Saccade launch site as a predictor of fixation durations in reading: Comments on Hand, Miellet, O'Donnell, and Sereno (2010)

Timothy J. Slattery
University of South Alabama
Adrian Staub
University of Massachusetts, Amherst
and
Keith Rayner
University of California, San Diego

Correspondence to:
Timothy J. Slattery
Department of Psychology
University of South Alabama
5871 USA Dr. North LSCB 320
Mobile, AL 36688-0002
slattery@usouthal.edu
Running Head: Frequency and Predictability Effects

Abstract

An important question in research on eye movements in reading is whether word frequency and word predictability have additive or interactive effects on fixation durations. A fair number of studies have reported only additive effects of the frequency and predictability of a target word on reading times on that word, failing to show significant interactions. Recently, however, Hand, Miellet, O'Donnell, and Sereno (2010) reported interactive effects in a study that included the distance of the prior fixation from the target word (launch site). They reported that when the saccade into the target word was launched from very near to the word (within 3 characters), the predictability effect was larger for low frequency words, but when the saccade was launched from a medium distance (4-6 characters from the word) the predictability effect was larger for high frequency words. Hand et al. argued for the importance of including launch site in analyses of target word fixation durations. Here we describe several problems with Hand et al.'s use of ANOVAs in which launch site is divided into distinct ordinal levels. We describe a more appropriate way to analyze such data—linear mixed-effect models (LMMs) – and we use this method to show that launch site does not modulate the interaction between frequency and predictability in two other data sets.

An important issue in the study of reading is the extent to which word frequency and word predictability interact. While a number of lexical decision or naming studies (Stanovich & West, 1979, 1981, 1983) suggested that the two variables interact, many eye movement studies find that they yield additive effects on fixation times (Altarriba, Kroll, Scholl, & Rayner, 1996; Ashby, Rayner, & Clifton, 2005; Lavigne, Vitu, & d'Ydewalle, 2000; Miellet, Sparrow, & Sereno, 2007; Rayner, Ashby, Pollatsek, & Reichle, 2004; Rayner, Binder, Ashby, & Pollatsek, 2001). Thus, though these studies document main effects of frequency and predictability, with shorter fixation times on high than on low frequency words and on predictable than on less predictable words, evidence for an interaction between the two variables has not been found. The most critical study is by Rayner et al. (2004) since it more directly examined the possibility of a frequency by predictability interaction than the other studies.

More recently, Hand, Miellet, O'Donnell, and Sereno (2011) reported that when they examined fixation time on target words, as with prior research, they found additive effects of the two variables. However, they also reported a variant on the traditional analyses of target word fixations in which fixations on the target word were split into 3 groups based on the distance of the prior fixation from the target word (*launch site*). They reasoned that the closer the prior fixation to the target, the more preview of it the reader would have prior to fixating it. They defined 3 distinct ordinal levels of launch site: near (1-3 characters), medium (4-6 characters), and far (7-9 characters). When they included this 3-level ordinal variable in ANOVAs, a 3-way interaction of frequency, predictability, and launch site emerged. Specifically, they found when the saccade into the target word was launched from near to the word (within 3 characters), the predictability effect was significantly larger for low frequency words; when it was launched from a medium distance (4-6 characters) the predictability effect was numerically larger for high frequency words; and when it was launched from even farther away (7-9 characters) no interaction was evident. If reliable, this result has 3 important theoretical implications. First, Rayner et al.'s data were the impetus for changing the underlying parameter values reflecting the operation of frequency and predictability in the E-Z Reader model (Reichle, Fisher, Pollatsek, & Rayner, 1998; Pollatsek, Reichle, & Rayner, 2006) from an interactive function to an additive one. If there really is an interactive pattern,

it would have implications for E-Z Reader. Second, a frequency by predictability interaction would perhaps be expected within parallel lexical processing models like SWIFT (Engbert, Nuthmann, Richter, & Kliegl, 2005) and Glenmore (Reilly & Radach, 2006) since lexical processing is distributed over a wider range of words than in E-Z Reader. Finally, an interaction of frequency and predictability would have implications for the modularity (Fodor, 1983) of the word recognition system.

Here we raise several concerns regarding Hand et al.'s statistical analysis, which cast serious doubt on their findings. The issues fall into two basic categories: the handling of analysis cells with no data points (or with widely varying numbers of data points) and the treatment of continuous launch site data as though it were an ordinal predictor with only a few distinct levels. We thus performed a launch site analysis of data from another experiment in which frequency and predictability were factorially manipulated (Gollan, Slattery, Goldenberg, van Assche, Duyck, & Rayner, 2011), using the ANOVA technique employed by Hand et al. We show, using this data set, how strongly the results are influenced by employing Hand et al.'s technique for estimating values in cells with no data. We also show how very unbalanced the design is in general, and discuss some undesirable consequences. We then discuss the use of a linear mixed-effects model (LMM) to analyze these data, which avoids these problems, and allows launch site to be treated as a continuous predictor variable. We illustrate the use of such a model with a subset of the Gollan et al. data, as well as with data from an experiment (Staub, 2011) in which predictability was also factorially manipulated, while frequency varied continuously. In neither data set was there evidence of a three-way interaction between frequency, predictability, and launch site.

ANOVA-based launch site analysis of Gollan et al. (2011)

In Gollan et al., 72 University of California, San Diego students participated; half were monolingual English readers and half were English-Spanish bilinguals. Given that both groups showed main effects of frequency and predictability, we combined them for the purpose of this analysis. There were 90 target words (45 low and 45 high frequency words), and for each target, a high and a low predictable sentence frame was constructed. The high and low frequency targets were matched in length (mean 5.4 characters) and averaged

79.8 and 9.3 occurrences per million respectively in the COCA corpus (Davies, 2008). The cloze values for the high and low frequency targets in the high predictable sentences were 87.5% and 85.5% respectively; in the low predictable sentences they were 3% and 2% respectively. Subjects only read 60 of these targets (15 from each condition); the remaining 30 were used in a lexical decision task. The items were counterbalanced across the experimental conditions so that subjects saw an equal number of targets in each condition and didn't see the same target twice. Eye movements were monitored with an Eyelink 1000 tracker and subjects read silently for comprehension; they pressed a button on a keypad when they finished reading (see Gollan et al., 2011 for full methodological details, and details about data trimming, track losses, and unusable trials).

