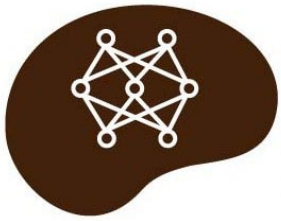

Language & Cognitive Dynamics Laboratory

Technical Report 2012.04
Version 1.0: September 2012

Aggregating fixation data across trials of different durations

Dan Mirman
Allison Britt
Pyeong Whan Cho

Abstract

When individual time series (i.e., trials) have differing durations, the researcher must decide how to aggregate data at time points where some trials do not have data. Some researchers argue that since there is no data for those trials, they should be excluded. Other researchers include those trials and "fill in" or "pad" the data with sensible values. Here we consider the specific case of gaze data from "visual world paradigm" studies and use Monte Carlo simulation to generate simplified gaze data to examine the consequences of different aggregation approaches. The results provide a concrete demonstration that excluding trials with no data is a form of selection bias that systematically distorts results. Unbiased data aggregation requires that the "denominator" (i.e., number of trials) remain the same for each time point in the analysis.

Aggregating fixation data across trials of different durations

In a typical "visual world paradigm" (VWP) eye tracking study of spoken language comprehension, participants view an array of images or words on a computer screen, and hear a linguistic stimulus relevant to one or more of the elements of the array. The trial-level data are strongly constrained by the physical mechanics of the oculomotor system, so these data are aggregated over trials by condition and participant (or item) in order to reduce the oculomotor contributions and extract the underlying cognitive contributions. One of the most basic and oft-ignored issues in analysis of eye tracking (and other time series) data is how to aggregate data across trials of different durations. Typically, a trial ends when the participant responds, which naturally leads to some trials that are shorter than others. It is hypothetically possible to restrict the analysis to the time window before any trials have terminated, but this is generally impractical because the time course of processing usually evolves over a substantially longer time window than the single shortest response time. So, when computing fixation proportions at later time points, should terminated trials be included or not? Three approaches are currently in use: (1) for each time bin, include all trials and count post-response frames as non-object fixations (i.e., the participant is done fixating all objects from this trial), (2) include all trials and count post-response frames as target fixations (i.e., if the participant selected the correct object, then consider all subsequent fixations to be on that object; note that, typically, any trials on which the participant made an incorrect response are excluded from analysis), (3) include only trials that are currently on-going and ignore any terminated trials since there is no data for those trials. It is important to note that researchers rarely even report which of these aggregation methods they used, so a formal survey of the literature is not possible.

In the domain of applied regression, the challenges of missing data are well-known, as are various approaches for handling them (e.g., Gelman & Hill, 2007, Chapter 25), but regression techniques have only recently penetrated into the world of eye tracking and psycholinguistics (e.g., Baayen et al., 2008; Barr, 2008; Mirman, Dixon, & Magnuson, 2008). Critically, aggregating data over trials of different lengths differs from other cases of missing data that may be more familiar in psycholinguistics. Focusing on eye tracking studies, data points may be missing for essentially random reasons such as equipment failures, blinks, etc., but these would affect all time points and conditions equally. In contrast, trial termination times are intrinsically related to cognitive processing – more difficult trials tend to last longer. Thus, the selective elimination of completed trials represents a form of selection bias. To our knowledge, the consequences of choosing one aggregation method over another have not been examined concretely. There are two points concerning the relative merits of the three methods of data aggregation described above. First, the two methods that include all trials are fundamentally the same – they will capture the same data and merely depict those data differently, in the same way that probability distribution curves and cumulative distribution curves depict the same underlying data in somewhat different ways. As we will show, depending on the researcher's goals, one visualization method or the other may be more effective or appropriate. Second, ignoring terminated trials represents a form of selection bias that will distort the data. This is because trials do not terminate at random, so as the time series progresses through the time window, the data move further and further from the complete, unbiased set of trials to a biased subset of only trials that required additional processing time. This bias will operate both between conditions (i.e., more trials from a condition with difficult stimuli than from a condition with easy stimuli) and within conditions (i.e., more of the trials that were difficult than that were easy within a condition). We will demonstrate both of these points using Monte Carlo simulation.

