False positive rates in standard analyses of eye movements in reading

Titus von der Malsburg[1], Bernhard Angele[2]
[1] University of California, San Diego, [2] Bournemouth University

April 28, 2015

Abstract: In research on eye movements in reading, it is common to analyze a number of canonical dependent measures in order to study how the effects of a manipulation unfold over time. Although this gives rise to the well-known multiple comparisons problem, i.e. an inflated probability that the null hypothesis is incorrectly rejected (Type I error), it is accepted standard practice not to apply any correction procedures. Instead, there is a widespread belief that corrections are not necessary because the increase in false positives is too small to matter. To our knowledge, no formal argument has ever been made to justify this assumption. Here, we report an investigation of this issue using Monte Carlo simulations. Our results show that false positives are in fact increased to unacceptable levels when no correction is applied, which casts doubt on the assumptions that Type I error rate increases are too small to matter in practice. We also tested two stricter alternative criteria for determining the reliability of an effect and found that the Bonferroni correction controls false positives effectively while only moderately reducing power. Thus, there is little reason why the Bonferroni correction should not be made a standard requirement for analyses of eye movement data in reading.

Introduction

A key advantage of using eye tracking methods in reading research is the wealth of the data that can be collected. The entire sequence of fixations that a participant produces during a trial is recorded and, in theory, available for analysis. In practice, some standard aggregate measures have been established to adequately summarize this wealth of data (Rayner, 1998). Most commonly, the analysis region for which these measures are computed will contain one word or phrase within a longer sentence. Over the last decades, many different measures have been proposed, but there are a handful of standard measures that almost every eye tracking study reports. In order to compute these measures, the fixation sequence is divided into the first pass, consisting of the fixations occurring when a reader first enters a word from the left, and the second pass, consisting of the fixations occurring when a word is revisited (see

Vasishth et al., 2013, for a discussion of the distinction between early and late measures). First pass and second pass fixations are then filtered and combined to form the standard aggregate measures:

(1) *First fixation duration (FFD)*: the duration of the very first fixation a reader made on a region, regardless of whether it was refixated thereafter or not.
(2) *Single fixation duration (SFD)*: the same as first fixation duration, but only counting cases where a region was not refixated during first pass.
(3) *Gaze duration (GZD)*: the sum of the duration of the first fixation and all refixations during first pass, that is, before the gaze left the word for the first time.
(4) *Go-past duration (GPD)*: also known as regression path duration, the sum of all fixations from when a reader's gaze first entered a word from the left, including all refixations and all regressive fixations to prior words, until it left the word towards the right.
(5) *Total viewing time (TVT)*: the sum of the durations of all fixation on the word, regardless of whether they occurred during the first or second pass.

Again, the above list of measures is not exhaustive, but the five measures described are those most commonly used. From the description of these measures it should already be apparent that they are all correlated to different degrees. These correlations, combined with the fact that several of these measures are potentially informative, leads to a problematic situation.

Let's assume that a researcher wants to test whether a certain manipulation (e.g., two presentation conditions) has an impact on reading behavior. He or she will then compute the condition means for each measure and, in order to test whether those means are significantly different, compute a series of ANOVAs or linear mixed-effects models, one for each measure. In the absence of a specific hypothesis about the impact of the manipulation on the reading process, the simplest and most commonly used decision strategy is to conclude that reading is affected by the manipulation if at least one of these analyses indicates a significant difference in means. Which measure is affected is not important under this decision strategy; at best, it might provide some extra information about the time course of the reading process.

The problem with this decision strategy is the following: Assuming that significance is determined using an *alpha* threshold of 0.05, each test produces a false positive with a probability of 5%. Assuming further that five independent eyetracking measures are tested, the probability that at least one of these tests produces a false positive result increases to $1 - 0.95^5 = 0.23$. This means that the probability of finding a spurious result would be 23% instead of the conventionally accepted 5%.

Of course, in reality, eye tracking are not quite independent as the different fixation time measures are typically highly correlated. For example, the correlation between first fixation duration and gaze duration is determined by the rate at which readers make refixations. If readers make no refixations at all, FFD and GZD are identical, in other words, they have a correlation of 1. The more first-pass refixations readers perform, the lower is the correlation between FFD and GZD. However, since the fixations that are used to compute FFD form a subset of those used to compute GZD, the correlation will typically not reach zero.

