Journal of Memory and Language 116 (2021) 104190

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

How reliable are individual differences in eye movements in reading?☆

Adrian Staub

Department of Psychological and Brain Sciences, University of Massachusetts Amherst, 430 Tobin Hall, Amherst, MA 01003, United States

ARTICLE INFO

Keywords:
Eye movements
Reading
Reliability
Individual differences

ABSTRACT

This study assessed the reliability of individual differences among fluent adult readers in the effects of four variables - word frequency, predictability, visual contrast, and font difficulty – on eye fixation duration measures, word skipping probability, and regression probability. Split-half reliability was computed in a reanalysis of data from two large, previously published experiments (Staub, 2020) by correlating simple effects in two halves of each experiment (e.g., Hedge, Powell, & Sumner, 2018) and by estimating, in the context of mixed-effects models, a correlation parameter between by-subject slopes for each half (Rouder & Haaf, 2019). The reliability of the effects was generally low, though the second of these methods revealed a few notable exceptions. First, the effects of visual contrast were quite reliable, as expected based on presumed individual differences in contrast sensitivity. Second, the frequency effect on gaze duration was also reliable, but only when raw (as opposed to log) gaze duration was used as the dependent measure. The effect of predictability demonstrated poor reliability for all dependent measures. Model comparison confirmed that model fit was improved by inclusion of by-subject slopes for those effects that showed substantial reliability. These results have implications for the feasibility of studies on individual differences in eye movements in reading, as only experimental effects that demonstrate substantial reliability are good candidates to be explored in individual difference studies.

Introduction

We have learned a great deal about the factors that influence how a reader's eyes progress. The eyes spend longer on an infrequent word than on a frequent one (e.g., Rayner & Duffy, 1986; Staub, White, Drieghe, Hollway, & Rayner, 2010), and longer on a word that is not predictable based on its preceding context than on one that is predictable (e.g., Ehrlich & Rayner, 1981; Staub, 2011). These two factors also influence the probability that a word is skipped altogether, rather than directly fixated. Fixations on a word are longer when text is faint (e.g., Reingold & Rayner, 2006; White & Staub, 2012), and when readers are denied parafoveal preview, i.e., when they are not able to see the word while fixating the previous word (e.g., Schotter, Angele, & Rayner, 2012). In addition, the forward progress of the eyes is rapidly disrupted when a word is not easily integrated into a representation of the sentence's syntactic structure (e.g., Frazier & Rayner, 1982) or meaning (e.g., Rayner, Warren, Juhasz, & Liversedge, 2004). All of these findings have informed theories of the perceptual, cognitive, and linguistic processes involved in reading.

These findings have been established by means of the within-subject experimental design that dominates cognitive psychology, in which each subject is exposed to a number of trials at each level of a critical experimental variable. For example, each subject will read a number of sentences containing a target word that is high in frequency, as assessed by corpus counts, and an equal number containing a target word that is low in frequency, where the target words are matched on other variables such as word length and predictability. The frequency effect on the mean of fixation duration measures, or on the probability that a word is skipped, is established by means of a statistical model in which experimental subjects are treated as levels of a random factor, allowing for statistical generalization to a population of readers.

Unsurprisingly, the group-level effect of a variable such as word frequency obscures variation in the effects shown by individual subjects. If, for example, the mean frequency effect on gaze duration (the sum of a reader's eye fixations on a word, on first pass reading) is 30 ms in a particular experiment, computed from a 250 ms mean for high

☆ Thanks to audiences at the Language and Cognition Brown Bag, Department of Psychological Sciences, University of Connecticut; the Language and Cognition Colloquium, Department of Psychology, Harvard University; and the UMass Psycholinguistics Workshop. Thanks also to Chuck Clifton and Brian Dillon for helpful discussion, and to Marc Brysbaert for exceptionally thoughtful comments on an earlier version. This work was supported by a grant from the U.S. National Science Foundation (BCS 1732008). Materials, data, and analysis scripts are available at: https://doi.org/10.17605/OSF.IO/BNAVZ.

E-mail address: astaub@umass.edu.

https://doi.org/10.1016/j.jml.2020.104190
Received 10 February 2020; Received in revised form 5 November 2020; Accepted 7 November 2020
Available online 21 November 2020

frequency words and a 280 ms mean for low frequency words, for some subjects the frequency effect will be very large, perhaps greater than 60 ms, and for some subjects the frequency effect will be rather small, or even nonexistent. Indeed, it is typical for a few subjects to show a reversed effect, with a longer average reading time on high frequency words.

While interest in the sources of this individual variability is not new (e.g., Ashby, Rayner, & Clifton, 2005; Calvo, 2001; Chace, Rayner, & Well, 2005; Kennison & Clifton, 1995), the number of studies focusing on this issue has increased dramatically in recent years. These studies have examined how various effects on eye movements in reading may be modulated by readers' working memory capacity (e.g., Kuperman & Van Dyke, 2011; Traxler, 2007; Traxler et al., 2012), their reading skill (e.g., Kuperman & Van Dyke, 2011; Slattery & Yates, 2018; Taylor & Perfetti, 2016; Veldre & Andrews, 2014, 2015a, 2015b, 2016), and their language experience (e.g., Falkauskas & Kuperman, 2015; Gordon, Moore, Choi, Hoedemaker, & Lowder, 2019; Schmidtke, Van Dyke, & Kuperman, 2018; Whitford & Titone, 2012). Early studies often used a two-group design, sometimes based on a median split of an individual difference variable (e.g., Ashby et al., 2005), but many recent studies have used mixed-effects regression models to explore the relationship between individual difference variables and experimental effects of interest, across the full range of the individual difference variables.

This research requires a critical assumption that is rarely made explicit. It is only possible to meaningfully assess the relationship between an individual difference variable, such as working memory or language experience, and the size of an experimental effect, such as the effect of word frequency, if the observed variability in the size of the effect reflects stable differences between individuals. In other words, differences between individuals in the size of the effect must be *reliable*, in the psychometric sense. The effect should demonstrate reasonably high *split-half reliability*, which is the correlation between the effects shown by individual subjects in two halves (divided by order or by, e.g., even vs. odd trials) of a testing session, or *test–retest reliability*, which is the correlation between individuals' effects in separate testing sessions. Poor reliability indicates that much of the observed variability in the size of the effect is not due to real differences between individuals, but to statistical noise. Researchers have offered guidelines, though not always consistent ones, regarding the level of reliability that is necessary in order for a measure to be useful in individual differences research. Enkavi et al. (2019) cite a 'common criterion' of .75 test-retest reliability, while Hedge, Powell, and Sumner (2018) remark that .6 is often considered 'good'. The quantitative relationship between reliability and the magnitude of observable correlations between measures has long been recognized (Nunnally, 1970; Spearman, 1904), as have the consequences of poor reliability for the statistical power of individual difference studies (e.g., Hedge et al., 2018).

The goal of the present study is to assess the reliability of individual differences in some of the most robust linguistic effects on eye movements in reading, the effects of word frequency and predictability, as well as the reliability of individual differences in the effects of visual contrast (e.g., White & Staub, 2012) and font difficulty (Rayner, Reichle, Stroud, Williams, & Pollatsek, 2006). We reanalyze data from two large-scale eye movement experiments in which these variables were factorially manipulated (Staub, 2020).

Several recent empirical and methodological developments inform the present work. In a recent but already widely cited study, Hedge et al. (2018) suggested that many extremely robust – i.e., replicable – cognitive effects, such as the Stroop effect, demonstrate disappointingly poor reliability. Hedge et al. had participants complete a very large number of trials in seven tasks (e.g., 240 in each of the congruent, incongruent, and neutral conditions of the Stroop task), in each of two sessions. The Intraclass Correlation Coefficient between the sessions (a measure of reliability on the same scale as the Pearson correlation, which we report in the present article) ranged widely across the critical effects. But even with such a large number of trials per subject these reliabilities rarely if ever reached the desirable range, as the maximum reliability of any RT effect was .70, for the Posner cueing task (Posner, 1980), and for many of the tasks the reliability of RT effects was very low; in the Navon global-local task (Navon, 1977), for example, the reliability of the RT effects was near 0.

Hedge et al. (2018) interpreted these findings as reflecting a 'reliability paradox': Effects that are large and highly replicable, at the group level, will often show poor test-restest or split-half reliability. This is for a simple mathematical reason. Part of what makes an effect replicable at the group level, and makes it a 'large' effect based on standardized measures of effect size such as Cohen's *d*, is that subjects do not show much variation in their response to the experimental manipulation. As a result:

> [I]t should not be assumed that robust experimental paradigms will translate well to correlational studies. In fact, they are likely to be sub-optimal for correlational studies for *the same reasons* that they produce robust experimental effects. Our findings, as well as observations from elsewhere in the literature, indicate that this challenge currently exists across most domains of cognitive psychology and neuroscience (p. 1177; original italics).

But as Hedge et al. also point out, it is not impossible for an effect that is highly replicable, at the group level, to show reliable between-subject variation; it is an empirical question, for any effect of interest, whether this situation does obtain.

Recently, however, Rouder and Haaf (2019; see also Kliegl, Wei, Dambacher, Yan, & Zhou, 2011; Rouder, Kumar, & Haaf, 2019) have suggested that this picture is unnecessarily pessimistic. Rouder and Haaf critique the method used to assess reliability by Hedge et al. (2018) and many others, suggesting that this method systematically underestimates the true reliability of experimental effects. They argue that an alternative statistical approach is both theoretically justified and likely to result in higher reliability estimates.

Hedge et al. (2018) deployed a standard method in which the data for each subject are aggregated into a single value for each session. For example, each subject's Stroop effect is measured as a single difference score: mean incongruent RT – mean congruent RT. The correlation between subjects' difference scores in the two sessions is the measure of reliability. It is important to note that with this method, the observed reliability of an experimental effect is strongly related to the number of trials that each subject completed in each condition, with reliability increasing as the number of trials increases. This is because of the relationship between sample size and the variability of sample means. As sample size (i.e. number of trials per condition) increases, the subject's mean RT in each of the two critical conditions will vary less around the subject's 'true' mean, and the difference score that measures the Stroop effect will vary less around the subject's 'true' difference score. Even if an effect would be highly reliable in the limit - when each subject is exposed to a very large number of trials in each condition - reliability may be quite poor with a smaller number of trials, as the measure of the effect in each session, or in each half of a session, will vary substantially around the subject's true effect. See Miller and Ulrich (2013) for extensive discussion and modeling of how the reliability of a response time effect depends on trial numbers, under varying assumptions about the mechanisms underlying the effect.

Rouder and Haaf (2019) propose, instead, to estimate individual subject effects using hierarchical models such as the linear or logistic mixed-effects models that are familiar to many psycholinguists (Baayen, Davidson, & Bates, 2008). When the random effects in such a model include by-subject slopes for an experimental effect, each slope reflects the model's estimate of how the effect for an individual subject differs from the overall effect across subjects, i.e., the fixed effect. Critically, the model simultaneously estimates trial noise: trial-to-trial variability that is not related to the subject differences themselves. As a result, the model's estimates of individual subject effects will be less variable than

the non-model-based estimates of the same effects, as the model attributes some of the variability in subjects' effects to trial noise. This is known as model-based *shrinkage* or *regularization* (Efron & Morris, 1977). At the same time, because trial noise has been excluded from these model-based estimates, the estimates can be expected to be more reliable than non-model-based estimates such as a difference between condition means, and to be less dependent on the number of trials than the corresponding non-model-based estimates. Reliability of model-based estimates of individual subject effects can be directly estimated by a separate model parameter estimating the correlation between by-subject effects in two sessions, or two halves of a session. If the hierarchical model is Bayesian, we will also have information about the precision of this reliability estimate in the form of the full posterior distribution of the correlation parameter. Note that a model-based estimate of a correlation between subject effects, as suggested by Rouder and Haaf (2019), is not expected to be identical to the hand-computed correlation between model estimates of subject effects. Kliegl, Masson, and Richter (2010) illustrate by means of simulation that the former method accurately recovers known correlations, while the latter method exaggerates these correlations.

Rouder and Haaf (2019) demonstrate these points by means of a reanalysis of the Hedge et al. (2018) data. When Stroop effect reliability is estimated based on a small quantity of data - only one of the five blocks of trials in each session - the original non-model-based method deployed by Hedge et al. reveals almost no reliability at all (.10), while the posterior mean of the correlation parameter in a Bayesian hierarchical model reveals modest reliability (.31). When all the data are used, the non-model-based method reveals reliability of .55, while the model-based method reveals reliability of .72. It is important to note, however, that the uncertainty of the critical correlation parameter in Rouder and Haaf's models is substantial, especially when the quantity of data is limited. This point will be relevant to the present study, as well.