For the purpose of exploring the influence of launch site, we used only gaze duration data[1]. In order to assess the frequency and predictability effects in a true orthogonal design, for comparison to the launch site analysis that follows, gaze durations were first analyzed via 2 (frequency: high vs. low) X 2 (predictability: high vs. low) ANOVAs with subjects as a random effect variable (F1) and items as a random effects variable (F2). There was a main effect of frequency, $F1\ (1,71) = 55.52$, $MSE = 570.88$, $p < .001$; $F2\ (1,88) = 17.99$, $MSE = 1092.81$, $p < .001$, as gaze durations were shorter on high than low frequency words (see Table 1). There was also a main effect of predictability, $F1\ (1,71) = 22.56$, $MSE = 832.79$, $p < .001$; $F2\ (1,88) = 21.47$, $MSE = 679.00$, $p < .001$, as gaze durations were shorter on high than low predictable words. However, there was no hint of an interaction between frequency and predictability, $Fs < 1$. These results are therefore similar to those reported by Hand et al. in their initial analysis and those reported by Rayner et al. (2004) and the other studies cited above.

Next we analyzed the data based on launch site in the same manner as Hand et al. Splitting the data into near, medium, and far launch sites results in fewer observations per ANOVA analysis cell than when no such split is undertaken. In fact, it isn't uncommon for there to be no observations for a given subject (or item) in some analysis cells. When this occurred, we followed Hand et al.'s procedure and estimated values for the cells with no observations[2] with the following formula: estimate = condition mean + (subject mean – grand mean). In this way, the estimated values represent both the average influence of the experimental manipulation as well as

the relative reading speed of the individual subject whose data are being estimated. These data were then submitted to 2 separate 3 (launch site) X 2 (frequency) X 2 (predictability) ANOVAs (means and standard errors are presented in Table 1). This analysis yielded a significant main effect of launch site, $F1(2,142) = 54.55$, $MSE = 2816.97$, $p < .001$; $F2(2,176) = 40.73$, $MSE = 1875.47$, $p < .001$. All Bonferroni adjusted pairwise comparisons were significant ($ps < .01$) as gaze durations increased from the near to the far launch site. There were also main effects of frequency, $F1(1,71) = 36.98$, $MSE = 2041.36$, $p < .001$; $F2(1,88) = 11.91$, $MSE = 4626.85$, $p < .005$, and predictability, $F1(1,71) = 17.63$, $MSE = 3069.67$, $p < .001$; $F2(1,88) = 14.91$, $MSE = 2685.08$, $p < .001$. There was also a marginal frequency by predictability interaction in the subjects analysis, $F1(1,71) = 3.67$, $MSE = 1865.94$, $p = .06$; $F2 < 1$, as well as a hint of a launch site by predictability interaction in the items analysis, $F1(2,142) = 1.52$, $MSE = 2998.21$, $p = .222$. However, the three-way interaction with launch site that had been significant in the Hand et al. analysis was not significant in these data, $F1(2,142) = 2.02$, $MSE = 1947.19$, $p = .136$; $F2 < 1$.

Insert Table 1 and Table 2 here

Using the same procedure as Hand et al. we were unable to replicate the pattern of effects that they reported. With every failure to replicate, questions about power should be raised. To draw inferences from this failure to replicate, it is important to show that this analysis has at least as much power as the Hand et al. analysis. To determine whether we had similar power for such a launch site contingent analysis, we calculated the number of trials in each of the launch site groups (see Table 2). The data set had a similar number of observations across the three launch sites of interest (near, medium, and far), and in fact had slightly more data in the critical near region. Thus, we could not replicate the critical interactions reported by Hand et al., despite having more subjects. However, what is even more notable is that when splitting the data in this way we now find a marginally significant interaction between frequency and predictability in the subjects analysis that hadn't approached significance ($Fs < 1$) in the analysis that didn't split the data by launch site.

When assessing how the inclusion of launch site may have generated a spurious frequency by predictability interaction it is important to realize that the degrees of freedom for the ANOVA reported above

are incorrect. This is because the missing values that were estimated do not count as degrees of freedom. Therefore, it is essential to subtract 1 degree of freedom for every value we estimate. This was not done in the Hand et al. analysis and, since they did not report the number of values that they estimated, it is impossible to assess the actual degrees of freedom. In the present gaze duration data we had to estimate values for 4.6% of analysis cells (40 empty cells). Thus, the degrees of freedom in the denominator of the F1 statistic for the interaction between frequency and predictability should have been 31, not 71. If we recalculate the appropriate MSE given these adjusted degrees of freedom, the F value for this interaction drops from 3.67 to 1.60, which with 1 and 31 degrees of freedom is clearly not significant. It is at least somewhat comforting to realize that once the correct degrees of freedom are used, the spurious interactive trend is no longer marginally significant.

There are also major problems with the manner with which values for missing cells were estimated. The procedure used by Hand et al., and repeated in our demonstration, is not appropriate. There are a number of issues with estimating missing data (for a thorough discussion see Allison, 2002; Little & Rubin, 1987; Schafer, 1997). One is that there is less variability with estimated data than with real data which reduces the standard errors in estimates. This may have played a part in the numerous sphericity violations we encountered with the launch site contingent analyses as estimated values were more likely to occur in the near and far than in the medium launch site conditions. Additionally, the estimation technique used is a variant of marginal mean estimation which produces biased estimates of variances and covariances (Haitovsky, 1968). Better estimates of missing data can be obtained using maximum likelihood or multiple imputation procedures in many situations.

However, failing to use the correct degrees of freedom and estimation of missing values aren't the only problems with how the launch site contingent analysis was conducted. Specifically, the procedure of splitting by launch site and estimating missing values affected the pattern of means. The final row of Table 1 contains the means for each of the four cells of the original design, recalculated from the cell means in the analysis that includes launch site. Examining the first and last rows of Table 1 shows that this recalculation procedure didn't change any one condition mean by more than a few ms, but the interaction effect nearly doubled, going from 6

ms in the initial analysis to 11 ms in the launch site analysis. The reason for this distortion lies in the fact that ANOVA gives each cell of the data matrix equal weight. That is, an estimated cell value counts as much as one consisting of a mean of 5 values or one consisting of a mean of 15 values. Intuitively, our confidence in the value of a cell as an estimate of a corresponding population mean should increase with the number of data points that are averaged together to yield this value. Note that in a truly balanced orthogonal design, where every analysis cell consists of an average of the same number of data points, no such distortion would take place.