Methods

To examine the effect of different aggregation methods, one must know the true pattern of underlying data, so we used a Monte Carlo simulation procedure to generate simplified eye tracking data. The simulations were designed to model a simple spoken word recognition experiment with two conditions: "Easy" and "Hard", where response times are about 400ms slower in the Hard condition, as might result from a manipulation of word frequency, cohort density, or other lexical variables (e.g., Magnuson, Dixon, Tanenhaus, & Aslin, 2007). Response times were sampled randomly from the gamma distributions shown in the left panel of Figure 1 (shape = 4, scale = 1 for both distributions; to convert raw gamma distribution values to the RT range, they were multiplied by 200 and 800 ("Easy") or 1200 ("Hard") was added). For the target fixation analysis, we simulated the simplest case: participants make exactly one 500ms fixation on the target just before clicking on it (i.e., there was a target fixation for the 500ms interval ending with that trial's RT). This is schematically shown in the right panel of Figure 1 for a random subset of 50 trials (25 from each of the two conditions).

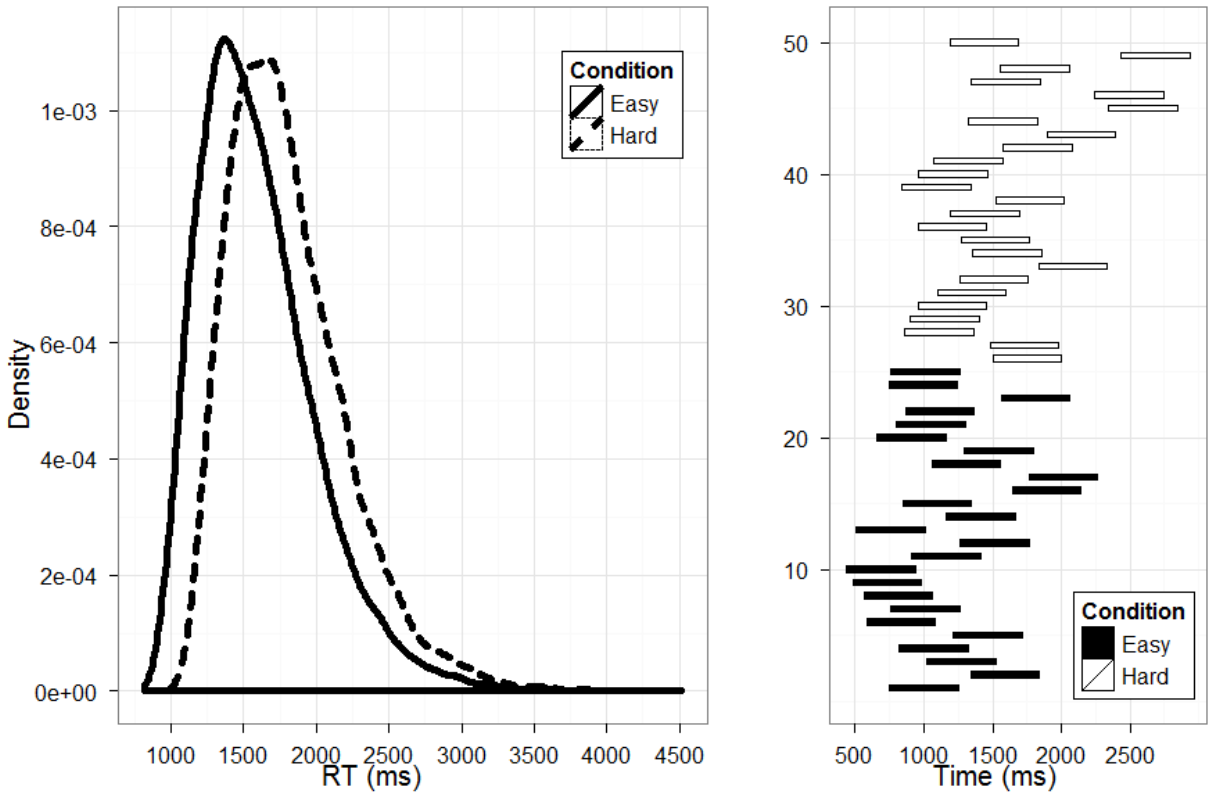

Figure 1. Left panel: RT distributions from Monte Carlo simulations of "Easy" and "Hard" conditions. Right panel: Time periods of target fixation for a subset of 50 trials (25 from each condition).