Informed by this recent literature, here we assess reliability in several ways. First, we use the traditional method of calculating the correlation between subjects' simple effects in two halves of each experiment, as in Hedge et al. (2018). Second, we fit Bayesian hierarchical mixed-effects models that estimate subject effects in each half and explicitly estimate the correlation between the effects for the two halves (Rouder & Haaf, 2019). Finally, we also directly assess the evidence for variation between subjects in the size of each effect by determining the extent to which including by-subject slopes improves the fit of hierarchical mixed-effects models, when compared to models without by-subject slopes. While the question of whether to include a set of random slopes in a mixed-effects model is usually considered in the context of attempts to optimize power and reduce Type I error rate in testing fixed effects (Barr, Levy, Scheepers, & Tily, 2013; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017), assessing the improvement in model fit when by-subject slopes are included also addresses the substantive question of whether there is meaningful variation between subjects in the size of an effect.

Before presenting the details of the present study, we discuss the one previous study of eye movements in reading that has directly investigated the reliability of individual subject effects. Carter and Luke (2018) analyzed test-retest reliability of frequency, predictability, and word length effects for 39 subjects reading 40 paragraphs from the Provo Corpus (Luke & Christianson, 2016); 20 paragraphs were read in each of two sessions, separated by about one month. Eye movement data for all words in the corpus were included in the analysis, with the exception of data contaminated by blinks and return sweep saccades and fixation durations removed by outlier exclusion. The effects of each variable on individual subjects' eye movements were assessed by means of by-subject slopes extracted from mixed-effects models of each eye movement measure, fit separately to the data from each of the two sessions. The correlation between by-subject slopes extracted from the separate models of the two sessions was the reported measure of reliability. As noted above, this method is not equivalent to model-based estimation of a correlation parameter (Kliegl et al., 2010).

The central results were as follows. For first fixation and gaze duration, the effects of all three predictors showed reliability above $r = .6$, with the highest reliability ($r = .78$) appearing for the effect of predictability on first fixation duration. The reliability of the effects on skipping probability and regression probability were all above $r = .5$, with the effect of predictability on regression probability demonstrating the highest reliability ($r = .64$). In sum, based on these results it would appear that individual differences in the effects of length, frequency, and predictability are all at least moderately reliable, for all eye movement measures.

However, the interpretation of these reliabilities should be qualified by an important caveat. The three variables assessed by Carter and Luke (2018) are strongly intercorrelated in natural texts. Shorter words are both more frequent and more predictable, and more frequent words are more predictable (e.g., Piantadosi, Tily, & Gibson, 2011; Smith & Levy, 2011). Indeed, Luke and Christianson (2016) demonstrated that within the Provo Corpus itself, the predictability of a word is significantly related to both word frequency and word length. Moreover, in natural texts these variables are also correlated with other factors such as part of speech and position in the sentence, as also demonstrated by Luke and Christianson (2016). However, Carter and Luke (2018) assessed the reliability of each effect in a separate statistical model, e.g., one model included length as the only fixed effect, along with random slopes for length, while a second model included frequency as the only fixed effect, along with random slopes for frequency. Given the correlation between the variables, any one of these models will not uniquely identify the effect of one variable, but will instead identify some combination of the effects of all three variables. Even if more complex models were used to simultaneously estimate the reliability of the effects of the three critical variables, interpretation would be complicated by, for example, the fact that almost all of the fifty most frequent words in English are very short closed-class words such as prepositions, determiners, pronouns, and conjunctions. In the present study, by contrast, the frequency, predictability, and contrast/font difficulty variables are entirely uncorrelated, and are uncorrelated with other variables such as part of speech, as the present study reanalyzes data from experiments in which the variables were factorially manipulated. Specifically, the predictability manipulation was implemented by presenting the same word in two different sentence contexts, while the frequency manipulation compared high- and low-frequency words that were closely matched in predictability, as well as word length and part of speech.

The results of one other existing study might also be taken to suggest, albeit indirectly, that the frequency effect is highly reliable. Schilling, Rayner, and Chumbley (1998) addressed the question of how word frequency effects in the lexical decision and naming tasks compare to frequency effects on eye movements in reading, by having each individual subject complete a pair of tasks. Schilling et al. found correlations of about .54 between the size of a subject's frequency effect on mean gaze duration and both the size of the frequency effect on lexical decision RT and the size of the effect on naming RT. These correlations were obtained despite the fact that subjects read only 24 words at each level of word frequency, in each task. Observed correlations of this magnitude are only possible if individual differences in each of the two correlated effects (e.g., the frequency effect on gaze duration and on lexical decision RT) are highly reliable; we return to this point in the General Discussion.

However, there is an idiosyncratic feature of the Schilling et al. (1998) study that has, to our knowledge, not been noted in the literature. Subjects did not merely read sets of high-frequency and low-frequency words in each of two tasks; they read the very same 24 high-frequency and 24 low-frequency words in the two tasks. Thus, what is demonstrated to be consistent across tasks, in the Schilling et al. study, is not subjects' frequency effect, in general, but rather their difference in responses to a specific set of high- and low-frequency words. Arguably, it is unsurprising that a subject who responds to a particular word quickly

(or slowly) in one task also responds to the same word quickly (or slowly) in another task.

The remainder of this paper proceeds as follows. In the next section, we provide an overview of the data sets from Staub (2020) that we reanalyze in the present study. In the following section, we illustrate that individual differences in basic eye movement measures such as mean first fixation duration (as opposed to experimental effects on these measures) show very high reliability, confirming several previous studies. In the next three sections we carry out the three reliability analyses that we have outlined above. In the General Discussion, we turn to the implications of the results of these analyses for our understanding of experimental effects on eye movements in reading, and for individual difference studies.

Overview of data sets

The two data sets that we use to assess reliability of eye movement effects come from recent experiments published in Staub (2020). These experiments were designed to investigate potential interactions between word frequency, predictability, and visual contrast of the text (Experiment 1) and between word frequency, predictability, and font difficulty (Experiment 2). Data and experimental items are available at https://doi.org/10.17605/OSF.IO/BNAVZ.

These data sets have several desirable characteristics for present purposes. First, both experiments demonstrated large, statistically significant group-level effects of all of the manipulated variables, replicating a number of well-established effects in the literature. Below we discuss the details of these group-level effects. Second, each experiment included 80 trials at each level of each variable (e.g., 80 high-frequency and 80 low-frequency target words), which is many more than in most eye movement studies; we know of only a few studies that have used similar trial numbers to investigate effects of frequency (e.g., Sheridan & Reingold, 2013) or predictability (Staub & Benatar, 2013). Thus, the split-half reliabilities that we compute here may be seen as establishing an upper bound on the split-half reliabilities that would be obtained in a typical experiment. They may even be regarded as optimistic estimates of test-retest reliabilities, as in many eye movement studies there are fewer than 40 trials at each level of a manipulated variable.

Third, the two experiments are identical with respect to their frequency and predictability manipulations, as they use the same experimental items. In effect, we are able to internally replicate reliability estimates for the effects of frequency and predictability. Finally, each experiment had a large number of participants, N = 80 in Experiment 1 and N = 92 in Experiment 2, and the participants in these experiments were in many respects typical of adult participants in eye movement experiments. Participants were native English-speaking undergraduates at the University of Massachusetts Amherst, who self-reported no history of reading or language disorder, and normal or corrected-to-normal vision. UMass Amherst is a large, moderately selective public university that in 2019 admitted 64% of applicants. The first and third quartiles of the SAT score distribution of admitted students in 2019 were 1220 and 1380, respectively.

Subjects in both experiments read 160 critical sentences, which were adopted from Kretzschmar, Schlesewsky, and Staub (2015). In each sentence, a target word's frequency and predictability were manipulated, such that each subject read 80 target words at each of two levels of frequency, 80 at each level of two levels of predictability, and 40 at each level defined by the interaction of these variables. In Experiment 1, half of the 40 sentences at each of these levels were presented in clear text and half in faint text; in Experiment 2, half were presented in Times New Roman font and half in Old English font. In Experiment 1, 4.0% of trials were removed due to blinks, track loss, or other error, leaving a total of 12,285 trials for analysis. In Experiment 2, 4.5% of trials were removed, leaving 14,064 for analysis.

For the purposes of the split-half analyses that we present here, we divide the sentences based on odd vs. even item numbers in the experimental script. Given the construction of these experimental scripts,[1] this split results in two halves that are perfectly balanced for each subject, i.e., each half includes 80 items whose distribution matches that of the 160 items as a whole, with 40 items at each level of each variable. The even/odd item split does not reflect alternating presentation of trials, as items were presented in a random order to each subject, with the exception of blocking of the visual contrast and font difficulty variables.

For the purposes of reliability analyses, we focus on four eye movement measures. These include two reading time measures: first fixation duration (the duration of a reader's initial first-pass eye fixation on a word) and gaze duration (the sum of all first-pass fixations on a word, before leaving it). On many trials, first fixation duration and gaze duration are identical; they will differ only if readers made an additional fixation after the first, before leaving the word. We also analyze two binary saccadic measures: skipping probability (the probability that a word is skipped on first pass reading rather than directly fixated) and regression probability (the probability that the reader's initial inspection of a word ends with a saccade to the left, rather than to the right). Together, these four measures present a fairly complete description of a reader's eye movements on her initial encounter with the target word. If a word is skipped on first pass reading, that trial is excluded from the computation of regression proportion as well as the fixation duration measures. Thus, while the skipping analysis is based on an average of 154 trials per subject in Experiment 1 and 153 trials in Experiment 2, after data exclusion, the other analyses, which also exclude word skips, are based on an average of 120 trials per subject in Experiment 1, and 115 trials per subject in Experiment 2.

The group-level effects on the four eye movement measures, as assessed based on the mixed-effects linear or logistic regression models in Staub (2020), are shown in Table 1. The critical results are as follows. The effects of frequency and predictability replicated many previous studies: In both experiments, both variables had sizable and significant effects on word skipping probability (in the range of a 4–5% effect, for both variables, for both experiments; note that parameter estimates in Table 1 are in log odds), and on both first fixation duration (approximately 20 ms effects, for both variables, in both experiments) and gaze duration (effects between 28 and 38 ms). Low frequency and low predictability words were less likely to be skipped and induced longer reading times. In addition, predictability, but not frequency, significantly influenced regression probability in both experiments, with low predictability words inducing about 5% more regressions. As in most previous studies (see Staub, 2015, for review), the frequency and predictability effects did not significantly interact in any measure.

The visual contrast manipulation in Experiment 1 strongly influenced both of the reading time measures, by about 51 ms in first fixation duration, and 62 ms in gaze duration, but did not significantly affect either skipping probability or regression probability. There were also statistically significant contrast-by-frequency interactions in the skipping measure and the gaze duration measure. However, these were very small effects, in opposite directions (the frequency effect on skipping was smaller with faint text, but the frequency effect on gaze duration was slightly larger with faint text), and neither would be significant if a correction for multiple comparisons were applied. Thus, these interactions should be interpreted with caution.

The font difficulty manipulation in Experiment 2 influenced the reading time measures, by about 16 ms in first fixation duration and 32

[1] Specifically, items 1–20 and 81–100 were low frequency, high predictability; 21–40 and 101–120 were low frequency, low predictability; 41–60 and 121–140 were high frequency, high predictability; and 61–80 and 141–160 were high frequency, low predictability. Items 1–80 were presented in a random order in the first block, and 81–160 in the second. The ordering of visual contrast and font difficulty blocks was counterbalanced across subjects, e.g., half of subjects read clear text first, and half read faint text first.

Table 1
Mixed-effects model parameter estimates from Staub (2020). First Fixation and Gaze estimates are in ms; Skipping and Regressions estimates are in log odds; ***$p < .001$; **$p < .01$; *$p < .05$.

