Excerpt from Response to Reviews of Initial Version of
Kliegl, Risse, & Laubrock (2007, JEP:HPP)

OVERVIEW OF GENERAL CHANGES

There are several major changes from the original submission in response to the reviews.

First, we re-analyzed all raw data to be as compatible with Rayner et al. as possible with a stricter selection criterion on delay times for the display change. Specifically, in the previous version, we ignored the starting and landing point of the saccade that triggered the display change. In the new version we included trials only if the saccade triggering the display change started on word n and also crossed the boundary in a forward direction during this saccade. (As before the display change still had to be completed during the saccade triggering the display change.) This eliminated a few cases where, e.g., triggering occurred but the eye drifted back to word n afterwards. We compensated the loss in trials by giving up the binocular constraint that trials were included only if fixations were assigned to the same word in both eyes. Now we use all right-eye data--like all other labs. Consequently, the means and number of observations are slightly different.

Second, we switched from F1/ F2-ANOVAs to linear mixed-effects models for statistical inference; the justification is in the first response to the Editor. We also added power analyses.

Third, we deleted statistics relating to initial landing positions to make room for results from the additional skipping analysis requested by all reviewers (also documented in a new Table 2).

These changes removed some ambiguities present in the original submission; they did not change the main pattern of results.

RESPONSE TO EDITOR AND REVIEWERS

EDITOR

As a small extra point, when I was reading the paper I was not clear what you were making of some of the cases where effects are reliable across items but not participants - perhaps you could clarify this?

Answer:

This is a big extra point! We reported F1 and F2 statistics for reasons of compatibility with Rayner et al. Perhaps even more than you, we were bothered by the differences between F1 and F2 statistics in our data. Therefore, we had also carried out linear mixed-effects analyses (lme). Indeed, the observed mean differences always mapped best on the lme statistics.

Given your comment, we decided to deal with this problem in what we consider to be the most rational way. The problem arises because F1/ F2-ANOVAs are simply not the best analyses for such data. First, for unbalanced designs, hardly avoidable in eye tracking, there is dramatic loss of statistical power in F1/F2-ANOVAs (see references below). Such loss can affect F1 and F2 statistics in different ways. Second, there is a long history of exchanges on when (a) only F1, (b) F1 and F2, or (c) minF should be used. For counterbalanced designs like the present experiment, the recommendation actually is to use only F1 (see Raaijmakers, 1999, JMemLang).

There is new research by Quené and van den Bergh (2004), Pinheiro and Bates (2000), and Baayen (2004, in press) showing that linear mixed-effects models fare much better than F1/F2 -ANOVAs, especially when the usual assumptions are violated and when the design is unbalanced. Under some conditions F1/F2-ANOVAs are even prone to produce spurious significance. Again, lme appears to be less susceptible to such problems.

Most importantly, the lmer program allows us to specify subjects and items as crossed random effects, meaning that one can control between-subjects and between-items effects in the same analysis. Consequently, we decided to part with tradition and go with the lme procedure, but attach a table with F1 and F2 statistics to this Reply so you and the reviewers can see that we are not sweeping anything under the rug (see Appendix).

What are the differences? Given Raaijmakers (1999), we do not trust F2-statistics unless they are also significant in F1 or lme. If we take only the F1-ANOVA as criterion, we miss the effect of lexical status of word n+1 on first-fixation and gaze durations on word n+1. Words n+1 were obviously very homogeneous: All of them are 3-letters long and in the function-word category of lexical status they were repeated. This may be one reason why the effect shows up only in the F2 and lme analyses. Moreover, skipping probability is a very reliable individual-difference variable and manifests itself in skipping of short words. This may have contributed strongly to between-subject variance for word n+1, reducing the chance to see the effect in the F1-ANOVA.

2. The reported effects are small (see first reviewer's comments). Are you convinced that you have the power to test for interactions given these small effect sizes? Can you provide any more formal analysis of power?

Answer:

We know from prior research that given enough power we can resolve effect sizes of about 7 ms for single fixations. We added a paragraph in a new Analysis section (p. 7) detailing the power analysis for first-fixation durations. Basically, we fixed main effect sizes to 7 ms and the interaction to 14 ms and used the variance estimates from the statistical analysis to simulate first-fixation durations of our experiment for word n, n+1, and n+2 (i.e., 30 subjects, 160 items). After randomly deleting the appropriate proportion of missing data, the power of our tests each based on 1000 simulations were .86 for word n and word n+2 and between .58 and .63 for word n+1 which was skipped more frequently. Power for word n+1 is not great, but we do not want to argue the null hypothesis anyway.

We have less experience with gaze durations. For the same effect sizes and with estimated gaze-duration variances (which are, of course, larger than those from first fixations) we obtained powers of .51/.52 for word n and word n+1 and powers of .63/.64 for word n+1. Note that there were not many refixations for 3-letter words n+1, so these statistics are not very different from first-fixation durations (also the variance estimates were smaller for gaze-durations on word n+1). Thus, for gaze durations on word n and n+2 we need larger effects for comparable power. As effects are typically larger for gaze than for first-fixation durations, this is not really a problem.

In the reviews, the power issue was raised in the context of the pattern of means looking like an interaction. This interaction is now reliable for gaze durations (b=-13) on word n+1 but there is no evidence of this kind of interaction for first-fixation durations. This is in good agreement with the means. Note, however, that the significant interaction is actually opposite in direction to our expectation. Therefore, we refrain from its interpretation (see p. 11).