To illustrate the effect of this phenomenon on the present data, consider Tables 3 and 4, which present the number of trials averaged together for each analysis cell for 6 subjects. Consider subject 2 in these tables. This subject had 9 valid gaze durations on high frequency, highly predictable words in the study. In a traditional 2X2 analysis, each of these observations would count equally toward the mean of the high frequency, high predictability condition. However, for this condition in the launch site analysis, this subject only had a single observation in each of the near and medium launch site cells, with 6 trials averaged together in the far launch site cell (1 trial had an even farther launch site, which was not included in the Hand et al. analysis). So in the launch site contingent analysis, when calculating the high frequency high predictability condition mean, the two observations that occur in the near and medium launch site cells are each weighted 6 times as much as the observations that occur in the far launch site cell, and 3 times as much as they had been weighted in the non launch site contingent 2X2 analysis. Subject 2 can be contrasted with subject 3, who had 13 observations in the high frequency and high predictability condition of the 2X2 analysis, but 9, 4, and 0 in the near, medium, and far launch site cells, respectively. Clearly, near launch sites were more representative of subject 3's data than of subject 2's. However, subject 2's single observation in the near launch site, high frequency and predictability condition is weighted 9 times as much as each of subject 3's observations when computing the mean for this condition. In sum, splitting each cell of the 2x2 design into 3 cells, based on a factor such as launch site, results in some experimental observations having a wildly disproportionate influence on condition means. Of course it

is not possible to say in advance which observations these will be; it will depend, for each subject, on how the observations in each of the 4 cells of the initial design happen to be distributed into the 3 sub-cells.

Insert Tables 3 and 4 here

There are additional issues when missing data are not evenly distributed across conditions, as different conditions will be more or less influenced by the estimation procedures. This will be the case for eye movement data in which launch site is included as a factor. The reason is that the probability of skipping the target word is much greater when the previous fixation (launch site) is closer to the target (see Table 5). Over 65% of skips come from near launch sites in the Gollan et al. (2011) data, as well as in the Staub (2011) data discussed below. Consider also Figures 1a-1c, which show the distribution of forward saccade lengths from the near, medium, and far launch sites in the Gollan et al. data set. The solid lines represent all forward saccades regardless of where these saccades landed. The dashed lines represent the lengths of only those saccades that ended in a fixation on the target. As seen, the distribution of all saccade lengths does not appear to be different across the 3 launch sites, with approximate means and ranges of 7 and 20 characters respectively. However, looking at only those saccades that land on the target, we see a shift from shorter saccades in the near launch site positions to much longer saccades in the far launch site positions, for obvious reasons. These figures illustrate that skipping (the area between the 2 lines on the right side of the distributions) is far more common with near launch sites. Additionally, with the far launch sites it is far more common to make another fixation prior to fixating or skipping the target word. When the target word is skipped during first pass reading, there is no single fixation duration or gaze duration for the trial, which constitutes a major source of missing data for such experiments.

Insert Table 5 here and Insert Figures 1a-1c here

This last observation raises some broader issues related to the treatment of missing data in experimental studies. There are essentially three classifications of missing data. Data can be considered to be *missing completely at random* (MCAR) in which case the likelihood of data point (X_i) being missing is entirely unrelated to the value of X_i (were it known) or to any of the experimental conditions. This may occur for

instance if the eye-tracking equipment randomly failed to function. Data can also be considered to be *missing at random* (MAR) in which case the likelihood of data being missing is related to one or more of the experimental variables, but after controlling for these variables the likelihood of data point (X_i) being missing is still unrelated to the value of X_i. For example, with eye-tracking this might be the case if blinks are more likely to occur in certain luminance conditions. However as long as it can be assumed, after controlling for the influence of luminance, that the likelihood of there being missing data due to blinks is unrelated to the fixation durations that would have resulted had these blinks not occurred, then the missing values can be considered MAR. The last category is *missing not at random* (MNAR). In this case, the value of the missing data point X_i (were it known) would be related to its being missing. With eye-tracking this seems to be the most likely situation when target word data are missing due to word skipping (i.e., the target word is never fixated). While an argument could be made that missing data due to target word skipping could be classified as MAR, target words are known to be skipped more often when their processing is easier (see Rayner, 1998, 2009). Therefore, the duration of fixations in these cases would be expected to be quite short had the word not been skipped, and this situation seems best categorized as MNAR. It should be noted that even the maximum likelihood and multiple imputation procedures mentioned earlier cannot in principle estimate missing values when those values cannot be assumed to be represented by the existing data. Such estimation would require incorporating additional assumptions concerning the nature of the missing values. However, such a treatment is beyond the scope of the current article. In either the MAR or MNAR situations, the simple least squares approximations used in ANOVA will cause problems with the covariance structure and will provide biased estimates.

In summary, analysis of Gollan et al.'s data, including launch site in the ANOVAs, as in Hand et al., did not reveal a launch site-contingent frequency x predictability interaction. More importantly, this exercise illustrated problems associated with this analysis. Critically, if cell means are to be estimated, the associated degrees of freedom must be corrected. Equally critically, the inherent variability in the number of observations in each cell will result in some observations counting much more than others, with unpredictable effects on the

overall analysis. This problem is likely to arise whenever observations are split based on a behavior that is not under the experimenter's control, and is therefore likely to be unequally distributed across experimental conditions[3]. It is also important to note that the issues described here will have influenced the results for all eye movement measures reported by Hand et al. in which they included launch site.

Launch site as a continuous predictor in a linear mixed-effects model

In this section we suggest that data sets with continuous predictors such as launch site should be analyzed using linear mixed-effects models (LMMs, see Baayen et al., 2008; Gelman & Hill, 2007; Luke, 2004; Quené & van den Bergh, 2004; Raudenbush & Bryk, 2002). This type of analysis solves the problems we have enumerated: It eliminates the need to estimate the value of missing data and gives each individual observation equal weight. Before describing this method in detail, however, we point out another, more conceptual problem associated with Hand et al.'s analysis, which is also solved by LMM. In the Hand et al. analysis, the inherently continuous values of launch site were divided into 3 distinct bins. The actual division is entirely arbitrary; Hand et al. offer no theoretical motivation for treating trials with launch sites 1-3 as alike, and as different from trials with launch sites 4-6. The results of the analysis would seem to imply that, with respect to the frequency by predictability interaction, trials with launch site 3 are functionally just like trials with launch site 1, but trials with launch site 4 are functionally distinct. However, considering the distribution of launch sites for the data in Figure 2, there appears to be no such functional distinction between the these regions. We assume, instead, that if launch site interacts with frequency and/or predictability, this must be a graded effect, rather than involving a sudden change at an arbitrary boundary between, for example, launch sites 3 and 4.

Insert Figure 2 here

As applied here, LMM models each individual observation rather than the means of the observations in a cell of the ANOVA design[4]. In the earlier ANOVA of the Gollan et al. data set, each subject contributed 12 data points to the analysis (1 for each condition mean), with extremely unequal numbers of observations contributing to these 12 points. However, in the LMM each subject contributes approximately 60 data points (1 for each valid

trial) with each point weighted equally. Additionally, this procedure does not need to estimate missing data but instead estimates the effects of predictor variables from the data it has and uses listwise deletion for trials without fixation data (skips). The random effects of subjects and items can be estimated at the same time, and continuous predictors can be used rather than creating arbitrary divisions over the variable. In the next section, we conduct an LMM analysis of the Gollan et al. gaze duration data introduced above. In the subsequent section, we illustrate how this method can be extended to data sets with multiple continuous predictors, by applying it to a data set with predictability as an experimental factor and frequency as a continuous covariate. To anticipate, in neither data set does this analysis reveal any hint of the interaction reported by Hand et al.