	Experiment 1				Experiment 2			
	First Fix	Gaze	Skipping	Regressions	First Fix	Gaze	Skipping	Regressions
Predictability	19.45***	28.33***	−.33***	.38***	19.25***	30.92***	−.35***	.46***
Frequency	19.06***	33.66***	−.30*	−.01	20.68***	37.91***	−.47***	.01
Contrast/Font Difficulty	51.25***	61.95***	−.12	−.12	16.15***	32.34***	−.41***	−.08
Pred × Freq	−6.52	−1.45	.07	−.03	−1.32	5.56	.06	−.17
Pred × Contrast/Font	3.76	7.19	.04	−.18	5.08	16.12**	−.08	.01
Freq × Contrast/Font	6.13	9.68*	.23*	−.00	7.46*	20.06**	−.11	.02
Pred × Freq × Contrast/Font	−.81	−2.50	−.01	.20	−7.49	−6.13	−.45**	.08

ms in gaze duration, as well as word skipping probability, by about 5%. In addition, this manipulation interacted significantly with both predictability and frequency in the gaze duration measure, with both variables having a larger effect on gaze duration when the text was presented in Old English font. These interactions were substantial in size (16–20 ms), and were also present in the go-past time measure, which we do not analyze here. Thus, we regard these interactions as likely to reflect real effects. However, we do not attempt to interpret an unpredicted three-way interaction in the skipping measure.

Basic eye movement measures: Distributions, reliability, and intercorrelations

Figs. 1 and 2 illustrate, for Experiments 1 and 2 respectively, the distributions, across subjects, of skipping proportion, regression proportion, and mean first fixation duration and gaze duration. There is substantial variation in subjects' saccadic behavior. Some subjects almost never skipped the target word, and some subjects skipped it over 50% of the time; some subjects almost never regressed from the target word, and others regressed 40% of the time. There is also substantial variation in fixation durations, with the lowest and highest mean gaze durations differing by about a factor of two.

Code for reliability analyses in this and subsequent sections is available at: https://doi.org/10.17605/OSF.IO/BNAVZ. The scatterplots in Figs. 3 and 4 illustrate the reliability of individual subject differences in each of the four measures, in each experiment. All measures show high split-half reliability; the lowest value is for regression proportion ($r = .75$ in Experiment 1; $r = .76$ in Experiment 2), while the other three measures show split-half reliability of at least .83 in both experiments. In sum, there is substantial variability between subjects in their basic eye movement behavior – the duration of their eye fixations, the frequency with which they skip words, and the frequency with which the make regressions - and these individual differences are highly reliable. These conclusions are entirely consistent with the conclusions of

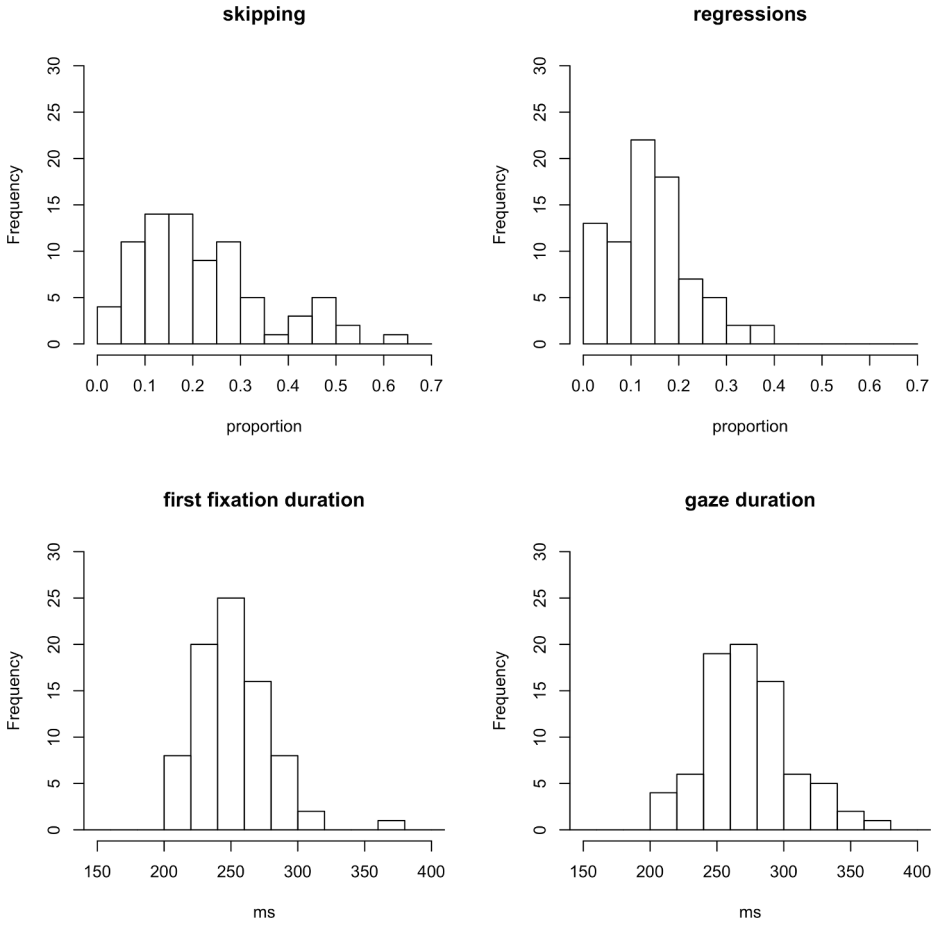

Fig. 1. Experiment 1 distributions of subject skipping and regression proportions, and first fixation and gaze duration means.

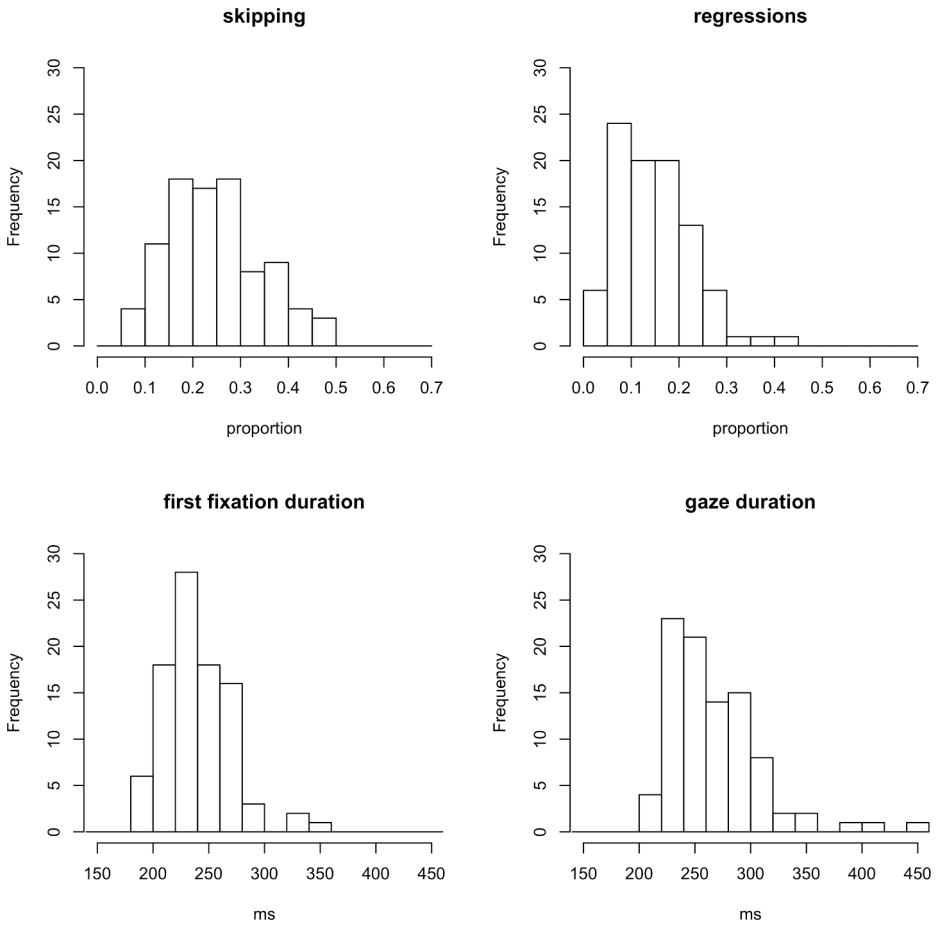

Fig. 2. Experiment 2 distributions of subject skipping and regression proportions, and first fixation and gaze duration means.

several other studies, which have found reliable differences between individuals in parameters such as mean fixation duration and saccade amplitude, both across tasks (e.g., in reading and scene viewing; Henderson & Luke, 2014; Rayner, Li, Williams, Cave, & Well, 2007) and across separate sessions of a reading task (Carter & Luke, 2018; Dirix, Brysbaert, & Duyck, 2019). Staub and Benatar (2013) also found substantial correlations between subjects' mean fixation duration on a target word in each experimental sentence and their fixation durations on preceding and following words.

An interesting question is whether all four measures demonstrate high reliability partly because of correlations among the measures themselves. For example, if mean first fixation duration is strongly correlated (either positively or negatively) with the tendency to skip words, then as long as one of these measures is highly reliable, the other will be, too. On the other hand, if fixation durations and skipping are only weakly correlated at the subject level, or even uncorrelated, then it is more notable that both measures are highly reliable, as this would suggest that there is meaningful variation between readers on two separate dimensions. To our knowledge, this issue has not been previously addressed.

Unsurprisingly, subjects' mean first fixation duration and gaze duration are very highly correlated (Experiment 1 $r = .92$, $p < .001$; Experiment 2 $r = .89$, $p < .001$), due to their mathematical relationship, i.e., these measures are identical on many trials. Of more interest are the relationships between the fixation duration measures and the saccadic measures. The fixation duration measures show a modest negative correlation with skipping proportion (Experiment 1 $r = -.20$; $p = .07$ and $r = -.31$; $p < .01$ for first fixation duration and gaze duration, respectively; Experiment 2 $r = -.27$; $p < .05$ and $r = -.41$; $p < .001$ for first fixation duration and gaze duration, respectively); readers who skip more also have somewhat shorter fixation durations. Fixation durations show essentially no correlation with regression proportion, in either experiment, with r ranging between -.01 and .13. Finally, skipping proportion and regression proportion show a modest positive correlation (Experiment 1 $r = .33$; $p < .01$; Experiment 2 $r = .17$; $p = .10$), with readers who skip more often also regressing somewhat more often.

In sum, there are weak correlations across subjects between skipping and fixation durations, and between skipping and regressions; though these relationships do not always reach significance, the trends are consistent. Fixation durations and regressions appear to be largely independent. The fact that these relationships between measures are either weak or nonexistent implies that the reliable differences between readers on the skipping, regressions, and fixation duration measures are not simply a reflection of a reliable difference on a single underlying dimension. Readers show reliable differences on several largely independent dimensions.

Reliability of Experimental Effects: Non-Model-Based Analysis

We now turn to estimation of the split-half reliability of individual differences in the magnitude of the experimental effects: word frequency, predictability, and visual contrast in Experiment 1, and word frequency, predictability, and font difficulty in Experiment 2. In this section we present non-model-based analyses, as in Hedge et al. (2018).