For the competitor fixation analysis, 75% of the way through each trial there was a 50% probability of a 500ms fixation to the "Competitor" (i.e., related) distractor and a 25% probability of a 500ms fixation to the "Unrelated" distractor. For each analysis, 50,000 trials were simulated (25,000 for each condition) and the results aggregated according to the three different methods described above: (a) consider all post-response data as a fixation somewhere other than the critical objects ("non-object fixation"), (b) consider all post-response data as target fixation, or (c) consider only on-going trials. Note that for the competitor fixation analysis, methods (a) and (b) are equivalent (because the target is, by definition, not a competitor object) so they will not be distinguished in the presentation of the results. In VWP experiments, researchers typically exclude some trials from analysis (e.g., error responses). For

simplicity, we did not exclude any trials from this analysis. The analyses were conducted in R version 2.13 (R Development Core Team, 2011).

Results

Target Fixations

The target fixation time courses based on each of the three aggregation methods are shown in Figure 2. Both of the first two methods, which included all trials in the analysis, correctly captured the approximately 400ms difference between the Easy and Hard conditions that was in the RT data. As mentioned above, these methods reflect the same underlying data but depict them somewhat differently. Researchers can choose the visualization method that best fits their goals. For example, considering post-response data as non-object fixations produces target fixation proportion curves that most closely mimic the RT distributions (compare Figure 2 and Figure 1) and the differences between the conditions were slightly more visible in the 2000-2500ms time range; on the other hand, considering post-response data as target fixations produces curves that more intuitively map on to activation curves from computational models such as TRACE (McClelland & Elman, 1986), which is important for many VWP studies of spoken word recognition (e.g., Allopenna et al., 1998; Dahan, Magnuson, & Tanenhaus, 2001; Mirman et al., 2008).

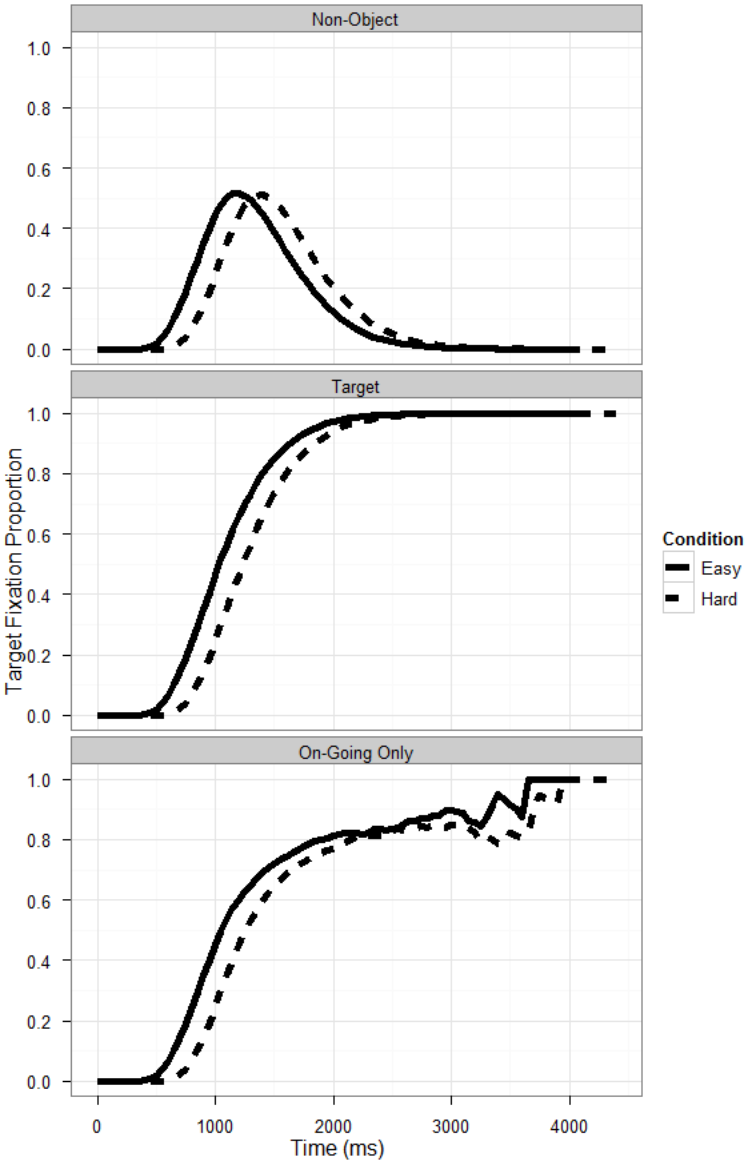

Figure 2. Target fixation time course based on three different aggregation methods.

