions. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 1032–1046. http://dx.doi.org/10.1037/a0030432

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

Drieghe, D., Rayner, K., & Pollatsek, A. (2005). Eye movements and word skipping during reading revisited. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 954–969. http://dx.doi.org/10.1037/0096-1523.31.5.954

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning & Verbal Behavior, 20,* 641–655. http://dx.doi.org/10.1016/S0022-5371(81)90220-6

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112,* 777–813. http://dx.doi.org/10.1037/0033-295X.112.4.777

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

Jainta, S., Nikolova, M., & Liversedge, S. P. (2017). Does text contrast mediate binocular advantages in reading? *Journal of Experimental Psychology: Human Perception and Performance, 43,* 55–68. http://dx.doi.org/10.1037/xhp0000293

Kretzschmar, F., Schlesewsky, M., & Staub, A. (2015). Dissociating word frequency and predictability effects in reading: Evidence from coregistration of eye movements and EEG. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41,* 1648–1662. http://dx.doi.org/10.1037/xlm0000128

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

Leon, A. C., & Heo, M. (2009). Sample sizes required to detect interactions between two binary fixed-effects in a mixed-effects linear regression model. *Computational Statistics & Data Analysis, 53,* 603–608. http://dx.doi.org/10.1016/j.csda.2008.06.010

Li, X., & Pollatsek, A. (in press). An integrated model of word processing and eye-movement control during Chinese reading. *Psychological Review*.

Liceralde, V. R. T., & Gordon, P. C. (2018). Consequences of using power transforms as a statistical solution in linear mixed-effects models of chronometric data. *PsyArXiv*. http://dx.doi.org/10.31234/osf.io/f73mh

Liu, P., Li, X., & Han, B. (2015). Additive effects of stimulus quality and word frequency on eye movements during Chinese reading. *Reading and Writing, 28,* 199–215. http://dx.doi.org/10.1007/s11145-014-9521-4

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

Luke, S. G. (2018). Influences on and consequences of parafoveal preview in reading. *Attention, Perception & Psychophysics, 80,* 1675–1682. http://dx.doi.org/10.3758/s13414-018-1581-0

Masson, M. E., & Kliegl, R. (2013). Modulation of additive and interactive effects in lexical decision by trial history. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 898–914. http://dx.doi.org/10.1037/a0029180

Masson, M. E., Rabe, M. M., & Kliegl, R. (2017). Modulation of additive and interactive effects by trial history revisited. *Memory & Cognition, 45,* 480–492. http://dx.doi.org/10.3758/s13421-016-0666-z

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94,* 305–315. http://dx.doi.org/10.1016/j.jml.2017.01.001

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287–330. http://dx.doi.org/10.1037/0033-295X.86.4.287

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88,* 375–407. http://dx.doi.org/10.1037/0033-295X.88.5.375

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics, 17,* 578–586. http://dx.doi.org/10.3758/BF03203972

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76,* 165–178. http://dx.doi.org/10.1037/h0027366

O'Malley, S., Reynolds, M. G., & Besner, D. (2007). Qualitative differences between the joint effects of stimulus quality and word frequency in reading aloud and lexical decision: Extensions to Yap and Balota (2007). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 451–458. http://dx.doi.org/10.1037/0278-7393.33.2.451

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Research, 46,* 4646–4674. http://dx.doi.org/10.1016/j.visres.2006.04.023

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

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology, 16,* 65–81. http://dx.doi.org/10.1016/0010-0285(75)90005-5

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z Reader model. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 720–732. http://dx.doi.org/10.1037/0096-1523.30.4.720

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14,* 191–201. http://dx.doi.org/10.3758/BF03197692

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging, 21,* 448–465. http://dx.doi.org/10.1037/0882-7974.21.3.448

Reicher, G. M. (1969). Perceptual recognition as a function of meaninfulness of stimulus material. *Journal of Experimental Psychology, 81,* 275–280. http://dx.doi.org/10.1037/h0027768

Reichle, E. D., & Drieghe, D. (2013). Using E-Z reader to examine word skipping during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1311–1320. http://dx.doi.org/10.1037/a0030910

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105,* 125–157. http://dx.doi.org/10.1037/0033-295X.105.1.125

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26,* 445–476. http://dx.doi.org/10.1017/S0140525X03000104

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review, 16,* 1–21. http://dx.doi.org/10.3758/PBR.16.1.1