We begin by illustrating the distributions of the effects, across subjects, in the data as a whole; see Figs. 5 and 6. For the fixation duration measures, an effect is computed as the difference between condition means, e.g., mean first fixation on low frequency target words minus mean first fixation on high frequency target words. For the saccadic measures, an effect is computed as the difference in proportions between

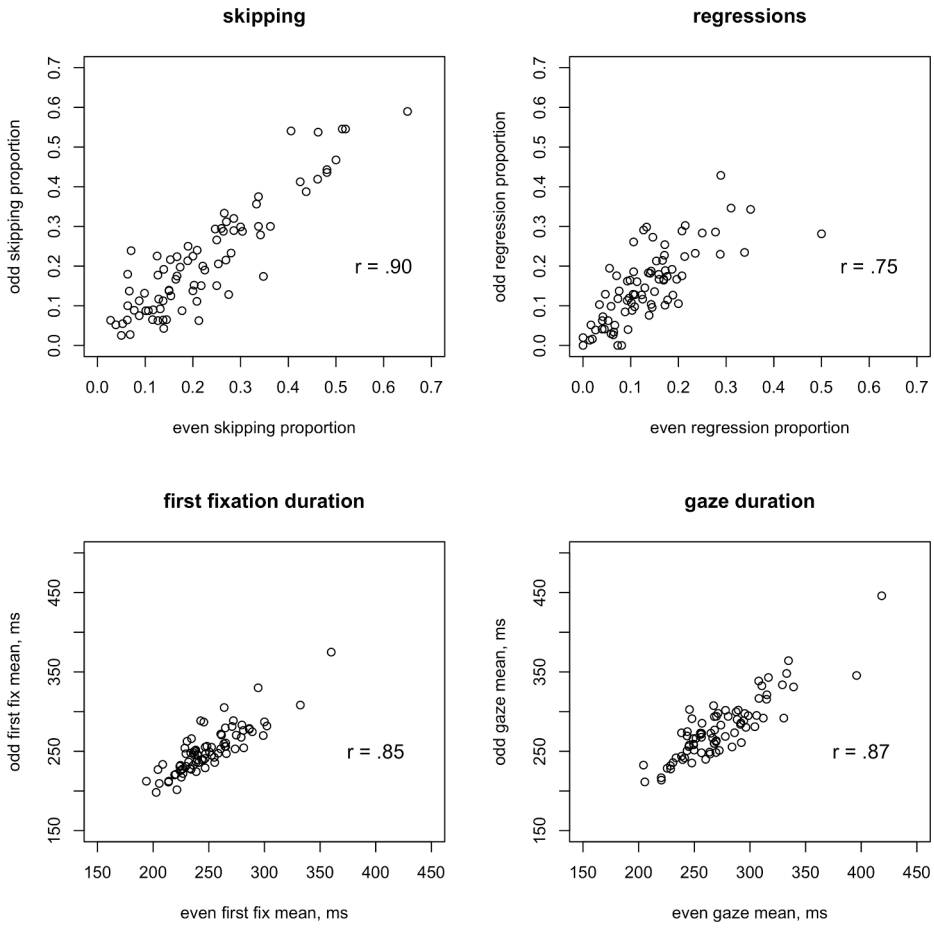

Fig. 3. Relationships in Experiment 1 between subjects' even- and odd-item skipping and regression proportions, and first fixation and gaze duration means.

the conditions, e.g., proportion of skips of low frequency words minus proportion of skips of high frequency words; in this case, the effect tends to be negative.

In all cases, the distribution of effects across subjects is unimodal and fairly symmetrical. The variance of the experimental effects is consistently much smaller than than the variance of the means of the fixation duration and saccadic measures themselves, shown in Figs. 1 and 2. With the exception of the visual contrast effect on reading times in Experiment 1 (where every single subject showed a numerical effect in the expected direction), in all cases there were some subjects who showed reversed effects, e.g., longer reading times for high-frequency words than low-frequency words. However, it is reasonable to assume that all of these effects demonstrate what Rouder and Haaf (2018) call *dominance*: For all subjects, the true effect is in the same direction. We assume that no reader actually reads low-frequency words faster than high-frequency words, or reads less predictable words faster than more predictable words. If dominance does hold, then the observed effects must overestimate the variability across subjects in the influence of the each of these variables. For example, the effect of word frequency on mean gaze duration in these experiments is about 35 ms, and if dominance holds, the distribution of true subject effects is constrained to have a lower bound of 0 ms. If we also assume that the distribution of true subject effects is fairly symmetrical, the largest true subject effect would then be about 70 ms. Thus, the range of the observed effects would overestimate the range of the true effects by a factor of two or more.

Split-half correlations of these effects were computed based on the even/odd item number split described above. In this non-model-based analysis, the effect in each half was again computed simply as a difference in condition means or proportions, and split-half reliability is simply the correlation between these effects in the two halves. These correlations are shown in the *Non-model-based* columns of Table 2, as well as on the scatterplots in Figs. 7 and 8. These split-half reliabilities are uniformly low, with .47 (for the contrast effect on skipping) as the maximum, and with several very near zero. Moreover, there is no particular relationship between split-half reliability and the significance or size of an effect at the group level. The numerically highest reliabilities are for the contrast effects on skipping and regressions, neither of which reached significance at the group level. Some significant and extremely large group-level effects show very low non-model-based reliability. For example, visual contrast demonstrated a 51 ms effect group-level effect on mean first fixation duration, while this effect has split-half reliability of .22.

Reliability of experimental effects: Model-based analysis

To implement a model-based reliability analysis (Rouder & Haaf, 2019), we constructed a set of Bayesian mixed-effects models using the *brms* package in R (Bürkner, 2017). The two fixed effects in each model were the effects of a single experimental manipulation in even items and in odd items. The effects in each half were coded with contrasts that assigned .5 to one level of the critical factor and -.5 and the other, and assigned 0 to the trials that were in the other half. For example, for the effect of frequency in the even items, the even, low-frequency trials were coded with .5, the even, high-frequency trials with -.5, and the odd trials with 0. The random effects in each model included a by-subject intercept, by-subject slopes for each of the two fixed effects, and parameters for the three correlations among the intercept and the two slopes. Thus, the general schema was as follows:

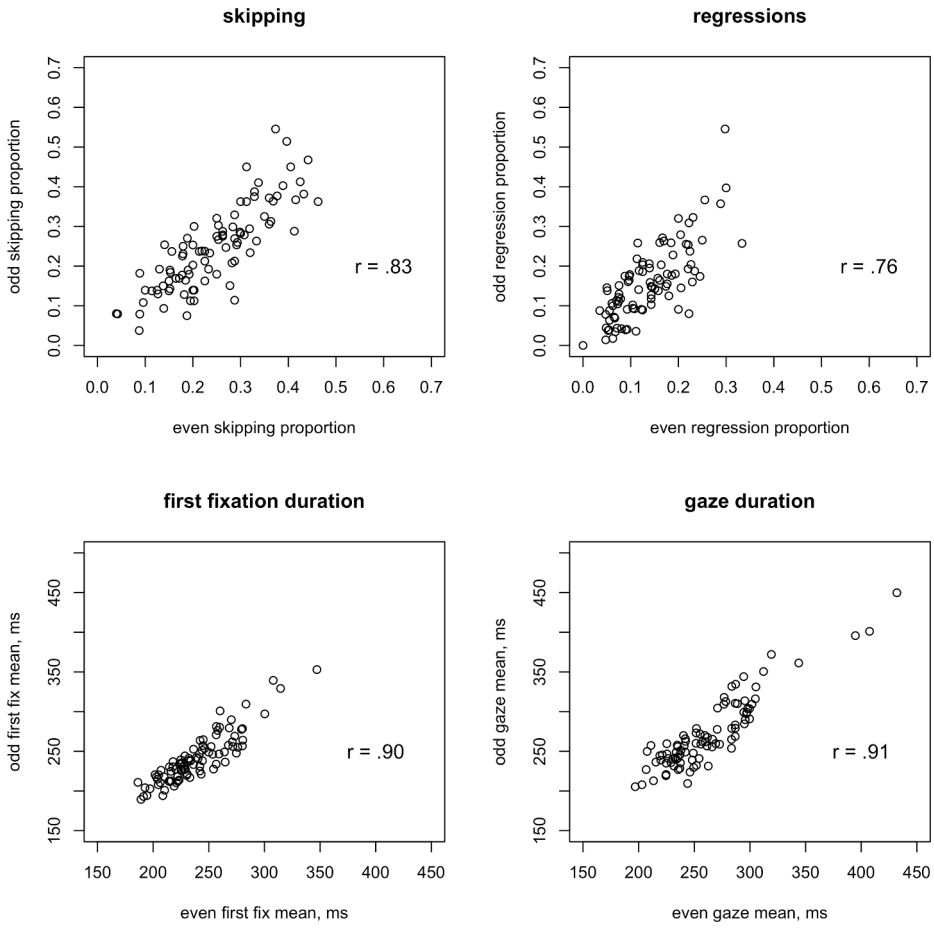

Fig. 4. Relationships in Experiment 2 between subjects' even- and odd-item skipping and regression proportions, and first fixation and gaze duration means.

`brm(measure ~ even.effect + odd.effect + (1 + even.effect + odd.effect|subject))`

For the reading time measures these were linear models, and for the skipping and regressions measures these were logistic models, i.e., with a binomial link function. All models were fit with 4 chains, each with 4000 iterations, of which 2000 were warmup. The default priors were used; in the case of the critical correlation parameters, this default is an LKJ prior (Lewandowski, Kurowicka, & Joe, 2009) with eta = 1. It is important to note that with this value of the eta parameter, the entire range of correlations between −1 and 1 has roughly uniform probability. With higher values of eta, values near −1 and 1 are regarded as unlikely. Thus, the current choice of prior does not prevent very strong posterior correlations, if they are justified by the data. All parameter estimates had R-hat values no higher than 1.01, and all of the critical correlation parameters had R-hat values of exactly 1, indicated good convergence. The fixed effects in these models were not of primary interest, but it is worth noting that these models confirmed all the effects reported in Staub (2020), and as expected, there was no indication that the effects differed in size in the two halves of the data.

For the reading time measures of first fixation duration and gaze duration, we constructed two versions of each model, one using the raw measure, and the other using the log measure. We note that the log transformation clearly results in more nearly normal residuals, and posterior predictive checks confirmed that the model predictions from the log models were qualititatively good, while the predictions from the raw models were not, as the predictions do not capture the right skew in the raw fixation duration distributions. Nevertheless, we report both models, as it will become clear below that there are some interesting differences.

The posterior mean of the critical correlation parameter from each model (i.e., the correlation between the subject-level estimates of `even.effect` and `odd.effect`), along with the 95% HDIs of the posterior distribution, are shown in Table 2, in the *Model-based* columns. We note, first of all, that though the model-based reliability estimates are generally somewhat higher than the non-model-based correlations, this is not universally true, and in many cases the difference in reliability is quite small. We also note that the 95% HDIs are extremely wide, in many cases spanning more than half of the potential range of the parameter from −1 to 1, and in a few cases spanning almost the entire range. Thus, the models are highly uncertain about the value of the critical correlation parameter (see Rouder et al., 2019, for discussion and analysis of this phenomenon).

However, there are a few model-based estimates of reliability that are (a) much higher than the corresponding non-model-based estimates, (b) relatively high in absolute terms, and (c) have a 95% HDI that does not include 0. These have been bolded in Table 2. The estimates of the reliability of the effect of visual contrast on gaze duration were .64 in the raw model and .72 in the log model, and the estimate of the reliability of contrast on word skipping was .84. The effect of font difficulty on word skipping was similarly reliable (.83). Finally, the reliability of the effect of word frequency on raw gaze duration was relatively high in both experiments: .68 in Experiment 1, and .75 in Experiment 2. Notably, the reliabilities extracted from the models of log gaze duration are much lower, with HDIs that include a large region of negative values. In the General Discussion, we address the question of why reliability may be higher in models of raw gaze duration.

In sum, computing model-based reliability does provide an indication that certain effects may be more reliable than the non-model-based analysis would indicate. However, this is a limited set of effects. In particular, the effect of predictability was not shown to be very reliable,