The third method, which considered only on-going trials, mostly followed the same pattern, but the data became noisy and inconsistent in the later portion of the time window, in this example, after about 2000ms. There are two effects contributing to this. First, the reliability of any estimate (i.e., fixation probability) decreases as the sample size (i.e., number of trials) decreases. However, that this was not just restricted to the "tail" of the data: about 20% of the response times were greater than 2000ms. This would be a rather large amount of data to exclude; nevertheless, one might imagine that truncating the data at that point would avoid this problem. However, this brings up the second contributing effect:

these late data are not evenly distributed across the two conditions. About 25% of the Hard condition response times were greater than 2000ms, but only 15% of the Easy condition response times were greater than 2000ms. That is, considering only on-going trials creates a Catch-22: the late data potentially reflect interesting condition differences (e.g., one condition harder than another), but the late data are noisy and potentially contain spurious effects (e.g., the visually striking, but completely spurious, re-emergence of the Easy vs. Hard condition effect around 3000ms in Figure 2). In contrast, the two methods that consider all trials, smoothly reflect the progressive disappearance of differences between conditions (and the effects near the asymptotes may be more visible on a logit scale: Barr, 2008; Jaeger, 2008).

The distorting effect of considering only ongoing trials is even more striking if the difference between the "Easy" and "Hard" conditions is reflected in the skew of the RT distribution rather than a simple 400ms shift (e.g., Balota & Spieler, 1999; Balota, Yap, Cortese, & Watson, 2008). This is depicted in Figure 3, where the difference between the "Easy" and "Hard" conditions was in the shape parameter (4 vs. 6; as in the previous simulation, the raw gamma distribution values were multiplied by 200 and 800 was added for both conditions, so the difference in the shape parameter produced an approximately 400ms difference in mean RT between conditions). In this case, considering post-response data as target fixations accurately captured the processing speed difference between conditions, but considering only on-going trials caused it to look more like an asymptote level difference (i.e., for the red curves in Figure 3, target fixations in the "Hard" condition appear to reach a lower asymptote level than in the "Easy" condition). This demonstrates that reported asymptote differences based on considering only on-going trials (e.g., McMurray, Samuelson, Lee, & Tomblin, 2010) need to be interpreted with caution.