The fact that the canonical eye tracking measures are correlated implies that the probability of a falsely declared effect is not quite as high as the 23% in the example above. We suspect that this is one reason why many reading researchers draw conclusions as if the false positive rate had only been 5%. However, whether this simplifying assumption is warranted or not is unclear. If the correlations between fixation time measures were always 1, multiple comparisons would indeed not be cause for concern; all tests would necessarily produce the same result such that no test beyond the first could possibly produce additional false positives. However, in that case, analyzing different fixation time measures would be completely redundant. Eye movement researchers are clearly aware that different fixation time measures share some information, but that each measure also contributes unique information. Thus, the true false positive rate in an eye tracking study with five dependent measures is somewhere between 5% and 23%.

If the rate of false positives is inflated to unacceptable levels due to multiple comparisons, adjustments to the decision making process are needed. The text-book solution for that is the Bonferroni correction. All a researcher has to do in order to apply this correction is to divide the threshold for determining significance (typically 0.05) by the number of tests that were performed and to use the resulting number as the new threshold. In the present scenario, that means that significance would be determined using the threshold $\alpha = 0.05/5 = 0.01$. If the threshold is lowered in this way, the probability of finding at least one false positive result in any of the multiple statistical tests is 5%. Unfortunately, the Bonferroni correction is not entirely appropriate for analyses of eye tracking measures because it assumes that the statistical tests are independent, which, as we discussed above, is not the case. The Bonferroni correction may therefore potentially be too conservative and will effectively sacrifice more statistical power than necessary.

This concern about loss of statistical power may be another reason why reading researchers do not apply corrections for multiple comparisons such as the Bonferroni correction. However, this leads us to a dilemma: applying a Bonferroni correction will lead to an unnecessary loss of statistical power, but the alternative, applying no correction at all, will inflate the probability of false rejections of the null hypothesis. The question is thus: on which side should we err?

Simmons et al. (2011) argue that false rejections of the null hypothesis are a particularly costly type of error that has the potential to threaten the integrity of the scientific discourse more than false negative results. McElreath and Smaldino (2015) illustrate and substantiate this point using a formal model of the population dynamics of scientific discovery. Their work shows that false positives results can easily proliferate and clog up the scientific discovery because the resources for weeding them out are spread too thinly. However, whether more harm is done by correcting or by not correcting for multiple comparisons crucially depends on the true rate of false positives in analyses of reading measures and on how much power is really sacrificed when the Bonferroni is applied. In the following, we present the results of a Monte Carlo simulation study that sheds light on these questions.

Overview

The purpose of the simulations presented below was two-fold: First, we determined by how much the rate of false positives is elevated when multiple dependent measures are tested without correcting for multiple comparisons. Since we found that false positive rates are indeed a serious concern, we also tested two alternative decision strategies intended to limit false positive rates to conventionally accepted levels while preserving as much statistical power as possible.

The first alternative decision strategy was to lower the alpha threshold using the Bonferroni method and to declare an effect of the manipulation only if at least one of the eye tracking measures showed an significant difference between conditions with a p-value at or below that lowered level.

The second alternative decision strategy was a rule-of-thumb criterion implicitly used by some researchers when deciding how to interpret a result and by reviewers when deciding whether the results seem robust enough for publication. In order to declare an effect reliable, this criterion requires that at least two eye tracking measures show the effect with p-values at or below 0.05. The number of measures required to show a significant effect will vary between researchers, but we suspect a majority would consider an effect showing in two measures more reliable than one that only shows in one measure. While this is a stricter criterion than our baseline, which requires just one such effect, it is unclear if it is sufficiently strict or perhaps even too strict.

To assess false positive rates and statistical power under the baseline strategy (no correction) and the two alternative strategies, we employed a Monte Carlo simulation approach similar to that used in a recent study by Barr et al. (2013). The basic idea is to generate artificial data sets with properties resembling those found in real eye tracking data and to analyze these data sets in order to determine false positive rates and statistical power. This is possible because we know the population-level effect present in artificial data and thus can ascertain whether the outcome of a statistical test is accurate or a type I (false positive) or type II error (false negative).

During each iteration of the simulation, we generated a number of artificial data sets that contained the eye movement measures mentioned above (FFD, GZD, GPD, TVT; we did not use SFD since it is just a subset of FFD) and a number of artificial subjects and items. One of these data sets had no effect of the hypothetical manipulation and the other data sets had increasingly larger effects. This allowed us to check the rate of false positives in data sets in which no true effect was present and the rate of correctly detected effects in data sets with true effects. By iterating these simulations many times, we could calculate the rate of false positives and correct detections with high precision. In each iteration, the presence of an effect was determined using all three decision criteria discussed above. P-values were computed by fitting linear mixed models and conducting likelihood-ratio tests comparing models with the factor for condition vs. models without this factor (Bates et al., 2014).