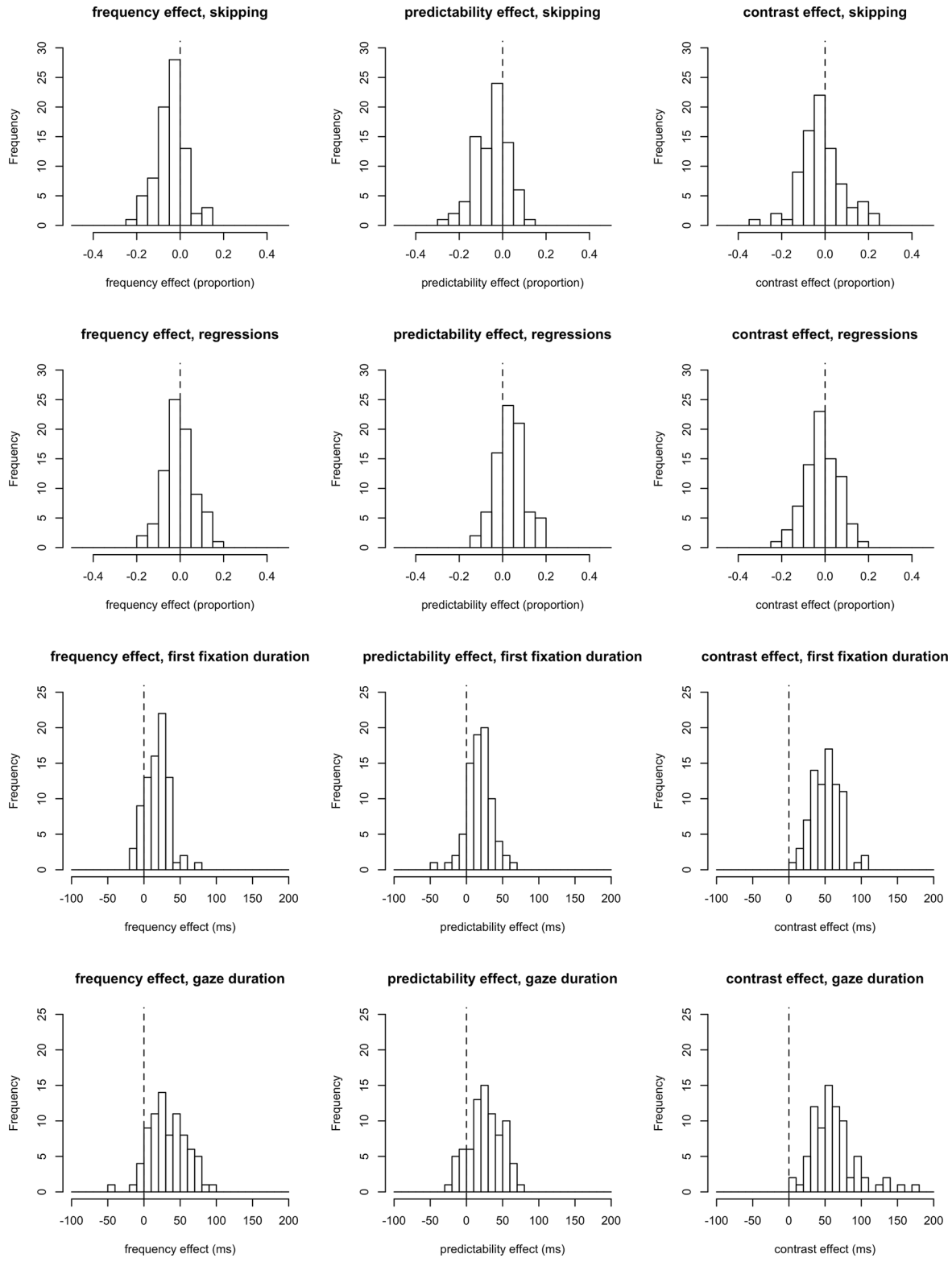

Fig. 5. Experiment 1 distributions of effects of each of the three experimental manipulations on subject skipping and regression proportions, and first fixation and gaze duration means.

using this analysis, and the effect of word frequency demonstrated substantial reliability only for the gaze duration measure - and even then only for raw, as opposed to log, gaze duration. Visual contrast demonstrated quite reliable effects on both gaze duration and skipping, and font difficulty demonstrated a reliable effect on skipping.

Do Random Slopes Improve Model Fit?

As noted above, the question of the appropriate random effect structure for a mixed-effects model is usually addressed in discussions of Type I error rate (Barr et al., 2013) or statistical power (Matuschek et al., 2017), when the researcher's primary interest is in testing the model's

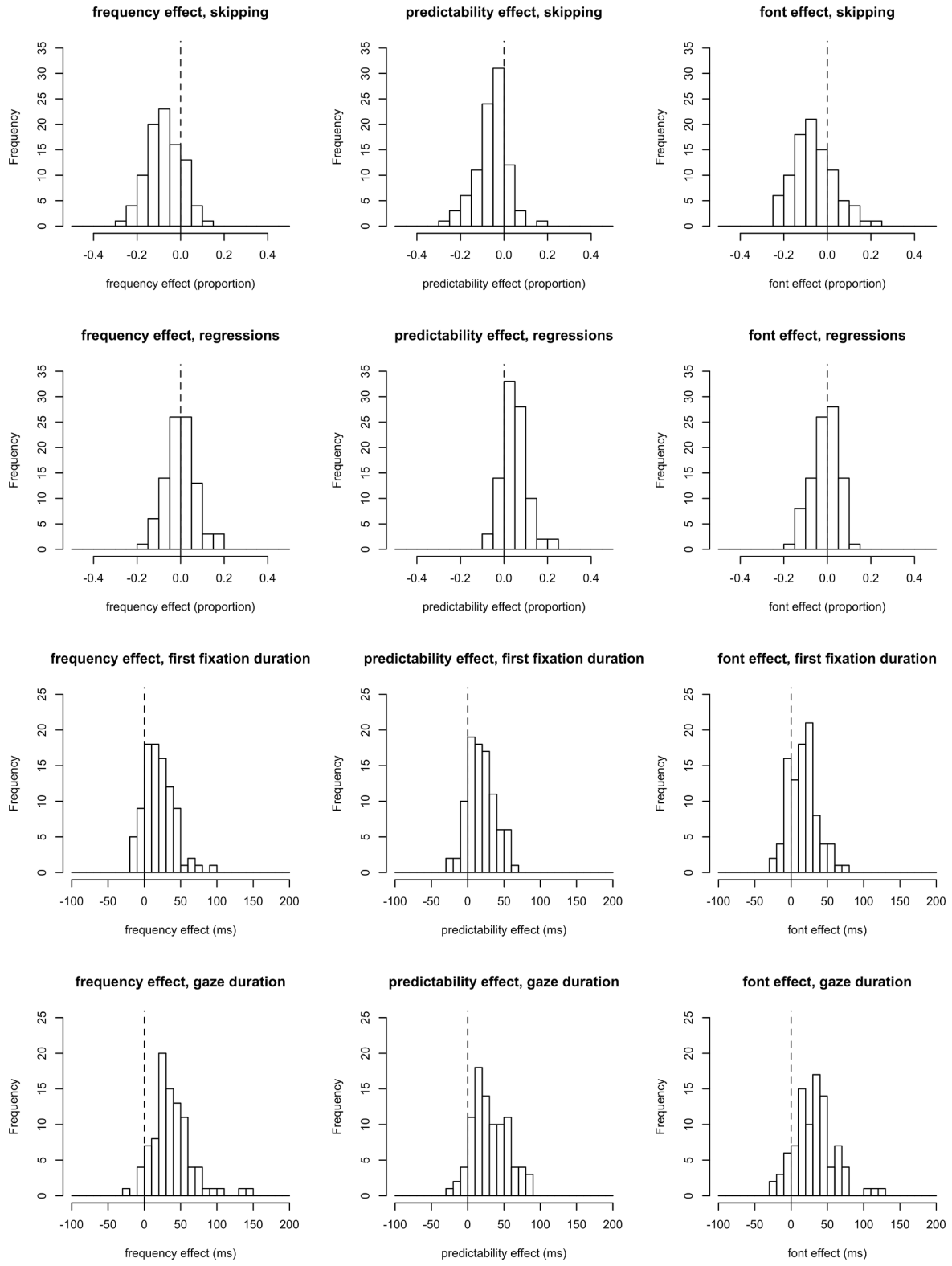

Fig. 6. Experiment 2 distributions of effects of each of the three experimental manipulations on subject skipping and regression proportions, and first fixation and gaze duration means.

fixed effects. In that context, it is of methodological but not theoretical interest whether a model that includes a given set of random effects results in an improvement in model fit over a model that does not. However, a comparison of the fit of nested models differing in the presence vs. absence of a set of by-subject slopes also addresses the substantive issue that is the focus of the present study: Is variability in the degree to which subjects show an effect attributable to genuine subject differences, as opposed to random sampling variability? The null hypothesis, exemplified in the model without by-subject slopes, is that all subjects show an effect to exactly the same degree, and that observed variability in effect size is due only to sampling variability.

The results of the analysis presented in the last section, which tested

Table 2
Split-half reliabilities of eye movement effects, computed based on correlations between simple effects in each half of the data, and based on correlation parameter in mixed-effects models, as described in the text. Model-based estimate is posterior mean and 95% HDI; bold indicates HDI that does not include 0.

Experiment 1	Frequency		Predictability		Visual Contrast	
	Non-model-based	Model-based	Non-model-based	Model-based	Non-model-based	Model-based
First Fixation	.22	Raw: .27[−.71, .93] Log: .07[−86, .90]	.37	Raw: .33[−.72, .95] Log: .17[−.82, .92]	.22	Raw: .54[−.02, .94] Log: .44[−.22, .92]
Gaze Duration	.36	**Raw: .68 [.06, .97]** Log: .39[−.68, .95]	.31	Raw: .40[−.66, .96] Log: .28[−.78, .95]	.36	**Raw: .64 [.15, .96]** **Log: .72 [.34, .96]**
Skipping	.02	.18[−.76, .91]	.37	.26[−.78, .94]	.47	**.84 [.51, .99]**
Regressions	.01	.07[−.83, .89]	.09	.05[−.85, .89]	.40	.59[−.31, .97]
Experiment 2	Frequency		Predictability		Font Difficulty	
	Non-model-based	Model-based	Non-model-based	Model-based	Non-model-based	Model-based
First Fixation	.23	Raw: .48[−.45, .95] Log: .51[−.53, .96]	.20	Raw: .32[−.70, .94] Log: .13[−.83, .90]	.11	Raw: .18[−.72, .88] Log: .16[−.80, .91]
Gaze Duration	.35	**Raw: .75 [.32, .98]** Log: .44[−.53, .95]	.14	Raw: .44[−.42, .95] Log: .19[−.79, .92]	.18	Raw: .47[−.33, .95] Log: .52[−.39,.96]
Skipping	.18	.31[−.71, .94]	.14	0[−.87, .87]	.39	**.83 [.50, .99]**
Regressions	.05	0[−.88, .87]	−.01	−.10[−.90, .83]	.03	0[−.88, .88]

the reliability of individual differences in each experimental effect, should be related to the results from a comparison of models with and without by-subject slopes. If there are consistent differences between subjects across the two halves of an experiment in how strongly they show an effect, then a model that explicitly takes account of individual differences in effect size should fit the data better than one that does not. However, if the reliability analysis does not conclude that there are consistent differences between subjects across the two halves of an experiment, modeling such differences in the data as a whole might not improve model fit. Thus, we might expect, for example, that adding by-subject slopes for the frequency effect to mixed-effects models of (raw) gaze duration should result in an improvement in fit, as this effect showed fairly reliable individual differences in the model-based analysis. However, by-subject slopes for the predictability effect on the same measure may not result in much improvement in model fit, as individual differences in the predictability effect did not demonstrate very impressive reliability.

For each measure, we constructed a base model, using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015), with the following structure:

```
lmer(measure ~ freq + pred + stim + (1|subject))
```

This model includes fixed effects of the three variables, with the levels of each variable coded as .5 and -.5, and subject intercepts. For the skipping and regressions measures, a logistic regression model was constructed. As in the model-based reliability analyses, we constructed models of both raw and log-transformed fixation duration measures. We omitted fixed effect interaction terms on the grounds that in our original analysis, there were no frequency-by–predictability interactions, and the only compelling interactions involving the contrast or font manipulations were for gaze duration in Experiment 2. This base model was then compared to three different larger models, each of which included by-subject slopes for one of the three variables, e.g.

```
lmer(measure ~ freq + pred + stim + (1|subject) + (0 +
freq|subject))
```

Each of these six-parameter models was then compared to the five-parameter base model by means of the *anova* function for model comparison.[2]

In Table 3, we present the estimated standard deviation of the by-subject slopes in the larger models, and the *p*-value for the likelihood ratio test (LRT) comparing each of the larger models with corresponding base model. We do not use these *p*-values to make a binary decision as to whether by-subject slopes are justified, but rather to quantify the strength of the evidence against the smaller model; the LRT *p*-value denotes the probability, under the null hypothesis of no true subject variation, of an improvement in likelihood as large as the observed improvement with the larger model. Looking across rows, it is evident that the standard deviations of the slopes and the corresponding LRT *p*-values are related: The larger the estimate of the standard deviation of subject slopes, the smaller the LRT *p*-value. For example, in the analysis of raw first fixation duration for Experiment 1 (the top row), the estimate of the standard deviation of the subject slopes is smallest for frequency and largest for visual contrast, and the model comparison *p*-value is largest for frequency and smallest for visual contrast. Predictability falls in the middle for both measures.

The general conclusions that emerge from this analysis are as follows. For the effect of predictability, while the model comparison favors the inclusion of by-subject slopes for some measures in Experiment 1, the comparisons are quite equivocal in Experiment 2. Importantly, there is no measure for which a model that includes by-subject slopes for predictability is clearly favored in both experiments. Turning to frequency, there is only one measure, raw gaze duration, in which the larger model is clearly favored in both experiments. This is precisely the measure that demonstrated a reliable frequency effect, in both experiments, in the model-based reliability analysis. Just as the reliability of the frequency effect was unimpressive with log gaze duration as the dependent measure, it is also the case that the LRT does not unequivocally favor the model with by-subject slopes for frequency when log gaze duration is the dependent measure, especially in Experiment 2.