LMM Analysis of Gollan et al. (2011) Data

The gaze durations from Gollan et al. were first analyzed collapsing over launch sites. This analysis used the LME4 package of the R statistical software (Bates & Maechler, 2010; R Development Core Team, 2010). The analysis included frequency and predictability as centered fixed effects, as well as random intercepts for both subjects and items. We also included random slopes for subjects and items when model comparisons indicated that these additions improved the model's overall fit. For each LMM, we indicate which random slopes were included in the final model. However, the estimates of the fixed effects were highly stable over all of the compared models. We report coefficients and standard error estimates for the fixed effects as well as estimated *t*-values, with *t*s greater than 2 indicating statistical significance. The final LMM with frequency and predictability as fixed effects included random intercepts for subjects and items as well as a random slope for predictability over items. This analysis indicated that gaze durations were shorter for high than for low predictable targets (b = 17.28, SE = 3.71, t = 4.65), and for high than for low frequency targets (b = 20.86, SE = 5.04, t = 4.14). However, there was no indication that target predictability and frequency interacted (b = -3.94, SE = 7.42, t < 1).

Next we fit an LMM that also included launch site as a continuous variable centered about its mean. Again, this model included a random slope for predictability over items as well as random intercepts for both items and subjects. There was an effect of launch site on gaze durations, (b = 6.32, SE = 0.54, t = 11.75) as target

gaze durations increased with increasing launch site distance. Gaze durations were also shorter for high than for low frequency targets (b = 20.99, SE = 5.18, t = 4.05), and they were also shorter on high than on low predictable targets (b = 18.26, SE = 3.76, t = 4.85). Launch site did not interact with frequency (b = 0.55, SE = 1.01, t < 1), or predictability, (b = 1.45, SE = 1.02, t = 1.42). Importantly, there was still no indication of a frequency by predictability interaction (b = -3.04, SE = 7.53, t < 1), nor was there a three-way interaction between frequency, predictability and launch site (b = 0.80, SE = 2.03, t < 1). Therefore, with the Gollan et al. data set we obtained additive main effects of frequency, predictability, and launch site within an LMM analysis.

As we have discussed, target word skipping is an important eye movement measure and a major source of missing data. Additionally, there is an assumption among many researchers that increases in target word skipping indicate increasing ease of processing. Thus, we explored the influence that word frequency and predictability had on target word skipping. As already mentioned, skipping rates for target words were strongly influenced by launch site, with the majority of skips coming from launch sites close to the target word (see Table 5). We conducted an LMM analysis that predicted skipping with binomial logistic regression, due to the binomial nature of skipping data, and we report z rather than t values here. The model included the same fixed and random effects predictors as the gaze duration model above. As expected, the closer the launch site was to the target, the greater the likelihood that the target would be skipped, (b = -0.400, SE = 0.024, z = -16.88, p < .001). Additionally, the likelihood of skipping the target word was significantly lower when it was unpredictable (b = -0.452, SE = 0.161, z = 2.80, p < .01). Finally, there was an interaction between launch site and predictability (b = -0.091, SE = 0.045, z = -2.02, p < .05). In order to understand this interaction, it is important to understand how to interpret the coefficient for the main effect of predictability. Since we centered the launch site variable, the coefficient for predictability is the effect for this variable at the average launch site. The interaction of these variables indicates that the effect of predictability changes with launch site; the predictability effect on skipping gets smaller as the launch site gets closer to the target. No other effects approached significance, ps > 0.1.

It may be noted that we included launch position only as a linear predictor, while Hand et al. reported what appears to be a non-linear 3-way interaction, in which predictability had a stronger effect on fixation durations for low frequency words for near launch sites and a stronger effect for high frequency words for mid launch sites, with no apparent interaction for far launch sites. To assess whether our linear model of gaze duration missed some non-linear interaction of this sort, we constructed a plot of the residuals from the model as a function of launch site (see Figure 3). If there were a non-linear interaction, some systematic trend in the residuals would be expected, with generally positive residuals appearing in some location along the range of launch sites, and generally negative residuals appearing somewhere else. It is evident that this is not the case; the residuals appear to be centered around zero along the entire range of launch sites. In other words, the linear model provides a fit that appears to be about equally good for all values of launch site[5].

Insert Figure 3 here

LMM Analysis of Staub (2011) Data

Next we examined the generalizability of the results obtained with Gollan et al.'s data by examining the interaction of predictability, frequency, and launch site in a second data set (Staub, 2011). This also illustrates how the LMM framework can be applied with any combination of categorical and continuous predictors. Unlike Gollan et al., Staub did not factorially manipulate frequency; rather, it varied continuously across the items. In the study, 31 University of Massachusetts students participated. All were native speakers of English. There were 50 target words, with each being read twice by each subject, once in a high-predictability context (mean cloze value 60%) and once in a low-predictability context (mean cloze value 2%); each subject contributed a maximum of 100 observations. The materials were presented in a pseudo-randomized fashion and each target word was read once in each half of the experiment; half of the sentences in each half of the experiment had a high predictability target and half had a low predictability target. The words were generally high in frequency, with a mean frequency of occurrence, based on the SUBTLEX corpus (Brysbaert & New, 2009), of 75.8 per million, but there was substantial variation, ranging from quite low (2.8 per million) to very high (641 per million). As in

Gollan et al., eye movements were monitored with an Eyelink 1000 tracker. Subjects read for comprehension, and answered a 2-choice comprehension question, by means of a button press, on approximately 1/3rd of trials.

As above, we first present the results collapsing across launch site. The LMM included fixed effects for log word frequency (as a continuous predictor centered about its mean) and predictability, and random intercepts for subjects and items as well as a random slope for predictability over items. Again, gaze durations were shorter for high than for low predictable targets (b = 23.84 SE = 5.39, t = 4.42), and were shorter for higher frequency words (b = 10.62, SE = 4.88, t = 2.18)[6]. Unlike in the Gollan et al. data, there was a hint of a predictability by frequency interaction, but it did not reach significance, (b = -14.97, SE = 8.82, t = 1.70). Note that this interactive trend is due, surprisingly, to a numerically smaller predictability effect for low frequency words.