Generation of artificial data

Our approach to estimating false positive rates and power requires that we generate large numbers of artificial data sets. For the results to be valid, it is important that these data sets have statistical properties similar to those found in real eye tracking data. Given the complex nature of eye movements in reading, this is not entirely trivial. After all, the ability to generate realistic eye movement data presupposes complete knowledge about the principles and mechanisms underlying the reading process. Hence, we have to make a number of simplifying assumptions for our generative model:

- The target word is fixated at least once.
- The target word can be refixated during first pass leading to gaze durations being longer than first fixation durations.
- A regressive eye movement can occur during the first pass leading to elevated go-past durations.
- A word can be refixated after the first pass leading to increased total viewing times.
- The artificial subjects differ in the average duration of their first fixations.
- The artificial items differ in the average duration of the first fixations they elicit.

A lot of characteristics of eye movements in reading are not captured by this model. However, as we will see below, the eyetracking measures generated with this model are sufficiently similar to real data. In particular, the correlations between the various eyetracking measures are largely preserved. This is important because, as discussed above, these correlations influence false positive rates. Note that the sole purpose of this model is to generate eyetracking measures for our simulation study. As such, it has a different scope and purpose than fully fledged models of eye movement control in reading such as E-Z Reader and SWIFT (Reichle et al., 1998; Engbert et al., 2005).

Parameters of the generative model

A complete definition of the generative model for eyetracking measures requires that we define a number of parameters:

- `n.subjects`: the number of subjects
- `n.items`: the number of items
- `p.refix`: the probability of a refixation
- `p.regr`: the probability of a regression
- `p.reread`: the probability of rereading after the first pass
- `mean.ffd`: the average duration of the first fixation
- `mean.gazediff`: the average duration of the refixation
- `mean.gopastdiff`: the average sum of the durations of all fixation occurring after a regressive saccade and before anything to the right of the word is fixated (duration of the regression path)

- `mean.tvtdiff`: the average sum of all fixations on the word occurring after the first pass
- `sd.ffd`, `sd.gazediff`, `sd.gopastdiff`, `sd.tvtdiff`: the standard deviations of the above durations
- `sd.subjects`, `sd.items`: the standard deviations of the random effects for subjects and items

The model samples durations from log-normal distributions which resemble the true distributions of the eyetracking measures more closely than normal distributions. All means and standard deviations are given as geometric means and geometric standard deviations.

Generation of eyetracking measures

Our model generates the following eyetracking measures: First fixation duration (FFD), Gaze duration (GZD), Go-past duration (GPD), and Total viewing time (TVT). First fixation durations were directly sampled from a log-normal distribution defined by $\mu =$ `mean.ffd` and $\sigma =$ `sd.ffd`. For each subject, a value was sampled from a normal distribution centered at zero and with standard deviation `sd.subjects`. This value was added to all first fixation durations of that subject. Likewise, differences between items were introduced by sampling from a normal distribution with standard deviation `sd.items`. Based on the parameter `p.refix`, the model determined whether or not there was a refixation in a trial. If yes, gaze duration was generated by sampling from the distribution defined by $\mu =$ `mean.gazediff` and $\sigma =$ `sd.gazediff` and by adding the corresponding first fixation duration. If there was no refixation, the gaze duration was the same as the first fixation duration. Go-past times and total viewing times were calculated similarly. The resulting data set also had a column for condition (coded as -0.5 and 0.5). Data sets were generated with true effects of the following sizes: 0 ms, 2.5 ms, 5 ms, 10 ms, 20 ms, and 40 ms. The function used for generating the data can be seen below.

```
new.etmeasures <- function(p) {

    n     <- p$n.subjects * p$n.items

    subj <- rep(1:p$n.subjects, each=p$n.items)
    item <- rep(1:p$n.items, p$n.subjects)
    cond <- rep(c(-0.5,0.5), p$n.subjects*p$n.items/2)
    d    <- data.frame(subj=subj, item=item, cond=cond)

    # Adding random intercepts and slopes for subjects:
    re.int.subj <- rnorm(p$n.subjects, 1, p$sd.subjects)
    re.int.subj <- rep(re.int.subj, each=p$n.items)

    # Adding random intercepts and slopes for items:
```

```
    re.int.item <- rnorm(p$n.items, 1, p$sd.items)
    re.int.item <- rep(re.int.item, p$n.subjects)

    # Sample ffds from log-normal:
    d$ffd <- (rlnorm(n, p$mean.ffd, p$sd.ffd) + re.int.subj + re.int.item)

    # Generate refixations, regressions, and rereads:
    d$refix  <- rbinom(n, 1, p$p.refix)
    d$regr   <- rbinom(n, 1, p$p.regr)
    d$reread <- rbinom(n, 1, p$p.reread)

    # Generate gaze durations, go-past times, and total viewing times:
    gazediff   <- rlnorm(n, p$mean.gazediff,   p$sd.gazediff)
    gopastdiff <- rlnorm(n, p$mean.gopastdiff, p$sd.gopastdiff)
    tvtdiff    <- rlnorm(n, p$mean.tvtdiff,    p$sd.tvtdiff)

    d$gzd <- d$ffd + ifelse(d$refix,  gazediff,   0)
    d$gpd <- d$gzd + ifelse(d$regr,   gopastdiff, 0)
    d$tvt <- d$gzd + ifelse(d$reread, gazediff,   0)

    d$trial <- 1:nrow(d)

}
```