Reingold, E. M., & Rayner, K. (2006). Examining the word identification stages hypothesized by the E-Z Reader model. *Psychological Science, 17,* 742–746. http://dx.doi.org/10.1111/j.1467-9280.2006.01775.x

Reingold, E. M., Reichle, E. D., Glaholt, M. G., & Sheridan, H. (2012). Direct lexical control of eye movements in reading: Evidence from a survival analysis of fixation durations. *Cognitive Psychology, 65,* 177–206. http://dx.doi.org/10.1016/j.cogpsych.2012.03.001

Roberts, S., & Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 611–653). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89,* 60–94. http://dx.doi.org/10.1037/0033-295X.89.1.60

Sanocki, T., & Dyson, M. C. (2012). Letter processing and font information during reading: Beyond distinctiveness, where vision meets design. *Attention, Perception, & Psychophysics, 74,* 132–145. http://dx.doi.org/10.3758/s13414-011-0220-9

Sheridan, H., & Reingold, E. M. (2013). A further examination of the lexical-processing stages hypothesized by the E-Z Reader model. *Attention, Perception, & Psychophysics, 75,* 407–414. http://dx.doi.org/10.3758/s13414-013-0442-0

Slattery, T. J., & Rayner, K. (2010). The influence of text legibility on eye movements during reading. *Applied Cognitive Psychology, 24,* 1129–1148. http://dx.doi.org/10.1002/acp.1623

Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. *Psychological Review, 125,* 969–984. http://dx.doi.org/10.1037/rev0000119

Staub, A. (2011a). The effect of lexical predictability on distributions of eye fixation durations. *Psychonomic Bulletin & Review, 18,* 371–376. http://dx.doi.org/10.3758/s13423-010-0046-9

Staub, A. (2011b). Word recognition and syntactic attachment in reading: Evidence for a staged architecture. *Journal of Experimental Psychology: General, 140,* 407–433. http://dx.doi.org/10.1037/a0023517

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass, 9,* 311–327. http://dx.doi.org/10.1111/lnc3.12151

Staub, A., & Goddard, K. (2019). The role of preview validity in predictability and frequency effects on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45,* 110–127. http://dx.doi.org/10.1037/xlm0000561

Staub, A., White, S. J., Drieghe, D., Hollway, E. C., & Rayner, K. (2010). Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 1280–1293. http://dx.doi.org/10.1037/a0016896

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica, 30,* 276–315. http://dx.doi.org/10.1016/0001-6918(69)90055-9

von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language, 94,* 119–133. http://dx.doi.org/10.1016/j.jml.2016.10.003

Wahlsten, D. (1991). Sample size to detect a planned contrast and a one degree-of-freedom interaction effect. *Psychological Bulletin, 110,* 587–595. http://dx.doi.org/10.1037/0033-2909.110.3.587

Wang, J., Li, L., Li, S., Xie, F., Liversedge, S. P., & Paterson, K. B. (2018). Effects of aging and text-stimulus quality on the word-frequency effect during Chinese reading. *Psychology and Aging, 33,* 693–712. http://dx.doi.org/10.1037/pag0000259

Warrington, K. L., McGowan, V. A., Paterson, K. B., & White, S. J. (2018). Effects of aging, word frequency, and text stimulus quality on reading across the adult lifespan: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44,* 1714–1729. http://dx.doi.org/10.1037/xlm0000543

Wheeler, D. D. (1970). Processes in word recognition. *Cognitive Psychology, 1,* 59–85. http://dx.doi.org/10.1016/0010-0285(70)90005-8

White, S. J., & Staub, A. (2012). The distribution of fixation durations during reading: Effects of stimulus quality. *Journal of Experimental Psychology: Human Perception and Performance, 38,* 603–617. http://dx.doi.org/10.1037/a0025338

Yap, M. J., & Balota, D. A. (2007). Additive and interactive effects on response time distributions in visual word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 274–296. http://dx.doi.org/10.1037/0278-7393.33.2.274

Yap, M. J., Balota, D. A., Tse, C. S., & Besner, D. (2008). On the additive effects of stimulus quality and word frequency in lexical decision: Evidence for opposing interactive influences revealed by RT distributional analyses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 495–513. http://dx.doi.org/10.1037/0278-7393.34.3.495

Ziegler, J. C., Perry, C., & Zorzi, M. (2009). Additive and interactive effects of stimulus degradation: No challenge for CDP+. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 306–311. http://dx.doi.org/10.1037/a0013738

Received December 18, 2019
Revision received June 10, 2020
Accepted June 11, 2020 ■