Next we added launch site, as a continuous variable centered about its mean, to the LMM. Again, this model included a random slope for predictability over items as well as random intercepts for both items and subjects. There was an effect of launch site on gaze durations (b = 6.17, SE = 0.57, t = 10.82), as they increased with increasing launch site distance. Gaze durations were also shorter for high than for low frequency targets (b = 13.18, SE = 5.00, t = 2.63), and they were shorter on high than on low predictable targets (b = 26.20, SE = 5.25, t = 4.99). Launch site did not interact with predictability (b = 0.42, SE = 1.08, t < 1), but did with frequency, (b = 2.24, SE = 0.91, t = 2.46) as the frequency effect increased with increasing launch distance. Also, there was a marginal 3-way interaction between frequency, predictability, and launch site (b = 3.41, SE = 1.83, t = 1.87). However, this marginal interaction is in the *opposite* direction of the one reported by Hand et al. (2010): Here, the effect of predictability on gaze durations on low frequency words increased, rather than decreased, with increasing launch distance. A plot of the residuals as a function of launch site appears in Figure 4. As with the Gollan data, these residuals appear to be centered around zero essentially throughout the range of launch sites.

Insert Figure 4 here

Finally, a logistic regression analysis of the probability of skipping the target word, as in the Gollan et al. data, revealed that the closer the launch site was to the target, the greater the likelihood that it would be

skipped, (b = -0.441, SE = 0.027, z = -16.20, p < .001); skipping was also less likely when the word was unpredictable (b = -0.662, SE = 0.153, z = -4.32, p < .001). The only other significant effect was an interaction between launch site and frequency (b = 0.121, SE = 0.044, z = 2.74, p < .01) wherein frequency had an increasing effect on the probability of a skip, as the launch site was farther from the word.

General Discussion

Across two comparable data sets to Hand et al. (2010), we found no evidence for an interaction of frequency and predictability when launch site of the saccade entering the target word was taken into account. Gollan et al.'s data revealed no frequency x predictability interaction overall and no 3-way interaction with launch site. Staub's data revealed a marginal 3-way interaction. However, it was due to a pattern opposite to that obtained by Hand et al.; rather than a pronounced predictability effect for low frequency words when the saccade launch site was near, the predictability effect for low frequency words was actually weakest for near launch sites. Thus, there was no reliable modulation of the predictability x frequency interaction by launch site; the data do, however, show main effects of frequency, predictability, and launch site. We suspect the conflicting results reported by Hand et al. occurred primarily because of inappropriate statistical analyses. The main problems involve the failure to adjust the degrees of freedom for the estimation of missing data, the undue influence wielded by certain data points when each cell in the 2 x 2 design is further subdivided into 3 cells (based on launch site), and the treatment of launch site as an ordinal variable with 3 discrete levels. As noted above, there is a strong relationship between launch site and the likelihood of skipping a target word, and this too complicates the analysis, by influencing where ANOVA cells with missing values, or very few observations, are most likely to appear. In Gollan et al.'s and Staub's data sets there were significant main effects of predictability with more skipping of high than on low predictable words, a common finding (see Rayner, 2009). Additionally, in Staub's data there was a frequency by launch site interaction effect on skipping. Given our assumptions about skipping and processing difficulty, this interaction may have played a role in the marginal effect of the 3-way interaction of launch site, predictability, and frequency on gaze durations in the same data.

Hand et al. also reported a word frequency effect on skipping, as well as a frequency by predictability interaction. The pattern of this interaction was the opposite of the pattern they reported for fixation duration measures in the near launch site condition—in which most of these skips would have been initiated. That is, they reported more skipping of high frequency words in the high than in the low predictable condition, with low frequency words showing no significant difference in skipping rates by predictability. The interpretation of this interaction would then be the opposite of the interpretation Hand et al. proposed for the interaction on fixation durations at the near launch sites. That is, if predictability eases processing more for low than for high frequency words, skipping should have increased more for low frequency words under the high predictability condition.

As these studies illustrate, word skipping represents a major source of missing data in eye movement studies, which can complicate statistical analysis of target word fixation duration measures. However, the fact that there was no fixation on a target word doesn't imply that it wasn't read; the general assumption is that the target word was processed (read) on the fixation immediately prior to or following the skipping saccade. Thus, it may be possible to obtain some type of fixation duration measure of processing even in the event that a word is skipped. Both EZ-Reader and SWIFT predict that when a word is skipped during first pass reading it is processed during the preceding and/or following fixations. For instance, EZ-Reader assumes that words can be skipped for two reasons. The majority of target word skipping will occur because, on the fixation prior to the skip, the word was processed to a great enough extent that it triggered the programming of a saccade to the next word. The other way in which EZ-Reader predicts word skipping is as the result of saccadic error. That is, saccades are not perfectly executed, and due to error the saccade may land beyond the target word. In this situation, the model predicts that attention is still located on the word that was skipped and it is processed during this post skip fixation. SWIFT also predicts that when a word is skipped it is processed during the preceding and following fixations. Additionally, SWIFT, like EZ-Reader, predicts that sometimes a word will be skipped accidentally due to saccadic error. However, there are important differences between these models where skipping and lexical processing are concerned. For instance, EZ-Reader predicts that the lexical processing will either take place

during the preceding fixation or take place during the following fixation but not both[7]. However, with SWIFT, processing of a skipped word will generally occur during a number of fixations both prior to and following skipping due to the distributed processing assumptions of the model. Additionally, in SWIFT the durations of the fixation prior to the skip will be uninfluenced by the frequency and/or predictability of the skipped word (except in cases of mislocated fixations that result from saccadic error). It may be possible to use such frameworks to devise a method for obtaining target word processing measures even in the event that the target word does not receive a direct fixation. However, further research will be required to determine the nature of such a measure.

Hand et al. (2010) are not the first researchers to analyze their data contingent on launch site. Indeed, a well-known study by Rayner (1975) used such a procedure. When that study was reported, the type of LMM statistical analyses discussed here was not widely used. Our claim is not that all studies reporting data contingent on launch site should be dismissed. In the case of the Rayner (1975) study, other studies using a gaze contingent naming procedure in which distance from the target word was experimentally manipulated yielded similar results (see Rayner, 1978; Rayner, McConkie, & Ehrlich, 1978; Rayner, McConkie, & Zola, 1980). Our main point is that investigators need to be very cautious in using analyses contingent on launch site and seek converging evidence if such analyses are used (hopefully, also using the procedures suggested in this article).

In summary, there are two important points made by the current study. First, regarding an interaction between frequency and predictability on fixation times, the majority of studies have failed to find an interaction (Altarriba et al., 1996; Ashby et al., 2005; Lavigne et al., 2000; Miellet et al., 2007; Rayner et al., 2001, 2004) and it is likely that the statistical issues with Hand et al.'s study resulted in Type 1 error. Even if their results for the near launch site were reliable, the issues with an opposite interaction for skipping rates raise serious questions as to how these interactions should be interpreted with regard to lexical processing. The second important point concerns the analysis of continuous variables like launch site in eye movement studies. Specifically, caution needs to be used when analyzing subsets of data that are split based on a dimension that is not under experimental control. Such analyses will routinely result in non-orthogonal data, and it is crucial that proper

statistical methods for the treatment of such data are followed to avoid issues with Type 1 error and biased estimates of population parameters. As we have shown, LMMs provide a statistical framework for including continuous predictors, and avoid the pitfalls that are present when analyzing such data with standard ANOVAs. However, even with the LMM framework, there are possible complications from missing data, due to word skipping, that need to be considered.