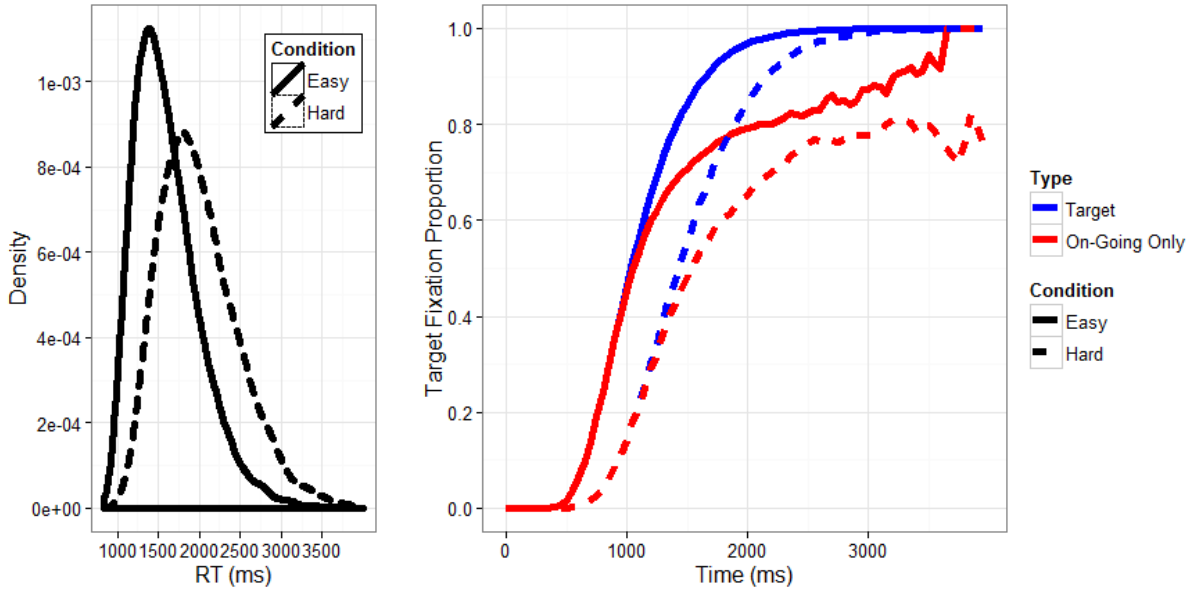

Figure 3. Left panel: RT distributions from Monte Carlo simulations of "Easy" and "Hard" conditions based on differences in the shape parameter. Right panel: Target fixation time course based on including only on-going trials (red) and considering post-response data as target fixations (blue).

Competitor Fixations

For analysis of competitor fixations, considering only on-going trials similarly led to distortion in the later time windows (Figure 4). When post-response data were considered as non-object (or, equally, target) fixations, the resulting competition effect was approximately symmetric (i.e., the rise and fall of the solid line in bottom panel of Figure 4 is approximately symmetric), reflecting the approximately symmetric

underlying procedure that produced the fixation data. In contrast, considering only on-going trials produced the appearance of a strongly asymmetric competition effect that persisted at an approximately stable level from 2500ms to 3500ms. Since we defined the algorithm that produced the data, we know that there was no late competition effect: fixations for both the related and unrelated competitor were tied to trial response times with a constant greater probability of fixating the related competitor than the unrelated competitor. Thus, the appearance that fixation of related competitors persisted longer than fixation of unrelated competitors is completely spurious.