We also plugged in Rayner et al.'s design parameters (i.e., 36 subjects, 40 items) and observed much lower statistical power. However, we do not think this is of more than exploratory value. There are too many differences between the experiments (e.g., they have fewer missing data because of a faster eye tracker) to turn this into a meaningful comparison.

In summary, we are confident that our design has sufficient power to recover main effects larger than 7-ms and interaction effects larger than 14 ms in first-fixation durations.

REVIEWER 1:

Instances of N+1 skipping, which should be relatively common for three-letter words, are of considerable theoretical interest. As noted in the Discussion, preview benefit for N+2 can now be expected according to the E-Z Reader model (and -in my view- according to the SWIFT model). Yet, such preview benefits do not emerge. Regrettably, this intriguing and important finding is mentioned in passing in a single sentence in the Discussion section, and there are no data to back up this claim. I suggest that the results of conditional N+2 analyses(N+1 skipping) be reported.

Answer: Done (top of p. 8, Table 2)

REVIEWER 2

First, with respect to word n+2, both studies find no effect of the nonword preview when word n+2 is fixated. Interestingly, they note (page 11) that a supplementary analysis revealed no effect when word n+1 was skipped. My inclination is that this supplementary analysis should be described in more detail. A footnote in Rayner et al. describes a similar analysis though they obtained values that look like there was some preview benefit from n+2 when n+1 was skipped (they simply didn't have enough data for a formal analysis; see also McDonald, in press - see point 6 below).

Answer: Done (top of p. 8, Table 2). We also refer to McDonald and footnote 3 of Rayner et al. on p. 13.

REVIEWER 3

… On page 4, KRL write "Aside from the language, our experiment [?and that of RJB] differed in two ways". The two ways relate to the length of word N+1 and its lexical status. In fact, KRL's manipulation differs in one other way. The RJB experiment defined a pre-target region as comprising two words, W1 and W2. A contingent boundary was placed either after W1 or after W2. The experiment was subsequently analysed in terms of inspection on the "pre-target zone" as a whole (i.e. two words) as a function of whether the boundary was placed after W1 or W2. My point is that analysing the pre-processing of word N+2 in this way may appear straightforward, but is actually incredibly complex. When the boundary was placed after W1, crossing it restored the target for the duration of any fixation on W2. It follows the measure of First Fixation duration would generally be acquired while the target was scrambled, but the aggregate measure of gaze would be derived from a situation where part of the time the target was scrambled and part of the time it was not. To assume that the restoration of the target had no interesting effects on the processing of W2 was, to my mind, misguided. For this reason I believe the conclusions arrived at by RJB (for example, regarding the apparent absence of parafoveal-on-foveal effects) are unwarranted. The manipulation in KRL (the present paper) is certainly better motivated, placing the boundary after W1, carefully controlling properties of W2, and measuring inspection time on both. I make this comment in the hope that this difference between the two experiments can be high-lighted. It is quite clear the authors are aware of all these points, but they are not made very explicit.

Answer: We added the difference between experiments in the introduction (p. 4). We also refer to it for discussion of the preview benefit on word n+1 (top of p. 11)

Table 1. Comparison of p-value statistic of F1-, F2-, and lme-analyses for each word n, word n+1, and word n+2 seperately.

DV	IV	word n			word n+1			word n+2		
		F1	F2	lme	F1	F2	lme	F1	F2	lme
FFD	lexical status of word n+1	**.009****	**.005****	**.009****	.116	<.001***	.006**	.354	.051 (*)	.327
	preview of word n+2	.197	.048*	.219	**.032***	**.001****	**<.001*****	.213	.694	.490
	lexstat x prevw	.426	.939	.424	.563	.621	.276	.881	.869	.772
GZD	lexical status of word n+1	<.001***	<.001***	<.001***	.178	.002**	.045*	.356	.542	.549
	preview of word n+2	.013*	.022*	.038*	.017*	<.001***	<.001***	.398	.506	.347
	lexstat x prevw	.038*	.226	.042*	.851	.822	.522	.295	.676	.430
SKP	lexical status of word n+1	.122	.102	.101	<.001***	<.001***	<.001***	.703	.910	.924
	preview of word n+2	.161	.768	.501	.416	.174	.100	.704	.939	.851
	lexstat x prevw	.906	.833	.839	.827	.414	.759	.836	.456	.894

FFD = first-fixation duration including single fixations   GZD = gaze duration   SKP = skipping probability

Table 2. Comparison of p-value statistic of F1- and lme-analyses for word n+2 conditional on skipping of word n+1.

	FFD		GZD	
	F1	lme	F1	lme
n+1 skipping (n1skip)	.012*	<.001***	<.001***	<.001***
lexical status of word n+1 (lexstat)	.070	.121	.014*	.033*
preview condition of word n+2 (prevw)	.402	.441	.439	.227
n1skip x lexstat	.033*	.026*	.210	.868
n1skip x prevw	.883	.298	.869	.557
lexstat x prevw	.379	.624	.095	.319
n1skip x lexstat x prevw	.135	.497	.241	.731

FFD = first-pass fixation including single fixations GZD = gaze duration