References

Allison, P.D. (2002). *Missing data*. Newbury Park, CA: Sage.

Altarriba, J., Kroll, J.F., Sholl, A., & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed language sentences: Evidence from eye fixation and naming times. *Memory & Cognition, 24*, 477-492.

Ashby, J., Rayner, K., & Clifton, C. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *Quarterly Journal of Experimental Psychology*, 58A, 1065-1086.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press, Cambridge, UK.

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.

Bates, D., Maechler, M., & Dai, B. (2009). Lme4: Linear mixed-effects models using S4 classes. R Package Version 0.999375–32.

Brysbert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977-990.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*, 193-210.

Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*, 777-813.

Gelman, A., Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press, Cambridge, UK.

Gollan, T.H., Slattery, T.J., Goldenberg, D., van Assche, E., Duyck, W. & Rayner, K. (2011). Frequency Drives Lexical Access in Reading but not in Speaking: The frequency lag hypothesis. *Journal of Experimental Psychology: General*, revision submitted.

Haitovsky, Y. (1968). Missing data in regression analysis.*Journal of the Royal Statistical Society,Series B,30*, 67-82.

Hand, C.J., Miellet, S., O'Donnell, P.J., & Sereno, S.C. (2011). The frequency-predictability interaction in reading:

It depends on where you're coming from. *Journal of Experimental Psychology: Human Perception and Performance*, in press.

Lavigne, F., Vitu, F., & d'Ydewalle, G. (2000). The influence of semantic context on initial landing sites in words. *Acta Psychologica, 104*, 191-214.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Luke, D.A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.

Miellet, S., Sparrow, L., & Sereno, S.C. (2007). Word frequency and predictability effects in reading French: An evaluation of the E-Z Reader model. *Psychonomic Bulletin & Review, 14*, 762-769.

Myers, J.L., Well, A.D., Lorch, R.F. Jr. (2010). *Research design and statistical analysis* (3rd edition). Routledge.

Pollatsek, A., Reichle, E.D., & Rayner, K. (2006). Tests of the E-Z Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology, 52*, 1-56.

Quené, H., & van den Bergh, H. (2007). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication, 43*,103-121.

R Development Core Team. (2010). R: A language and environment for statistical computing. Retrieved from http://www.R-project.org.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology*, 7, 65-81.

Rayner, K. (1978). Foveal and parafoveal cues in reading. In J. Requin (Ed), Attention and performance (Vol 7, pp 149-162). Hillsdale, NJ: Erlbaum.

Rayner, K., (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124, 372-422.

Rayner, K. (2009). The Thirty Fifth Sir Frederick Bartlett Lecture: Eye movements and attention during reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, 62, 1457-1506.

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

Rayner, K., Binder, K.S., Ashby, J., & Pollatsek, A. (2001). Eye movement control in reading: Word predictability has little influence on initial landing positions in words. *Vision Research, 41*, 943-954.

Rayner, K., McConkie, G. W., & Ehrlich, S. (1978). Eye movements and integrating information across fixations. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 529-544.

Rayner, K., McConkie, G. W., & Zola, D. (1980). Integrating information across eye movements. *Cognitive Psychology*, 12, 206-226.

Reichle, E.D., Pollatsek, A., Fisher, D.L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105*, 125-157.

Reilly, R., & Radach, R. (2006). Some empirical tests of an interactive activation model of eye movement control in reading. *Cognitive Systems Research, 7*, 34-55.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Stanovich, K.E., & West, R.F. (1979). Mechanisms of sentence context effects in reading: Automatic activation and conscious attention. *Memory & Cognition, 7*, 77-85.

Stanovich, K.E., & West, R.F. (1981). The effect of sentence context on ongoing word recognition: Tests of a two-process theory. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 658-672.

Stanovich, K.E., & West, R.F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General*, 112, 1-36.

Staub, A. (2011). The effect of lexical predictability on distributions of eye fixation durations. *Psychonomic Bulletin & Review*, *18*, 371-376.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition, 116,* 71-86.

Footnotes

1. The problems with this analysis that we will bring to light will actually influence analyses of single fixation durations even more than gaze durations, because there will be fewer instances of single fixation durations than gaze durations.

2. We thank Chris Hand for providing information regarding the treatment of missing data in the Hand et al. study.

3. In eye movement data, other examples would include analyses based on skipping vs. fixating, or regressing vs. moving forward.

4. See Gelman and Hill (2007) for a detailed introduction to mixed-effects modeling, and see, e.g., Demberg and Keller (2009) and Staub (2010) for some recent applications to eye movement data.

5. One other notable feature of this graph is the fact that there are clearly some outlying positive residuals, but there are not outlying negative residuals. This is simply due to the fact that the distribution of gaze durations includes some very high values that are not well fit by the model.

6. The parameter estimate for frequency reflects the predicted change in fixation duration corresponding to each change of 1 unit of log frequency.

7. Except in the extremely rare case of a mislocated target fixation that lands just to the left of the target word followed by a mislocated refixation that lands just to the right of the target word.

Acknowledgments

This research was supported by Grant HD 26765 from the National Institute of Child Health and Human Development. We would like to thank Caren Rotello and Arnold Well for their helpful comments concerning statistical issues addressed in this manuscript.

Table 1. Gaze durations from Gollan et al. (2011), with standard errors in parentheses

	HF-HP	HF-LP	LF-HP	LF-LP
Unsplit	221 (4.1)	240 (4.6)	244 (4.3)	257 (5.2)
Near	202 (5.6)	214 (6.0)	226 (7.3)	227 (5.8)
Medium	218 (5.4)	239 (6.5)	242 (5.8)	266 (7.6)
Far	241 (8.3)	273 (9.4)	267 (7.5)	272 (8.3)
Re-collapsed	221 (4.8)	242 (5.2)	245 (4.9)	255 (5.4)

Table 2. Number and percent of trials in each of the different launch site conditions

	Gollan et al. (2011)	Staub (2011)	Hand et al. (2010)
Near	1083 (25.1%)	643 (20.7%)	917 (16.3%)
Medium	1284 (29.7%)	823 (26.5%)	1356 (24.1%)
Far	746 (17.3%)	600 (19.4%)	1159 (20.6%)
Very far	344 (8.0%)	281 (9.1%)	804 (14.3%)
Skips	611 (14.1%)	530 (17.1%)	1054 (18.7%)
Excluded	252 (5.8%)	223 (7.1%)[1]	342 (6.1%)
Total	4320 (100%)	3100 (100%)	5632 (100%)

[1] Included in this cell are 54 trials in which there was a fixation on the target word, but no legal launch fixation from the prior material, either because there was no fixation on this material (in a few items, the critical word appeared relatively near to the beginning of the sentence) or because the launch fixation was eliminated based on duration cutoffs of 80 and 1000 ms.