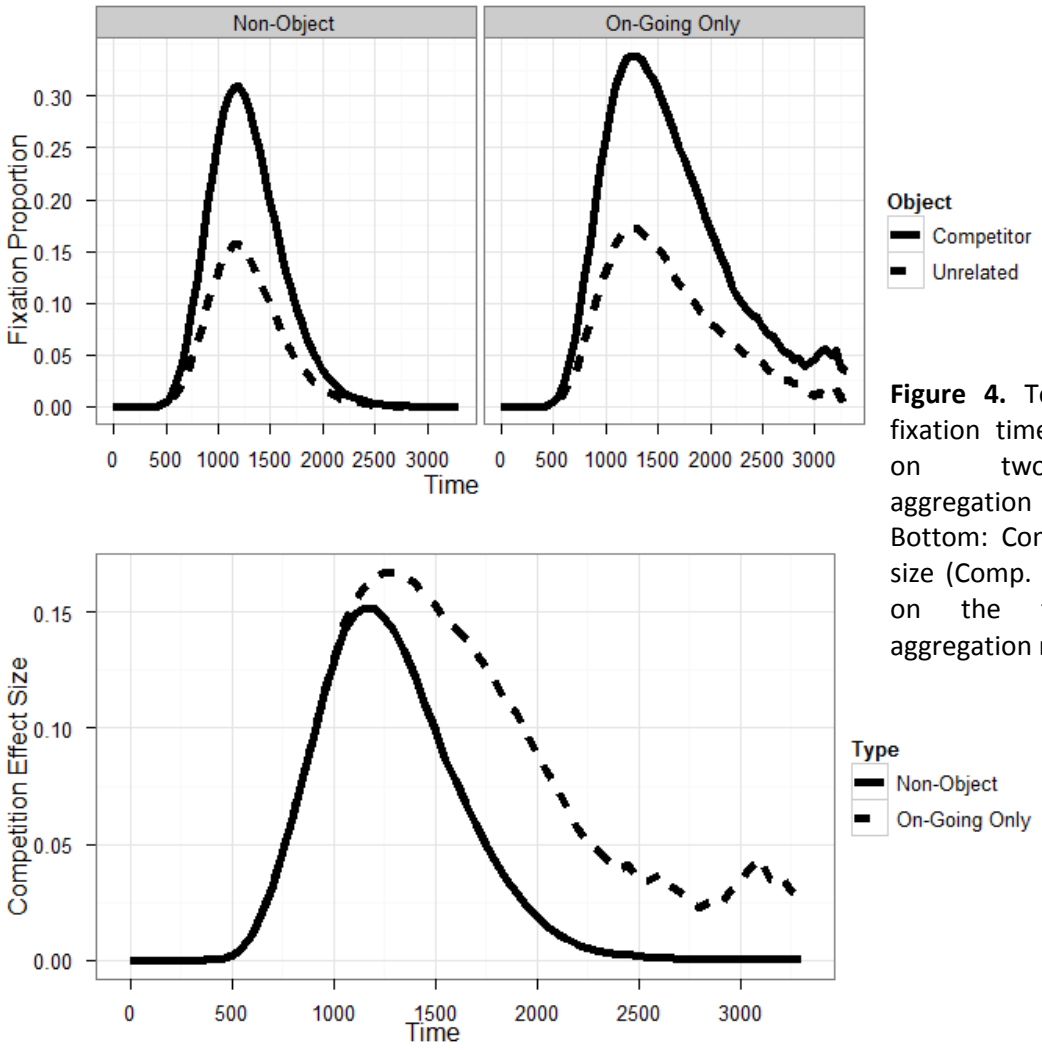

Figure 4. Top: Competitor fixation time course based on two different aggregation methods. Bottom: Competition effect size (Comp. – Unrel.) based on the two different aggregation methods.

As mentioned above, this spurious effect is due to biased sampling, not merely to noisier estimation due to a smaller number of data points. To demonstrate this concretely and to examine how this affects analysis of real visual world paradigm eye tracking data, we considered data from the taxonomic competition condition from a recent study (i.e., more looks to a member of the same semantic category than to unrelated objects; Mirman & Graziano, 2012). To simulate random data loss, we removed a random sample of 10% and 20% of the data points. Figure 5 shows the competition effect size based on the two aggregation methods for each of the three data sets (complete data, 90% of data points, and 80% of data points). Considering only on-going trials produced the appearance of a more persistent competition effect whereas random data loss had virtually no effect on the competition effect size. Thus, considering only on-going trials produces the appearance of an overly persistent competition

effect and this is intrinsic to the aggregation method, not due to simply having a smaller number of data points.