Calculation of model parameters

In order to obtain sensible values for our model parameters, we used a real-world eyetracking data set, namely Experiment 1 from Angele et al. (2013). This data set was ideal for our purposes since it contained all the measures described above (not all were reported in Angele et al., 2013) and because the design of that study was representative of the most common type of reading study. For calculating our model parameters, we only used data from the control condition. The number of subjects and items was set to values that are common in eyetracking experiments. The other parameters were calculated from the empirical data as listed in Table 1.

Evaluation of artificial data

For the present purpose, it is important that the artificial data exhibits similar variance components and correlations among eyetracking measures as found in the real data. Below, we examine the artificial data generated based on the parameters listed above. As a first step, we calculated the pair-wise correlations of the four eyetracking measures in the real data and in the artificial data (see Table 2).

Parameter name	Value	Parameter name	Value
n.subjects	40	sd.subjects	22
n.items	132	sd.items	21
p.refix	0.14		
p.regr	0.07		
p.reread	0.197		
mean.ffd	221	sd.ffd	1.3
mean.gazediff	194	sd.gazediff	1.4
mean.gopastdiff	238	sd.gopastdiff	1.8
mean.tvtdiff	236	sd.tvtdiff	1.5

Table 1: Parameters for our generative model of eyetracking measures as calculated using the Angele et al. (2013) data

	Real Data				Artificial Data			
	FFD	GZD	GPD	TVT	FFD	GZD	GPD	TVT
FFD	1.00	0.64	0.40	0.40	1.00	0.65	0.45	0.49
GZD	0.64	1.00	0.63	0.62	0.65	1.00	0.67	0.76
GPD	0.40	0.63	1.00	0.61	0.45	0.67	1.00	0.50
TVT	0.40	0.62	0.61	1.00	0.49	0.76	0.50	1.00

Table 2: Correlations of eyetracking measures in real data (left) and in artificial data (right)

A more detailed picture of the dependencies between eyetracking measures can be obtained from scatter plots that show each measure as a function of another measure (see Fig. 1 and Fig. 2). As we can see from the correlation coefficients and the scatter plots, our generative model captures the dependencies found in the real data quite faithfully. Some correlations are slightly off due to our simplifying assumptions (in most cases, the correlations in the artificial data set are slightly higher than in the real data), however, the overall pattern of correlations is preserved in the artificial data.

Monte Carlo Simulation

The simulation consisted of 100,000 iterations. In each iteration we generated six data sets, one for each tested effect size, and tested each of these data sets with the three decision criteria.