Table 3. Number of trials averaged into each cell mean for the 2 X 2 experimental design, for six representative subjects from Gollan et al. (2011).

Subject	HF-HP	HF-LP	LF-HP	LF-LP
1	12	12	9	13
2	9	10	13	9
3	13	14	15	14
4	12	14	15	15
5	13	11	12	14
6	12	14	13	12

Table 4. Number of trials averaged into each cell mean for the 3 X 2 X 2 quasi-experimental design, for the same six subjects from Gollan et al. (2011) included in Table 3. "HH" denotes High Frequency, High Predictability, etc.

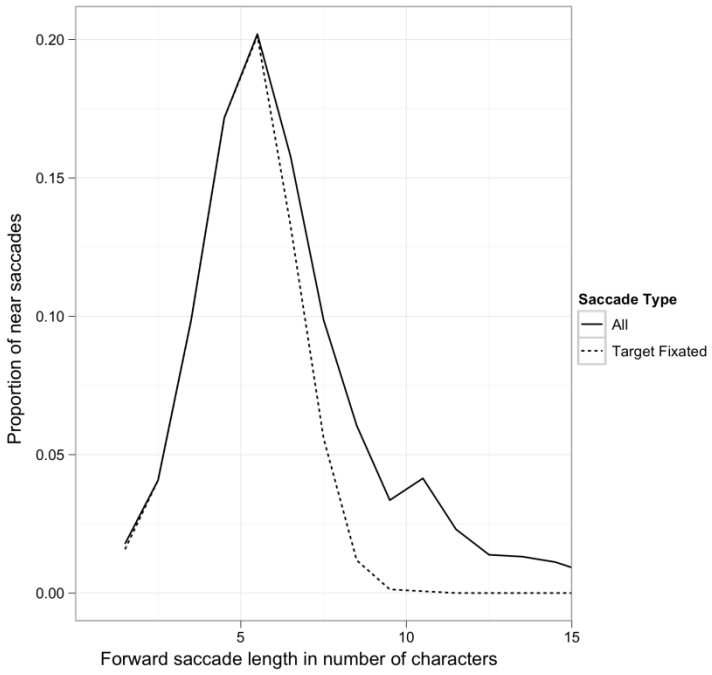
Subject	Near HH	Near HL	Near LH	Near LL	Med HH	Med HL	Med LH	Med LL	Far HH	Far HL	Far LH	Far LL
1	3	1	1	3	1	3	1	3	2	3	6	6
2	1	2	missing	2	1	7	3	7	6	2	6	missing
3	9	5	5	6	4	7	8	6	missing	3	1	2
4	4	4	3	4	4	6	5	6	3	4	5	4
5	1	3	2	5	7	3	4	3	2	4	4	3
6	1	3	1	1	6	3	6	8	5	2	5	2

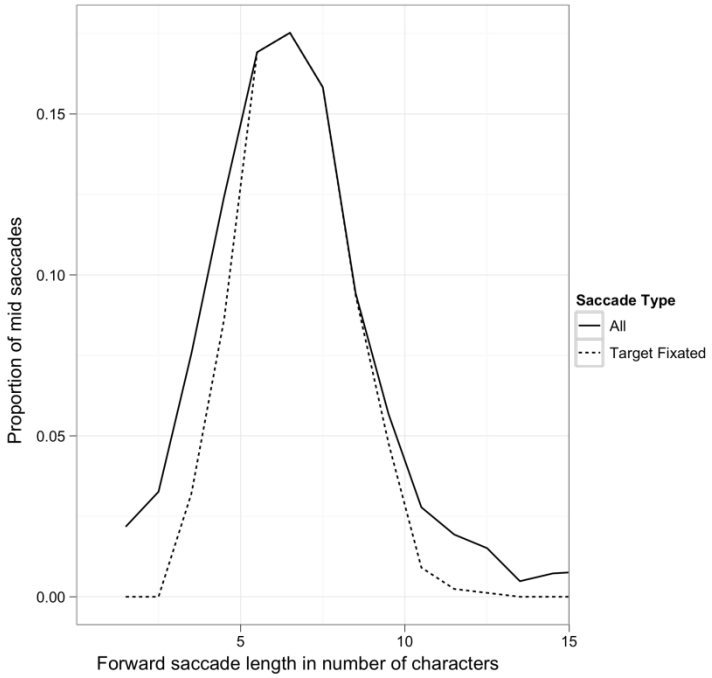
Table 5. Skipping cases broken down by launch site.

	Near	Medium	Far	Very Far
Gollan et al. (2011)	398 (65.1%)	149 (24.4%)	45 (7.4%)	19 (3.1%)
Staub (2011)	326 (66.4%)	121 (24.6%)	28 (5.7%)	16 (3.3%)

Note: Launch sites are relative to the target word and are computed by subtracting the length of the target region from the launch site of the post target fixation. Percentages given in parenthesis are for the percentage of total skips; for the Staub data, this excluded 39 trials in which there was no legal launch fixation prior to a skip.

Figures 1a-1c. Forward Saccade length distributions out of the near, middle, and far launch sites respectively, for Gollan et al. (2011) data.





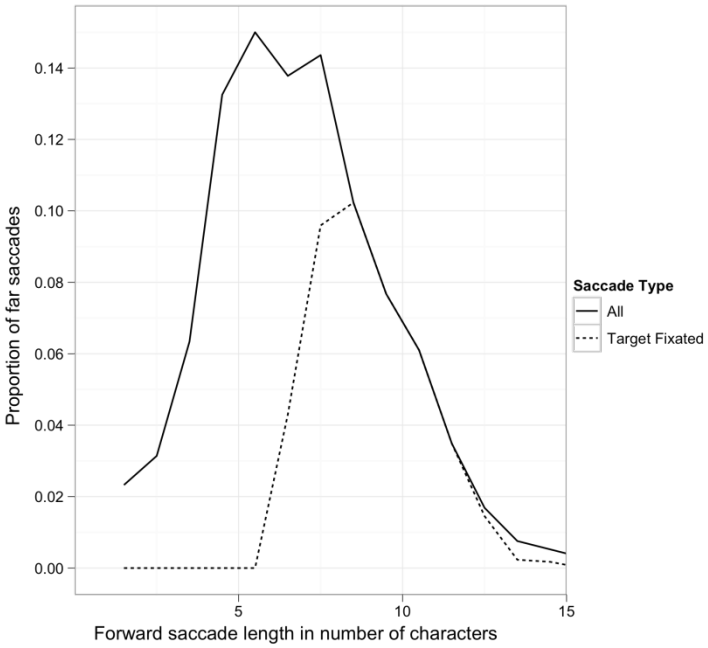

0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Proportion of far saccades
5
10
15
Forward saccade length in number of characters
Saccade Type
All
Target Fixated

Figure 2. Launch site distribution for Gollan et al. (2011) data.