The effects of visual contrast (Experiment 1) on gaze duration and skipping, and the effect of font difficulty on skipping (Experiment 2) demonstrated impressive model-based reliability. We also see that for these measures, the model that includes by-subject slopes is clearly favored over the base model. In addition, by-subject slopes for visual contrast are clearly favored in the model of first fixation duration, and by-subject slopes for font difficulty are clearly favored in the model of gaze duration. These results, too, are broadly consistent with the reliability analysis; though the model-based reliability estimates for these effects had HDIs that spanned 0, the mean of the posterior was relatively high, in the neighborhood of .5.

In sum, the present analysis and the model-based analysis of reliability provide converging evidence for a set of conclusions about the stability, or lack thereof, of individual differences in the size of the

[2] For comparability between the analyses presented in this section and the preceding sections, we omit item-level random effects. Including by-item intercepts in both the base model and each larger model does not result in notable changes to the qualitative patterns discussed in this section.

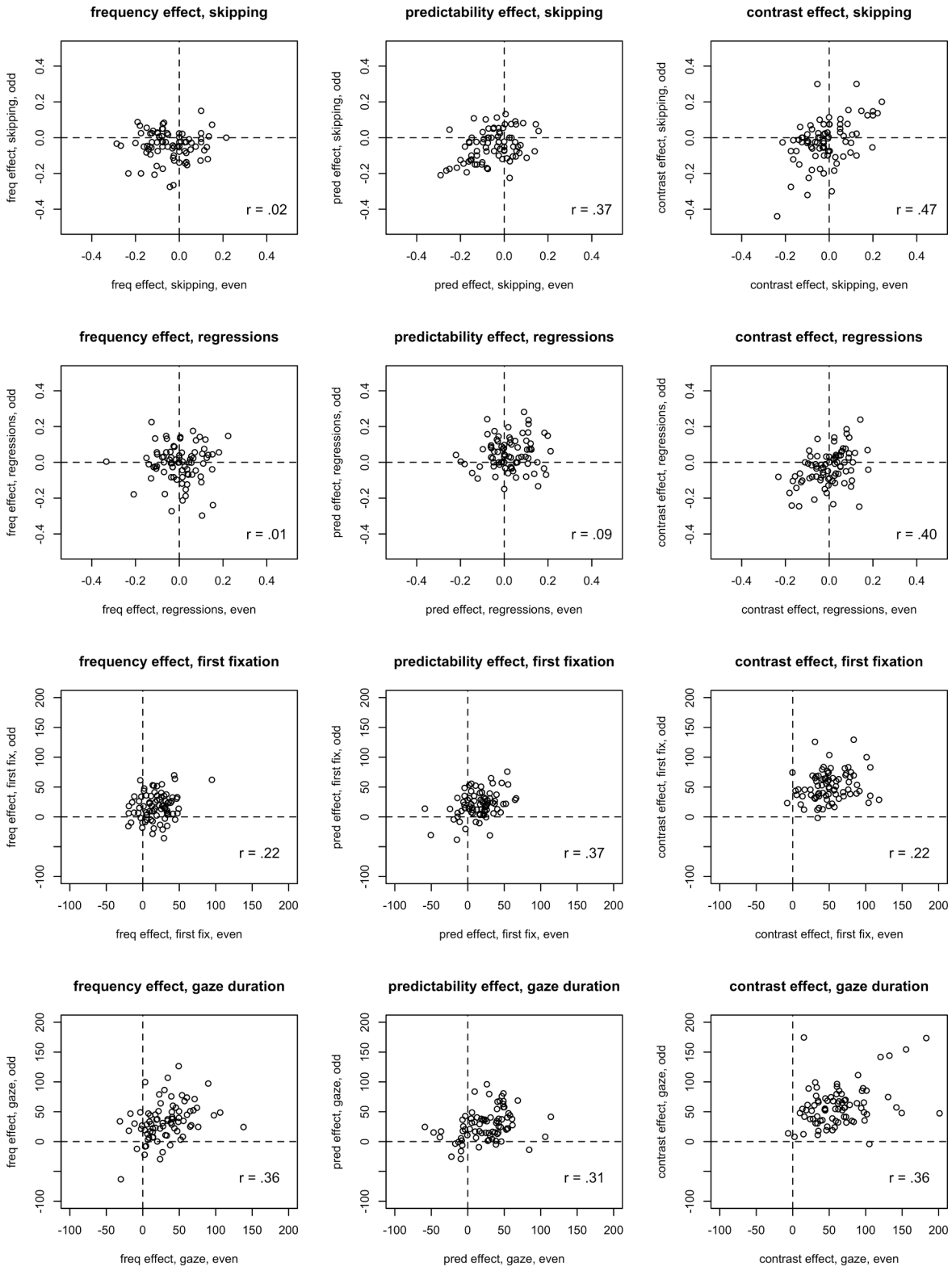

Fig. 7. Relationships in Experiment 1 between subjects' effects of experimental manipulations in even and odd items, based on non-model-based analysis.

various effects. Individual differences in the effects of predictability showed unimpressive reliability, and by-subject slopes for predictability do not consistently improve model fit, across the two experiments. Individual differences in the effect of frequency on raw gaze duration are quite reliable, and adding by-subject slopes for frequency to models of raw gaze duration does clearly improve model fit. However, effects of frequency on other measures were less reliable, and adding by-subject slopes resulted in only equivocal improvement in model fit. Both visual contrast and font difficulty demonstrated reliable effects on a number of measures, and by-subject slopes for these effects are clearly warranted.

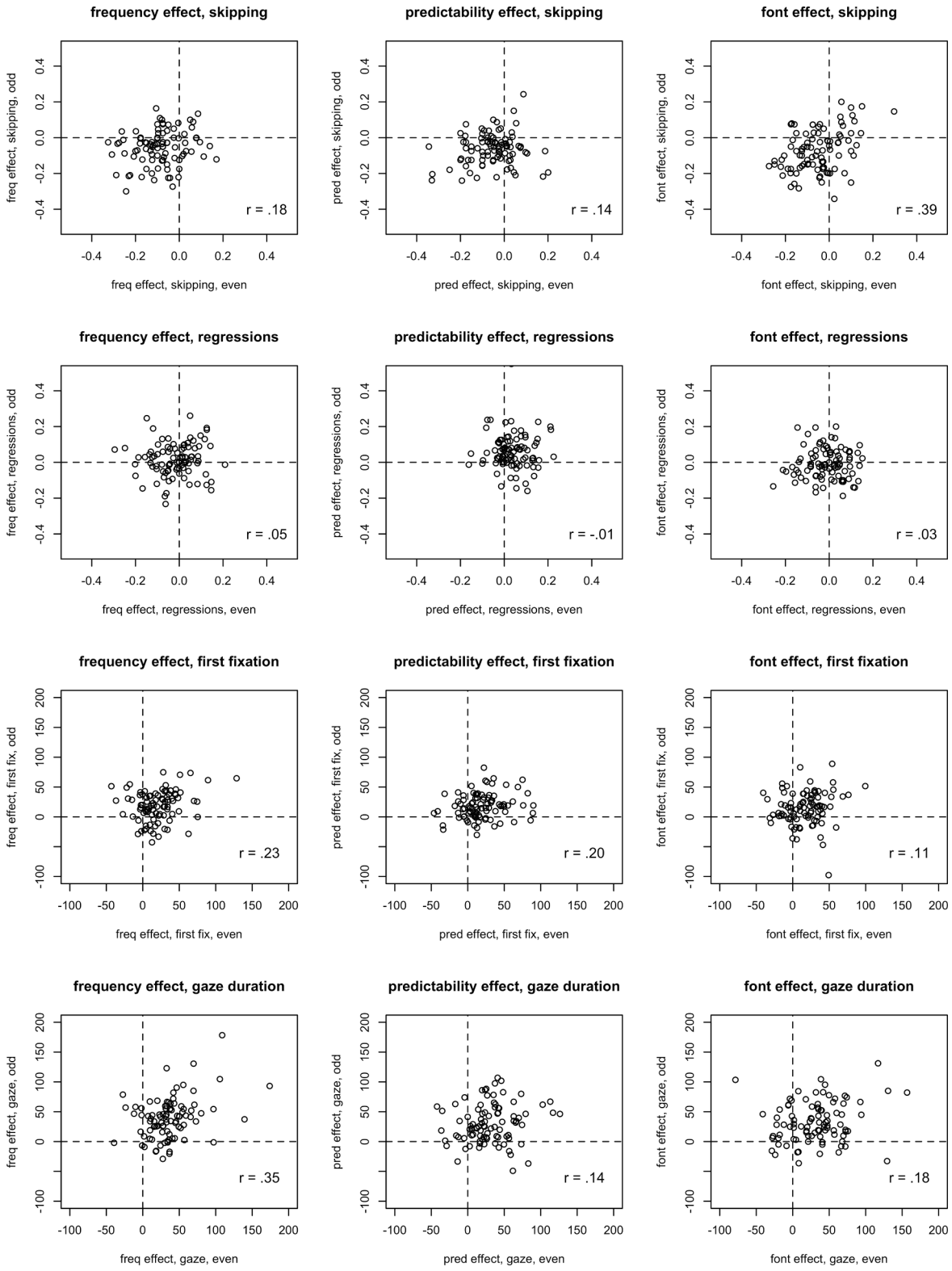

Fig. 8. Relationships in Experiment 2 between subjects' effects of experimental manipulations in even and odd items, based on non-model-based analysis.

General discussion

The present study confirms a number of previous studies (e.g., Carter & Luke, 2018; Dirix et al., 2019) in finding that individual differences in basic eye movement measures such as mean first fixation duration and skipping proportion are highly reliable. However, the present study finds quite poor split-half reliability of the effects of word frequency, predictability, visual contrast, and font difficulty, as computed by the traditional method of correlating simple effects in each half of the data. Computing split-half reliability by modeling the correlations between effects in each half, within the context of a mixed-effects model, also resulted in low reliability estimates for many effects, but did uncover substantially higher estimates in several cases: Effects of visual contrast and font difficulty on several eye movement measures, and the effect of

Table 3
Standard deviation of by-subject slopes for mixed-effects models, and LRT p-values for comparison of models with and without by-subject slopes, as described in the text.

Experiment 1	Frequency		Predictability		Visual Contrast	
	Slope sd	Model-comparison *p*	Slope sd	Model-comparison *p*	Slope sd	Model-comparison *p*
First Fixation	Raw: 6.46	Raw: .225	Raw: 9.60	Raw: .012	Raw: 14.76	Raw: <.001
	Log: .018	Log: .516	Log: .032	Log: .055	Log: .052	Log: <.001
Gaze Duration	Raw: 15.09	Raw: .001	Raw: 11.00	Raw: .059	Raw: 24.83	Raw: <.001
	Log: .039	Log: .029	Log: .04	Log: .021	Log: .072	Log: <.001
Skipping	.129	.505	.233	.074	.482	<.001
Regressions	.218	.279	.016	.995	.346	.036
Experiment 2	Frequency		Predictability		Font Difficulty	
	Slope sd	Model-comparison *p*	Slope sd	Model-comparison *p*	Slope sd	Model-comparison *p*
First Fixation	Raw: 11.57	Raw: .006	Raw: 9.03	Raw: .072	Raw: 9.91	Raw: .034
	Log: .046	Log: .001	Log: .028	Log: .18	Log: .033	Log: .064
Gaze Duration	Raw: 16.4	Raw: .001	Raw: 9.56	Raw: .229	Raw: 17.04	Raw: <.001
	Log: .033	Log: .172	Log: .035	Log: .124	Log: .055	Log: <.001
Skipping	.214	.052	.049	.908	.403	<.001
Regressions	.119	.741	0	1	0	.999

word frequency on raw gaze duration in both experiments. Likelihood ratio tests comparing mixed-effects models with and without a given set of by-subject slopes largely confirmed the model-based reliability analysis: Including by-subject slopes for visual contrast and font difficulty improved model fit for several measures, and among the predictability and frequency effects, the only by-subject slopes that were clearly justified in both experiments were slopes for the frequency effect on raw gaze duration.