Significance testing

We fit one linear mixed model for each dependent measure (FFD, GZD, GPD, TVT). All dependent measures were log-transformed to obtain normally distributed residuals. The models had random intercepts for subjects and items. Random slopes were not needed because we generated the data such that random slopes would not systematically vary among subjects and items. *P*-values for the manipulation factor were determined by fitting models with and without the factor and by comparing these models using likelihood-ratio tests. For each linear mixed model, we also recorded whether or not there was a convergence warning. No convergence warnings occurred in the whole simulation.

```
mer0.ffd <- lmer(ffd ~ 1    + (1|item) + (1|subj), nd, REML=F)
mer0.gzd <- lmer(gzd ~ 1    + (1|item) + (1|subj), nd, REML=F)
mer0.gpd <- lmer(gpd ~ 1    + (1|item) + (1|subj), nd, REML=F)
mer0.tvt <- lmer(tvt ~ 1    + (1|item) + (1|subj), nd, REML=F)

mer1.ffd <- lmer(ffd ~ cond + (1|item) + (1|subj), nd, REML=F)
mer1.gzd <- lmer(gzd ~ cond + (1|item) + (1|subj), nd, REML=F)
mer1.gpd <- lmer(gpd ~ cond + (1|item) + (1|subj), nd, REML=F)
mer1.tvt <- lmer(tvt ~ cond + (1|item) + (1|subj), nd, REML=F)

p.ffd <- anova(mer0.ffd, mer1.ffd)[2,8]
p.gzd <- anova(mer0.gzd, mer1.gzd)[2,8]
p.gpd <- anova(mer0.gpd, mer1.gpd)[2,8]
p.tvt <- anova(mer0.tvt, mer1.tvt)[2,8]
```