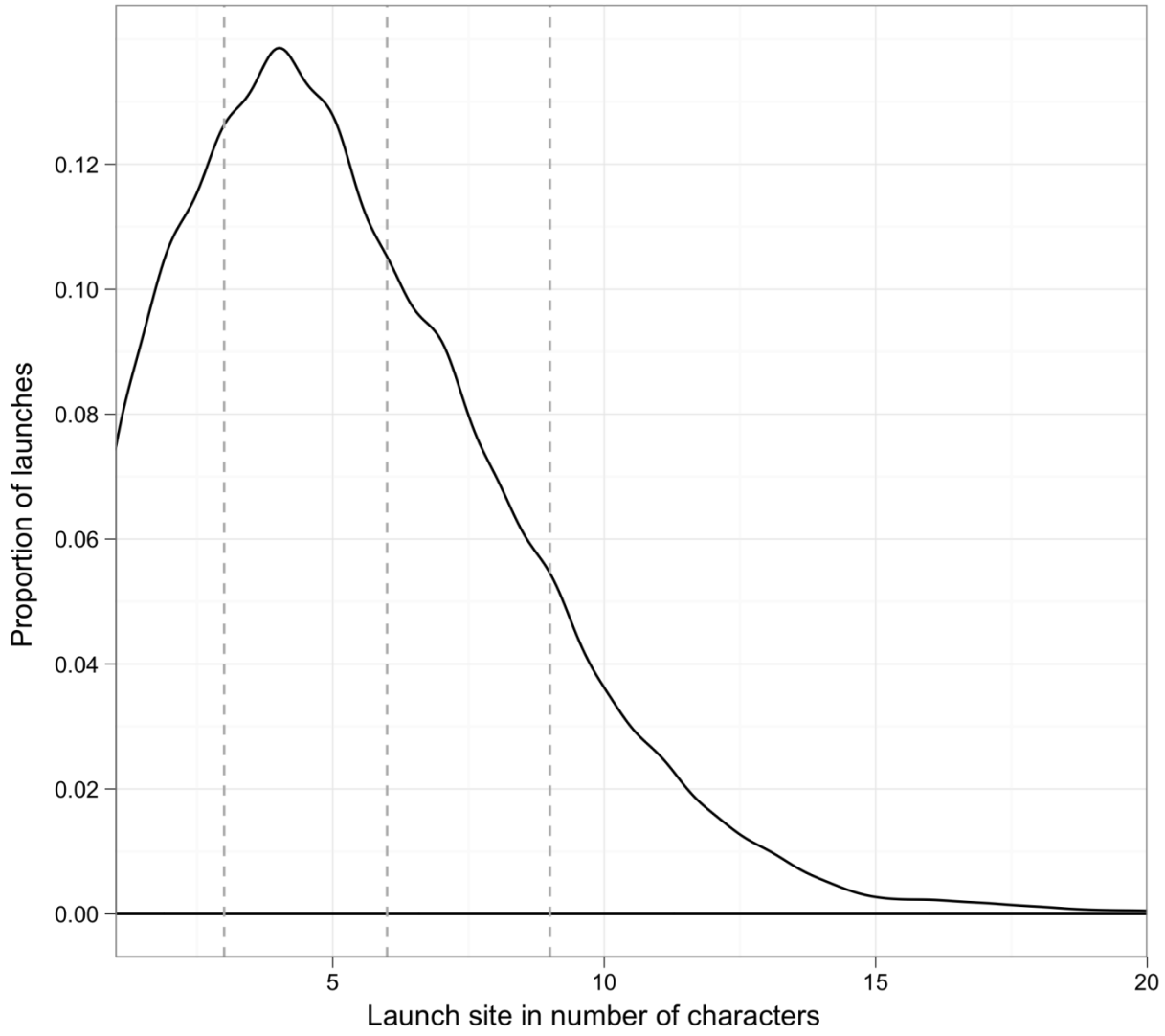


Figure 3. Plot of residuals over launch sites for the Gollan et al. (2011) data

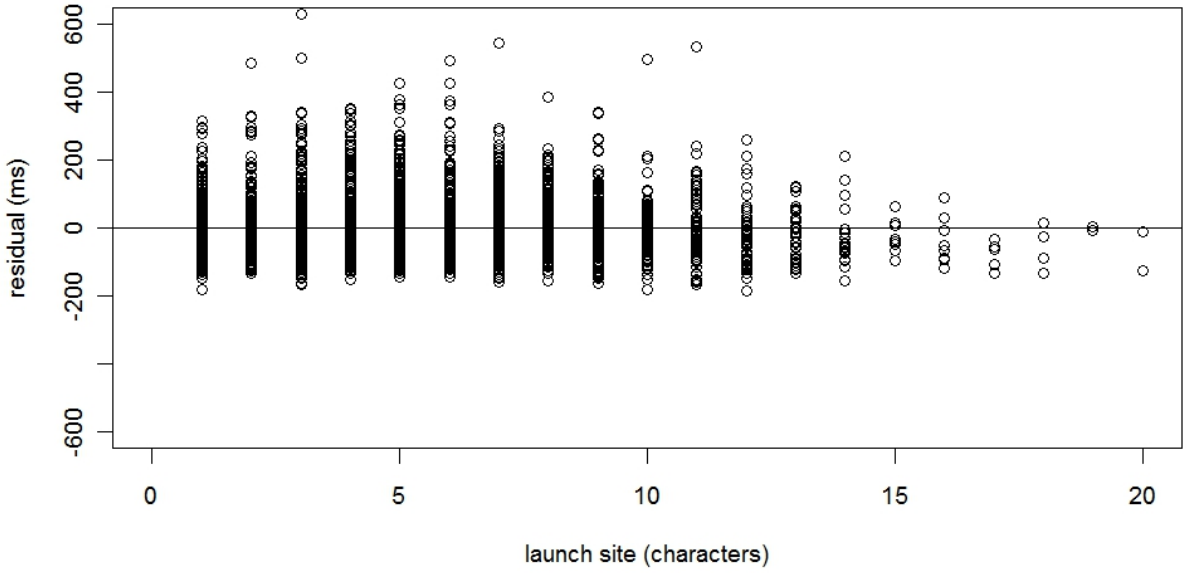


Figure 4. Plot of residuals over launch sites for the Staub (2011) data