Thus, we come to a quite different conclusion from the one previous study (Carter & Luke, 2018) that has investigated the reliability of individual differences in frequency and predictability effects on eye movements. While that study found that these variables, as well as word length, had moderately to highly reliable effects on a wide range of eye movement measures, we find that predictability effects are not very reliable, regardless of the measure in question, and the frequency effect is reliable only for raw gaze duration. This difference may be due to the fact that in the present study, the frequency and predictability variables were uncorrelated with each other and with other variables that are known to influence eye movements, enabling truly independent tests of the reliability of each effect. Subjects in the Carter and Luke (2018) study read natural texts where frequency, predictability, and word length are correlated with each other, as well as with other variables such as part of speech and position in a sentence.

We first note that our fairly high model-based estimates of the reliability of the visual contrast effects – as high as .84, for the effect of contrast on word skipping – may be seen as a successful 'sanity check'. The visual contrast manipulation should have different effects on different readers, due to genuine individual differences in contrast sensitivity (e.g., Legge, Rubin, & Luebker, 1987; Rubin & Legge, 1989). Had we been unable to identify reliable effects of visual contrast, we would have been skeptical about the suitability of our methods for assessing reliability at all. The fact that reliable effects of contrast were obtained in the model-based analysis, but not in the non-model-based analysis, reinforces the argument for model-based analysis of reliability (Rouder & Haaf, 2019).

It is less clear whether we should have expected reliable effects of font difficulty, predictability, or frequency. We focus our discussion on the latter two effects, which have been central to models of eye movement control (e.g., Reichle, Rayner, & Pollatsek, 2003) and to much theorizing about lexical processing in reading (e.g., Norris, 2006). We begin by discussing the effect of predictability. It has been argued that there are genuine differences between readers in how 'predictive' they are, and that these differences are related to reading skill or reading experience. On the one hand, some authors have suggested that the use of prediction in language processing increases with reading experience and skill (e.g., Huettig & Pickering, 2019). On the other hand, it has also been argued that less fluent readers rely strongly on prediction as opposed to bottom-up word identification (Stanovich, 1980). But the low reliability of individual differences in predictability effects on eye movements suggests that whatever the direction of the relationship, inter-reader differences in the use of prediction are very modest compared to the variability in the predictability effect from trial to trial. One reader may strongly predict the 'predictable' target word in one sentence, while only weakly predicting the 'predictable' target word in another sentence; these prediction strengths may be reversed for a different reader. Eye movement behavior when a reader encounters any specific target word may reflect the interaction of a specific reader's language experience with the context established by the item in question, more than global differences between readers in how strongly they predict upcoming words.

The low reliability estimates emerging from the present study imply that it should actually be quite difficult to detect any relationship between the size of a reader's predictability effect and other individual difference variables. (Below, we discuss in more detail the relationship between the reliability of an effect and the observable correlation between that effect and other measures.) Several studies have assessed whether predictability effects on eye movements are modulated by either reading skill (Ashby et al., 2005; Slattery & Yates, 2018) or age (Choi, Lowder, Ferreira, Swaab, & Henderson, 2017; Rayner et al., 2006). Ashby et al. (2005, Experiment 2) compared groups of skilled and average readers, and did not find significant group-by-predictability interactions. Slattery and Yates (2018) used a reading skill measure as a continuous predictor, and did find that less proficient readers showed a larger effect of predictability on gaze duration. However, no interaction was found in the skipping measure, and no interactions were observed with a spelling ability measure. Rayner et al. (2006) failed to find significant differences in predictability effects between older and younger readers. Choi et al. (2017), on the other hand, did report two just-significant ($t = 1.98$ and $t = 2.04$) interactions between predictability and age group, with older adults showing a larger effect of predictability on gaze duration and regression path duration, a measure that includes regressive re-reading. Interactions were not observed in other measures, including two other reading time measures and word skipping. Taken as a whole, it is questionable whether this literature supports any conclusion regarding the relationship between either reading skill or age and the predictability effect.

By contrast, the present results indicate that there are more reliable differences between readers in the effect of word frequency on their eye movements. However, it is only the gaze duration measure – and indeed, only raw gaze duration – that demonstrates such reliability. Why should

the effect of word frequency be more reliable than the effect of predictability? And why should it be in raw gaze duration, as opposed to other eye movement measures, that individual readers show reliable effects of frequency?

Evidence from a range of experimental paradigms (e.g., Chateau & Jared, 2000), including eye movement studies (Gordon et al., 2019; Kuperman & Van Dyke, 2011; Taylor & Perfetti, 2016), indicates that the word frequency effect diminishes as readers gain skill and experience. Explanations for this relationship vary; see Kuperman and Van Dyke (2013) for review of evidence and discussion of theoretical issues. One possibility is that skilled, experienced readers have developed fully specified orthographic representations even for low-frequency words, diminishing the functional difference between low- and high-frequency words. What is most relevant for present purposes is that this relationship is much better established than any relationship between reading skill or experience and predictability effects. If individual differences in the frequency effect on eye movements in reading were not at least moderately reliable, relationships between reading skill or experience and the size of a reader's frequency effect should have been difficult or impossible to detect.

Interestingly, there is evidence that it is the gaze duration measure that most clearly demonstrates a relationship between reading experience and the size of the frequency effect. Gordon et al. (2019) reanalyzed data from 546 subjects in eye movement experiments that were initially conducted to investigate a range of psycholinguistic questions. All subjects completed an Author Recognition Test (ART) and a Rapid Automatized Naming task (RAN). Gordon et al. assessed how effects of both word frequency and word length on eye movements may be modulated by subjects' scores on these two measures. They found significant interactions between the frequency effect on gaze duration and a subject's ART score, with subjects who recognized more authors (indicating more extensive reading experience) showing a smaller frequency effect. Notably, this interaction was not evident in any of their other reported measures, despite the very large size of the study, and despite the fact that the group-level frequency effect was evident in most of these measures. For example, both word frequency and subjects' ART scores had substantial effects on word skipping, but an interaction between the two was not in evidence. Based on the results of the present study, this discrepancy is expected, as it may be the case that only in the gaze duration measure are there reasonably reliable individual differences in the frequency effect.

Why gaze duration, as opposed to other measures, and why raw as opposed to log gaze duration? In both Experiments 1 and 2, raw gaze duration was the only measure in which the model-based reliability of the frequency effect was clearly on one side of zero, and where the mean of the posterior distribution for the estimate was in the range that is usually considered 'good': .68 in Experiment 1 and .75 in Experiment 2. Raw gaze duration was also the only measure in which by-subject slopes for the frequency effect clearly improved model fit in both experiments. By-subject slopes for first fixation duration improved model fit (based on model comparison p-value $< .05$) in Experiment 2, but not Experiment 1, and by-subject slopes for log gaze duration improved model fit in Experiment 1, but not Experiment 2.

In the E-Z Reader Model of eye movements in reading (Reichle et al., 2003), gaze duration is more consistently sensitive to word frequency than is any other measure. Regressions are not posited to be sensitive to frequency at all (see Abbott & Staub, 2015, for discussion), which is consistent with the lack of main effect of frequency on the regression measure in either experiment. Word skipping is sensitive to frequency because on a minority of trials, an early stage of word recognition known as L1, the duration of which is sensitive to frequency, may complete while the eyes are still on word n-1, and a planned saccade into word n may then be cancelled, resulting in a skip of word n. However, on most trials L1 does not complete rapidly enough for this to happen, regardless of word n's frequency, so word n receives a direct fixation. The duration of the first fixation on word n is then sensitive to word frequency on the majority of remaining trials, as this duration is a function of the duration of L1. However, on some proportion of trials the reader rapidly terminates the first fixation in order to refixate the word in a more ideal location, and in this case only gaze duration, not first fixation, will be sensitive to frequency. Thus, gaze duration is sensitive to frequency on every trial on which word n is fixated rather than skipped, while first fixation duration is sensitive to frequency on most of these trials, but not all. Based on this logic, it is not surprising that if one measure were to show reliable individual differences in the frequency effect, it would be gaze duration.

Why is substantially higher reliability observed when raw, as opposed to log, gaze duration is used as the dependent measure? We can think of two reasons why this might be the case. First, previous studies have shown that a substantial portion of the frequency effect resides in the right tail of fixation duration distributions (e.g., Staub et al., 2010). By eliminating the right skew of fixation duration distributions - which is a virtue, when it comes to satisfying the parametric assumptions of linear mixed-effects models - the log transformation may be reducing our ability to detect subtle differences between readers in the size of their frequency effect, precisely because that effect resides partially in the tail.

The second reason that individual differences in the frequency effect may be more reliable for raw gaze duration than log gaze duration is that this reliability may derive from what is essentially a measurement-related artifact. Several previous studies have suggested that the frequency effect, and other RT effects, may be larger for subjects who are slower responders overall (Faust, Balota, Spieler, & Ferraro, 1999; Schilling et al., 1998). Faust et al. point out that this relationship may be understood as arising from general processing rate differences. When an experimental manipulation modulates the quantity of processing work that is required, a subject who is a slower processor will demonstrate a larger experimental effect on RT. If the effect of word frequency is constant across readers in terms of the amount of extra cognitive work that is required in order to recognize a low-frequency word, a participant who is the slower processor will show a larger effect on RT.

In our Bayesian mixed-effects modeling we did see evidence in both experiments that the frequency effect on raw gaze duration was larger for slower readers. The mean of the posterior for the correlation between the subject intercept and the by-subject slope for word frequency in each half was between .66 and .87; readers who were slower overall showed larger frequency effects in each half of the data. Thus, it is possible that part of the reliability of the frequency effect derives simply from its relationship to overall reading speed.

On this account of the reliability of the frequency effect on raw gaze duration, it is expected that the log transform would reduce this reliability, as the log transform would suppress the relationship between the frequency effect and overall reading speed. The log transform tends to remove superadditive interactions (e.g., Lo & Andrews, 2015); a large raw RT difference between two relatively long times may be equivalent, in log space, to a smaller raw RT difference between two relatively short times. For example, the 50 ms difference between fixation durations of 300 and 350 ms is almost exactly the same, after log transformation, as the 35 ms difference between fixation durations of 215 and 250 ms. Thus, the log transform of gaze duration will tend to 'undo' a general relationship between effect size and overall reading speed. As expected on this account, in our Bayesian mixed-effects models of log gaze duration, subject intercepts were less strongly related to the by-subject slopes for frequency in each half, with the mean of the posterior of the correlation parameters ranging between .37 and .53 in the two experiments.

In sum, we cannot rule out the possibility that the primary reason that the frequency effect on raw gaze duration demonstrates relatively good reliability is simply that slower readers show a larger effect. The log transformation would suppress this relationship, diminishing the apparent reliability of the frequency effect itself.

We turn now to a broader methodological point, and a practical

suggestion. As noted in our introduction, the quantitative relationship between measurement reliability and the magnitude of an observable correlation is well known (Nunnally, 1970; Spearman, 1904). Given a true correlation $r_{AB.true}$ between variables A and B, and known test–retest reliabilities for the measurements of the variables relA and relB, the observable correlation between the variables is given by:

$$r_{AB.obs} = r_{AB.true} \times \sqrt{relA \times relB}$$

The implications of this equation are striking. If two variables have true correlation $r = .5$, and both are measured with test–retest reliability of .8, the observable correlation is .4. In order to obtain statistical power of .8 to detect a correlation of .4, we would need 46 subjects. But if one of the two variables is measured with reliability of .3, the observable correlation is now .245. Now, we would need 128 subjects for power of .8. The poor reliability of one of the measures has dramatic consequences for statistical power; see Hedge et al. (2018) for further discussion.

We are also able to estimate $r_{AB.true}$, if $r_{AB.obs}$ is known, along with the two reliabilities:

$$r_{AB.true} = \frac{r_{AB.obs}}{\sqrt{relA \times relB}}$$

If we have an observed correlation of .5, and reliabilities of .8 for both measures, the true correlation between the variables is estimated to be .625. But an observed correlation of .5, if one measure has reliability of .8 and the other has reliability of .3, corresponds to an estimated true correlation of just over 1; this is to say that such an observed correlation does not plausibly correspond to a true correlation. Vul and colleagues (Vul & Pashler, 2012; Vul, Harris, Winkielman, & Pashler, 2009) have referred to observed correlations that should not be mathematically possible, given the reliabilities of the correlated measures, as 'voodoo' correlations. In effect, knowing the reliabilities of the two measures establishes an upper limit on the correlations that we should expect to see in the data; this upper limit will be quite low when the reliability of one or both measures is low. If we see correlations that are higher than our measures' reliabilities should allow, we should suspect Type I error. Vul et al. (2009) point out that such 'voodoo' correlations are likely to arise in individual difference studies when researchers assess multiple correlations and do not appropriately correct for these multiple comparisons.