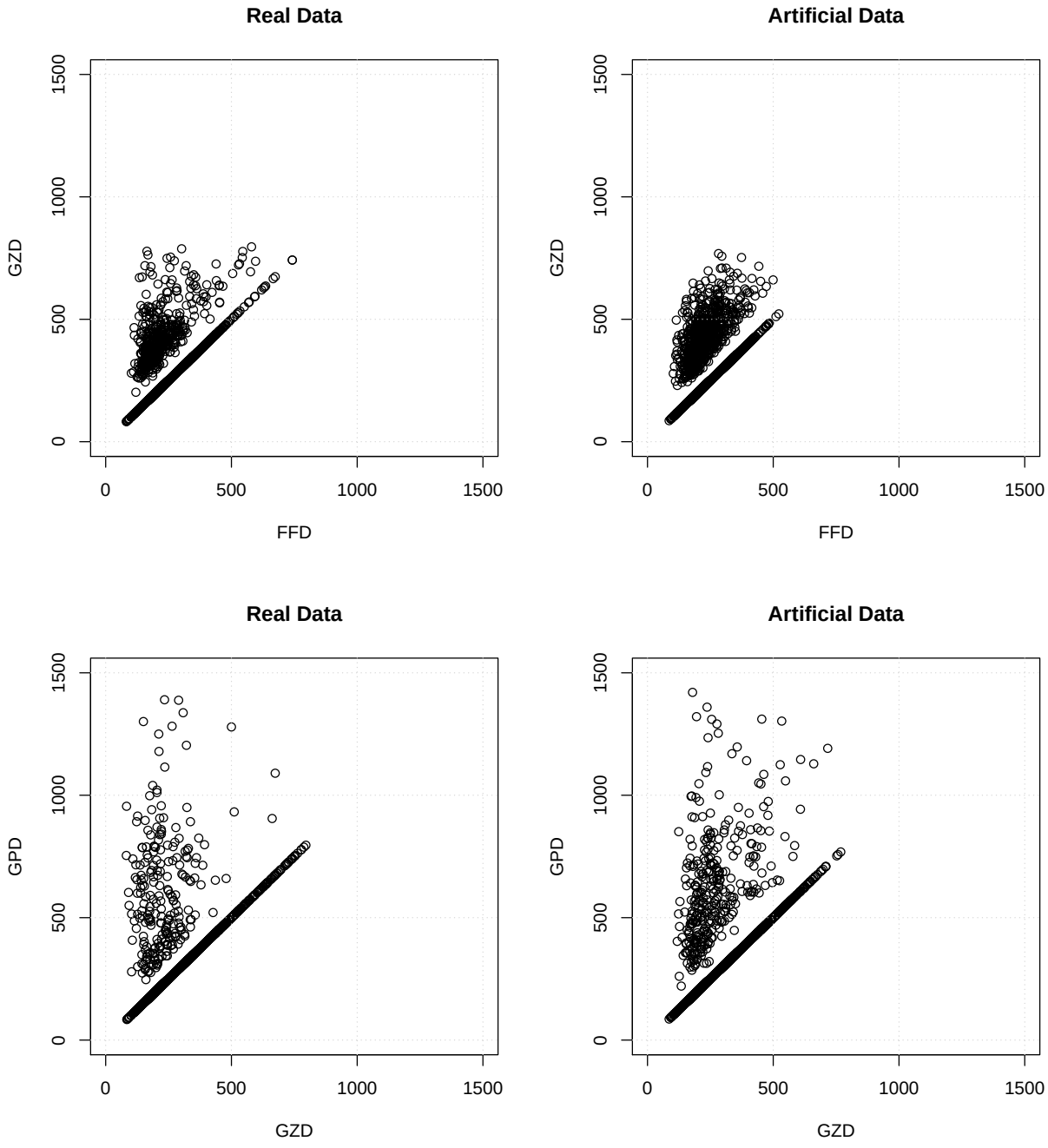

Figure 1: Gaze duration as a function of first fixation duration (first row) and go past time as a function of gaze duration (second row).

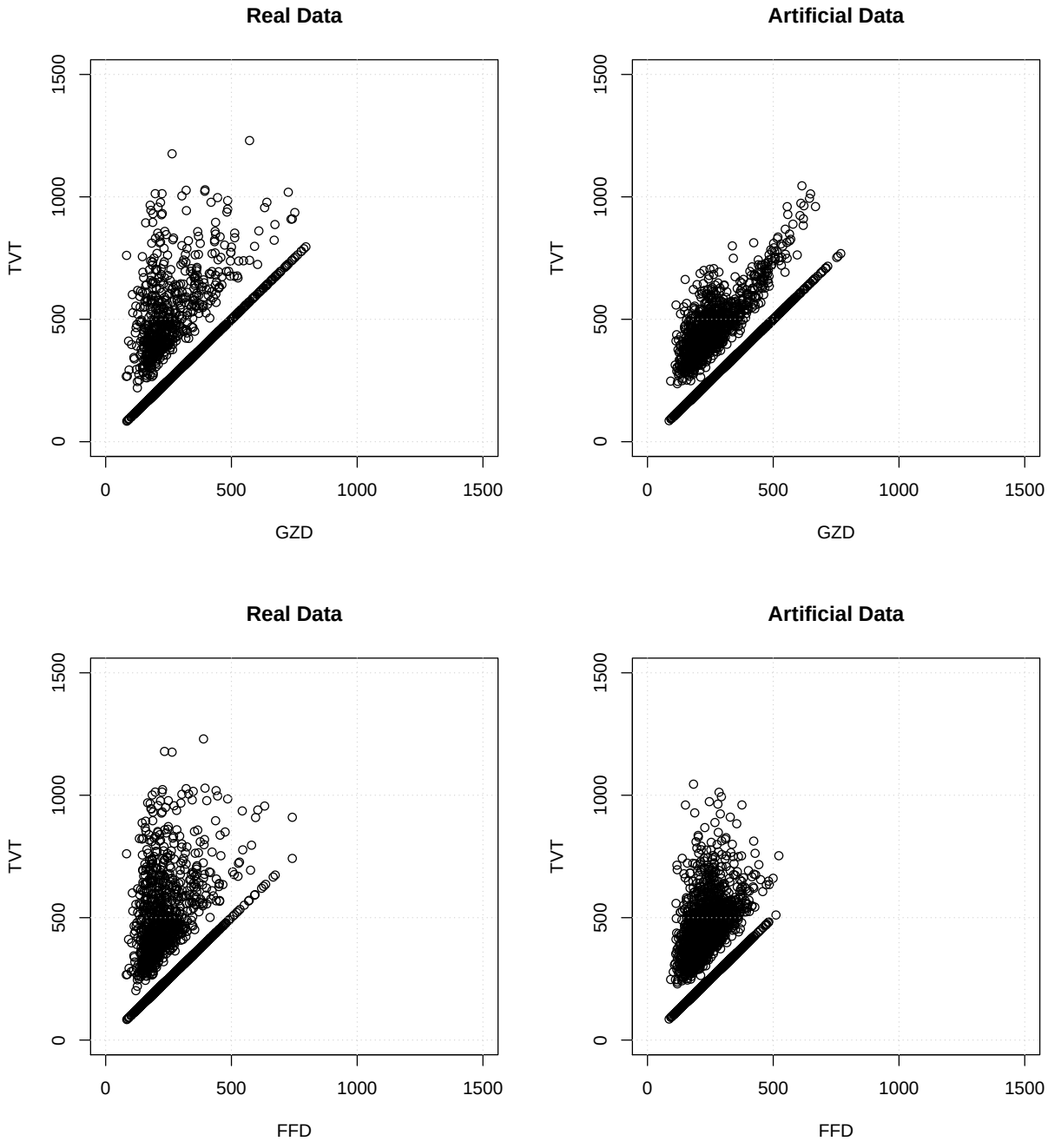

Figure 2: Total viewing time as a function of gaze duration (first row) and total viewing time as a function of first fixation duration (second row).