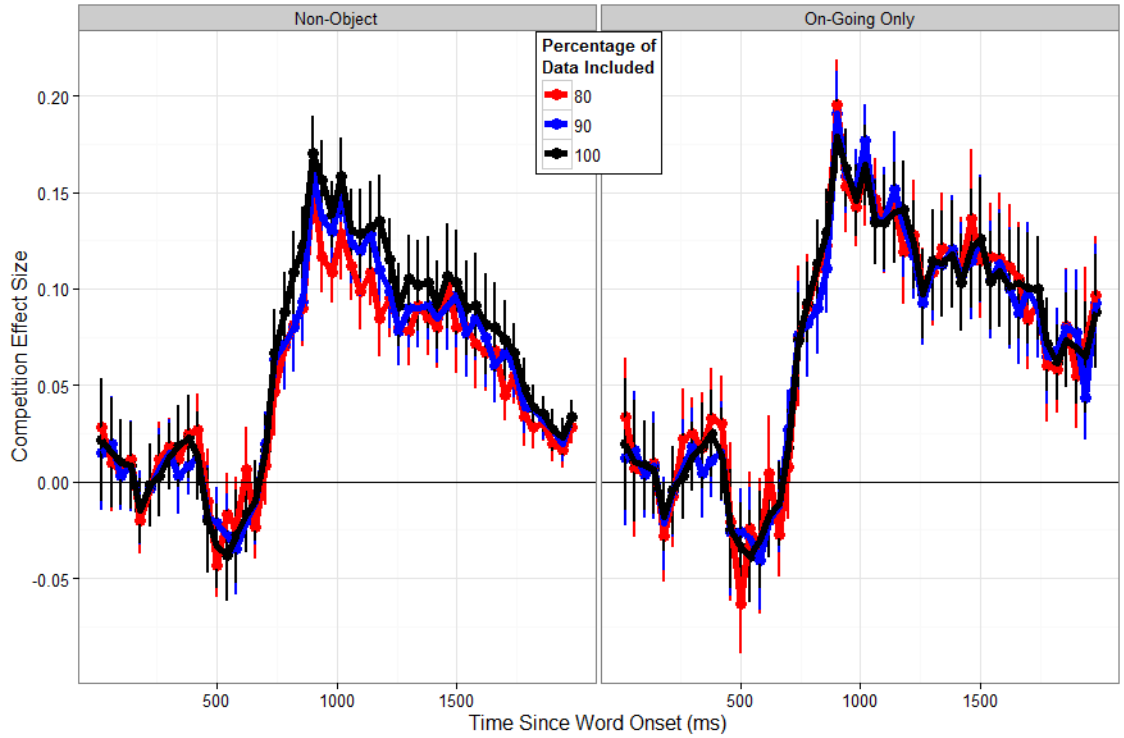

Figure 5. Effects of moderate data loss and different aggregation methods on taxonomic competition effect sizes (error bars represent ±SE).

Discussion

We used Monte Carlo simulation to examine the consequences of different methods of aggregating data across time series (i.e., trials) of different durations. This issue is particularly important for studies -- such as typical "visual world paradigm" (VWP) experiments -- in which trial durations are determined by the participant (e.g., the trial ends when the participant makes a response) and thus typically have different durations. Using Monte Carlo simulation to generate pseudo-VWP fixation data allowed us to precisely determine the underlying data and effects so that we could identify distortions due to aggregation methods. Additional analyses of real VWP data demonstrated that the observed patterns hold for real data, not just simulated data with particular parameters. We considered three approaches to dealing with data from terminated trials: (1) treat all post-response data as a fixation somewhere other than the critical objects ("non-object fixation"), (2) consider all post-response data as target fixation, or (3) consider only on-going trials. The simulations illustrated that there are no substantive differences between the first two methods because they are equivalent for competitor fixation analyses and are merely different depictions of the same underlying data for target fixations (probability distribution functions vs. cumulative distribution functions). In contrast, considering only on-going trials caused distortions at later time points because terminated trials were being selectively removed from analysis. Direct comparisons with random data loss demonstrated that it is indeed a selection bias and not merely noisier estimation due to fewer data points.

Since slow response times are typically causally related to competition (i.e., competition slows down responses) or other processing difficulty, distortion in the late time window can have serious theoretical consequences. To understand why this is the case, imagine that we want to evaluate the response rate

over time to a drug for a deadly disease. We enroll 100 participants in the trial and administer the drug. At first, only 50% of the participants respond to the drug. As the trial progresses, the non-responders begin to, unfortunately, die. After 6 months, only 75 participants are alive and participating in the trial and the same 50 are responding to the treatment. At this point, is the response rate the same 50% or has it risen to 67%? Would it be accurate to conclude that responsiveness to the treatment increases after 6 months? This example, hopefully, makes it intuitively clear why considering only on-going trials is a form of selection bias that will systematically distort the results. In other words, unbiased data aggregation requires that the denominator of the proportion calculation remain the same over the full time course.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory & Language*, *38*(4), 419-439.

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

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: beyond measures of central tendency. *Journal of Experimental Psychology: General*, *128*(1), 32-55.

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (2008). Beyond mean response latency: Response time distributional analyses of semantic priming. *Journal of Memory & Language*, *59*(4), 495-523. doi:10.1016/j.jml.2007.10.004

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory & Language*, *59*(4), 457-474. doi:10.1016/j.jml.2007.09.002

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*(4), 317-367.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY, USA: Cambridge University Press.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory & Language*, *59*(4), 434-446.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*, 1-24.

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cognitive Psychology*, *60*(1), 1-39.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory & Language*, *59*(4), 475-494. doi:10.1016/j.jml.2007.11.006

Mirman, D., & Graziano, K. M. (2012). Individual differences in the strength of taxonomic versus thematic relations. *Journal of Experimental Psychology: General*. doi:10.1037/a0026451

R Development Core Team. (2011). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/