Together with the present empirical results, these mathematical relationships imply that eye movement researchers should be concerned about the power of individual difference studies to detect relationships between effects of variables such as word frequency or predictability and other variables such as reading skill or age, and about the possibility that when such relationships are observed, they are actually Type I errors. When carrying out such studies, a reasonable first step is to assess the reliability of individual differences in the experimental effects. Given that most researchers now use mixed-effects models to analyze eye movement data, an easy-to-implement check is to assess whether adding by-subject slopes for the effect in question does clearly improve model fit, relative to a model without these slopes. Clearly, this method does not definitively establish the reliability of an effect; for some of the effects that we have investigated here, this method favored a model with by-subject slopes, in one experiment or the other, even when the effect in question did not appear to be very reliable based on our model-based reliability analysis. However, if this model comparison *fails* to justify the assumption that observed differences in the size of an effect are attributable to genuine differences between subjects – if a model with by-subject slopes does not clearly fit the data better than a model that assumes that each subject demonstrates the same effect - it is not warranted to then ask whether these differences are modulated by some other factor. If we do not have clear statistical evidence that subjects do indeed differ in the extent to which they are affected by some variable, it does not make sense to explore how differences in the size of the effect are modulated by reading skill, working memory capacity, or age.

A final methodological point is that when explicit reliability calculations are carried out, the model-based estimates suggested by Rouder and Haaf (2019) are to be preferred over correlations between simple effects, as in Hedge et al. (2018). The model-based estimates of reliability that we obtained were largely consistent with theoretical expectations – e.g., visual contrast has highly reliable effects on a range of eye movement measures – while the simple effect correlations were low across the board. Moreover, the model-based estimates were largely consistent with the results of likelihood ratio tests establishing that inclusion of by-subject slopes do improve the fit of certain models. As noted by Rouder et al. (2019), model-based estimates of reliability, or of correlations between two effects, will often be highly uncertain in the absence of truly enormous quantities of data. Nevertheless, the estimates for certain effects in the present experiments were at least precise enough to ascertain that the reliability was on one side of zero.

Conclusions

The main substantive conclusion of the present study is that while individual differences in some effects of experimental variables on eye movements in reading are fairly reliable - effects of visual contrast, and the effect of word frequency on raw gaze duration - individual differences in other effects are much less reliable than might be hoped, from the perspective of individual difference studies. In particular, individual differences in the effect of a word's predictability are not reliable, and frequency effects on other measures are also not very reliable.

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. https://doi.org/10.1016/j.jml.2015.07.002.

Ashby, J., Rayner, K., & Clifton, C. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology Section A, 58*(6), 1065–1086. https://doi.org/10.1080/02724980443000476.

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

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

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

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

Calvo, M. G. (2001). Working memory and inferences: Evidence from eye fixations during reading. *Memory, 9*, 365–381. https://doi.org/10.1080/09658210143000083.

Carter, B. T., & Luke, S. G. (2018). Individuals' eye movements in reading are highly consistent across time and trial. *Journal of Experimental Psychology: Human Perception and Performance, 44*, 482–492. https://doi.org/10.1037/xhp0000471.

Chace, K. H., Rayner, K., & Well, A. D. (2005). Eye movements and phonological parafoveal preview: Effects of reading skill. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 59*, 209–217. https://doi.org/10.1037/h0087476.

Chateau, D., & Jared, D. (2000). Exposure to print and word recognition processes. *Memory & Cognition, 28*, 143–153. https://doi.org/10.3758/bf03211582.

Choi, W., Lowder, M. W., Ferreira, F., Swaab, T. Y., & Henderson, J. M. (2017). Effects of word predictability and preview lexicality on eye movements during reading: A comparison between young and older adults. *Psychology and Aging, 32*, 232–242. https://doi.org/10.1037/pag0000160.

Dirix, N., Brysbaert, M., & Duyck, W. (2019). How well do word recognition measures correlate? Effects of language context and repeated presentations. *Behavior Research Methods, 51*, 2800–2816. https://doi.org/10.3758/s13428-018-1158-9.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American, 236*, 119–127. https://doi.org/10.1038/scientificamerican0577-119.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*, 641–655. https://doi.org/10.1016/s0022-5371(81)90220-6.

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences, 116*, 5472–5477. https://doi.org/10.1073/pnas.1818430116.

Falkauskas, K., & Kuperman, V. (2015). When experience meets language statistics: Individual variability in processing English compound words. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 41, 1607–1627. https://doi.org/10.1037/xlm0000132.
Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125*, 777–799. https://doi.org/10.1037/0033-2909.125.6.777.
Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14*, 178–210. https://doi.org/10.1016/0010-0285(82)90008-1.
Gordon, P. C., Moore, M., Choi, W., Hoedemaker, R. S., & Lowder, M. W. (2019). Individual differences in reading: Separable effects of reading experience and processing skill. *Memory & Cognition*, 1–13. https://doi.org/10.3758/s13421-019-00989-3.
Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1.
Henderson, J. M., & Luke, S. G. (2014). Stable individual differences in saccadic eye movements during reading, pseudoreading, scene viewing, and scene search. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 1390–1400. https://doi.org/10.1037/a0036330.
Huettig, F., & Pickering, M. J. (2019). Literacy advantages beyond reading: Prediction of spoken language. *Trends in Cognitive Sciences, 23*, 464–475. https://doi.org/10.1016/j.tics.2019.03.008.
Kennison, S. M., & Clifton, C. (1995). Determinants of parafoveal preview benefit in high and low working memory capacity readers: Implications for eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 68–81. https://doi.org/10.1037/0278-7393.21.1.68.
Kliegl, R., Masson, M. E., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition, 18*, 655–681. https://doi.org/10.1080/13506280902986058.
Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology, 1*, 238. https://doi.org/10.3389/fpsyg.2010.00238.
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. https://doi.org/10.1037/xlm0000128.
Kuperman, V., & Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language, 65*, 42–73. https://doi.org/10.1016/j.jml.2011.03.002.
Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 802–823. https://doi.org/10.1037/a0030859.
Legge, G. E., Rubin, G. S., & Luebker, A. (1987). Psychophysics of reading—V. The role of contrast in normal vision. *Vision Research, 27*, 1165–1177. https://doi.org/10.1016/0042-6989(87)90028-9.
Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis, 100*, 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008.
Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6*, 1171. https://doi.org/10.3389/fpsyg.2015.01171.
Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology, 88*, 22–60. https://doi.org/10.1016/j.cogpsych.2016.06.002.
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. https://doi.org/10.1016/j.jml.2017.01.001.
Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: Modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic Bulletin & Review, 20*, 819–858. https://doi.org/10.3758/s13423-013-0404-5.
Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*, 353–383. https://doi.org/10.1016/0010-0285(77)90012-3.
Norris, D. (2006). The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*, 327–357. https://doi.org/10.1037/0033-295x.113.2.327.
Nunnally, J. C., Jr (1970). *Introduction to psychological measurement*. New York: McGraw-Hill.
Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences, 108*, 3526–3529. https://doi.org/10.1073/pnas.1012551108.
Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*, 3–25. https://doi.org/10.1080/00335558008248231.
Rayner, K., & Duffy, S. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*, 191–201. https://doi.org/10.3758/bf03197692.
Rayner, K., Li, X., Williams, C. C., Cave, K. R., & Well, A. D. (2007). Eye movements during information processing tasks: Individual differences and cultural effects. *Vision Research, 47*, 2714–2726. https://doi.org/10.1016/j.visres.2007.05.007.
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. https://doi.org/10.1037/0882-7974.21.3.448.
Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 1290–1301.
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.
Reingold, E. M., & Rayner, K. (2006). Examining the word identification stages hypothesized by the EZ Reader model. *Psychological Science, 17*, 742–746.
Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science, 1*, 19–26. https://doi.org/10.1177/2515245917745058.
Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review, 26*, 452–467.
Rouder, J., Kumar, A., & Haaf, J. M. (2019, March 25). Why most studies of individual differences with inhibition tasks are bound to fail. https://doi.org/10.31234/osf.io/3cjr5.
Rubin, G. S., & Legge, G. E. (1989). Psychophysics of reading. VI—The role of contrast in low vision. *Vision Research, 29*, 79–91. https://doi.org/10.1016/0042-6989(89)90175-2.
Schilling, H. E., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition, 26*, 1270–1281. https://doi.org/10.3758/bf03201199.
Schmidtke, D., Van Dyke, J. A., & Kuperman, V. (2018). Individual variability in the semantic processing of English compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*, 421–439. https://doi.org/10.1037/xlm0000442.
Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Attention, Perception, & Psychophysics, 74*, 5–35. https://doi.org/10.3758/s13414-011-0219-2.
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. https://doi.org/10.3758/s13414-013-0442-0.
Slattery, T. J., & Yates, M. (2018). Word skipping: Effects of word length, predictability, spelling and reading skill. *The Quarterly Journal of Experimental Psychology, 71*, 250–259. https://doi.org/10.1080/17470218.2017.1310264.
Smith, N., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In Proceedings of the annual meeting of the cognitive science society (Vol. 33, No. 33).
Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology, 15*, 72–101. https://doi.org/10.2307/1412159.
Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*, 32–71. https://doi.org/10.2307/747348.
Staub, A. (2011). The effect of lexical predictability on distributions of eye fixation durations. *Psychonomic Bulletin & Review, 18*, 371–376. https://doi.org/10.3758/s13423-010-0046-9.
Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language & Linguistics Compass, 9*, 311–327. https://doi.org/10.1111/lnc3.12151.
Staub, A. (2020). Do effects of visual contrast and font difficulty on readers' eye movements interact with effects of word frequency or predictability? *Journal of Experimental Psychology: Human Perception and Performance, 46*, 1235–1251. https://doi.org/10.1037/xhp0000853.
Staub, A., & Benatar, A. (2013). Individual differences in fixation duration distributions in reading. *Psychonomic Bulletin & Review, 20*, 1304–1311. https://doi.org/10.3758/s13423-013-0444-x.
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. https://doi.org/10.1037/a0016896.
Taylor, J. N., & Perfetti, C. A. (2016). Eye movements reveal readers' lexical quality and reading experience. *Reading and Writing, 29*, 1069–1103. https://doi.org/10.1007/s11145-015-9616-6.
Traxler, M. J. (2007). Working memory contributions to relative clause attachment processing: A hierarchical linear modeling analysis. *Memory & Cognition, 35*, 1107–1121. https://doi.org/10.3758/bf03193482.
Traxler, M. J., Long, D. L., Tooley, K. M., Johns, C. L., Zirnstein, M., & Jonathan, E. (2012). Individual differences in eye-movements during reading: Working memory and speed-of-processing effects. *Journal of Eye Movement Research, 5*(1).
Veldre, A., & Andrews, S. (2014). Lexical quality and eye movements: Individual differences in the perceptual span of skilled adult readers. *The Quarterly Journal of Experimental Psychology, 67*, 703–727. https://doi.org/10.1080/17470218.2013.826258.
Veldre, A., & Andrews, S. (2015a). Parafoveal lexical activation depends on skilled reading proficiency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 586–595. https://doi.org/10.1037/xlm0000039.
Veldre, A., & Andrews, S. (2015b). Parafoveal preview benefit is modulated by the precision of skilled readers' lexical representations. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 219–232. https://doi.org/10.1037/xhp0000017.
Veldre, A., & Andrews, S. (2016). Semantic preview benefit in English: Individual differences in the extraction and use of parafoveal semantic information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 837–854. https://doi.org/10.1037/xlm0000212.
Vul, E., & Pashler, H. (2012). Voodoo and circularity errors. *Neuroimage, 62*, 945–948. https://doi.org/10.1016/j.neuroimage.2012.01.027.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*, 274–290. https://doi.org/10.1111/j.1745-6924.2009.01125.x.

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. https://doi.org/10.1037/a0025338.

Whitford, V., & Titone, D. (2012). Second-language experience modulates first-and second-language word frequency effects: Evidence from eye movement measures of natural paragraph reading. *Psychonomic Bulletin & Review, 19*, 73–80. https://doi.org/10.3758/s13423-011-0179-5.