Criterion	Rate of false positives	95% Confidence interval
At least one effect with $p \leq 0.05$	0.121	[0.119, 0.123]
Bonferroni correction	0.032	[0.031, 0.033]
At least two effects with $p \leq 0.05$	0.042	[0.040, 0.043]

Table 3: Rates of false positives under the three decision criteria and binomial 95% confidence intervals

Results

The purpose of the simulation was to empirically investigate false positive rates and statistical power under the three decision criteria discussed in the introduction. Any significant effect of the hypothetical manipulation found in a data set that does not have a true effect is a false positive by definition. If a statistical test determines that the manipulation had a significant effect in such a data set, that result is a type I error due to random variation in the data. Thus, we simply counted the number of significant effects found under each decision criterion and checked whether the rate of false positives was higher or lower than the conventionally accepted 5%.

Similarly, we assessed statistical power using the data sets that contained a true effect. Using each of the three decision criteria we determined whether an effect was present or not in each data set and the rate of successfully detected effects then constituted our measure of power.

First, we present the false positive rates under the three decision criteria (see table 3). Applying the baseline criterion, which demands only one effect with $p \leq 0.05$, leads to a false positive rate of 12%. This rate is lower than the 18.5% we would expect if the tests were independent but it is also much higher than the conventionally accepted 5%. That shows that the issue of testing multiple dependent eyetracking measures is a very real concern. A failure to correct for multiple comparisons thus results in an unacceptably high risk that a result will not replicate.

As expected, the Bonferroni correction is an effective remedy against inflated false positives. Under the Bonferroni correction, we obtained false positives in 3% of the cases, which is also consistent with the suspicion that the Bonferroni correction might be more conservative than necessary. Requiring effects with $p \leq 0.05$ in at least two eyetracking measures (rule-of-thumb criterion), produced a false positive rate of %4 which is close to the desired 5%. A preliminary conclusion from these results is that the multiple comparison problem cannot be ignored and that both alternative decision criteria help to keep false positives under control.

Next, we look at the issue of statistical power. The goal of this was to determine which of the two alternative decision criteria performs better at detecting true effects in the data. As described above, we generated data sets with effect sizes of 2.5 ms, 5 ms, 10 ms, 20 ms, and 40 ms. Detecting effects in the data sets with effects of 2.5 ms was expected to be difficult with all three decision criteria, whereas effects of 40 ms were expected to be easy to detect.

Beyond that, we expected the three decision strategies to differ in how reliably they detect effects between these extremes.

Fig. 3 shows the detection rates for all three decision criteria as a function of effect size. For effect sizes greater than zero, the detection rate indicates the proportion of correctly detected true effects of the hypothetical manipulation. By *correctly detected* we mean that it was not enough that there was a significant effect, the effect also had to be in the correct direction. For effect sizes of zero, the detection rate is the rate of false positives (the data already reported above).

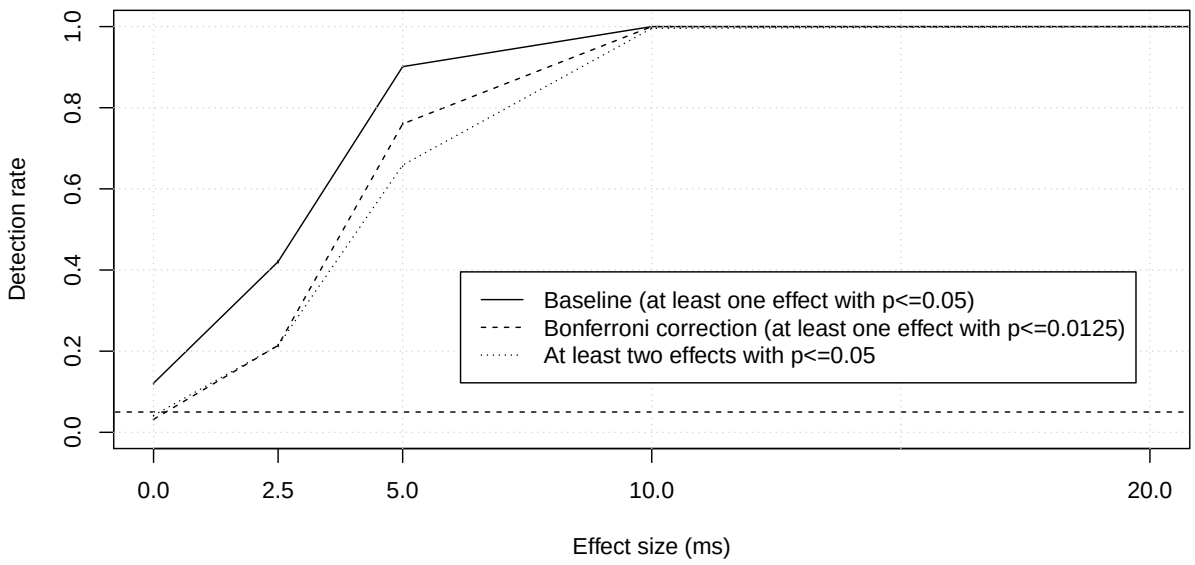

Figure 3: Detection rates under the three decision criteria. The dashed horizontal line indicates the conventionally accepted level of false positives (5%). Binomial confidence intervals (95%) are plotted but too small to be visible. At 0 ms the lines show the false positive rates obtained with each decision criterion.

All three decision criteria reliably detected effect with sizes of 10 ms and longer. This may seem surprising because 10 ms is typically considered to be a small effect which should be hard to detect. However, the study by Angele et al. (2013), after which our artificial data sets were modeled, was targeted at a small effect and therefore tested a high number of stimuli (132).

For true effect sizes of 2.5 ms and 5 ms, there were marked differences between the decision criteria. The baseline criterion, which requires an effect with $p \leq 0.05$ in just one of the four dependent measures, consistently had the highest power. However, from the previous analysis we know that this comes at the price of a greatly inflated probability of false positives. Among the two alternative criteria, the Bonferroni correction had better power than the rule-of-thumb criterion, which required at least two effects with $p \leq 0.05$.

Discussion

The purpose of the simulations reported above was to answer two questions. First, we wanted to determine the rate of false positive results obtained when four dependent eyetracking measures are tested for the effect of a hypothetical experimental manipulation. Second, we wanted to determine to what extent two alternative decision criteria are effective at keeping false positive positives at the desired level of 5% while preserving as much statistical power as possible.

Our results are quite clear: Declaring effects reliable when only one out of four dependent measures shows an effect with $p \leq 0.05$ is clearly anticonservative. A failure to address this issue may therefore considerably compromise the reproducibility of a result, and researchers and peer reviewers should not ignore this problem. Both alternative decision criteria did a good job at keeping the rate of false positives close to the conventionally accepted 5%. However, the Bonferroni delivered the better statistical power and is therefore preferred. In the following, we will discuss the alternative decision criteria in more detail.

Bonferroni correction

The decision criterion based on the Bonferroni correction was the most conservative and at the same time delivered high statistical power. Contrary to conventional wisdom, the Bonferroni correction therefore seems to be quite appropriate when multiple dependent eyetracking measures are tested. Since the Bonferroni correction was slightly too conservative, a researcher applying this correction may in some cases not find evidence for a true effect when a more precise correction may have allowed the detection of the effect. However, this type of error (type II, false negative) has much less potential to do harm than the greatly inflated false positive rate suffered without any form of correction. After all, researchers will typically draw strong conclusions when they believe to have found statistical evidence for an effect whereas a null-result (no statistical evidence for an effect) is inconclusive.

A researcher reporting an effect in just one dependent variable may argue that a Bonferroni correction is not needed because the effect was expected to appear in that dependent variable and not in others. If that were the case, the reader of such research may rightly ask why the researchers tested the other dependent variables in the first place. Even if the tested hypothesis predicts an effect in just one particular measure, it is often possible to produce a perfectly plausible explanation of why the effect could also appear in another dependent measure. Clearly, our general propensity to view an effect as entirely plausible after seeing the evidence for it does not help here. Thus, a researcher either has to demonstrate the a priori nature of the prediction (e.g., through pre-registration of the study design or through a very strong hypothesis that not only explains the presence of the effect in one measure, but also its absence in the other measures), correct for multiple testing, or conduct a replication of the experiment (see Gelman and Loken, 2013, for further discussion).

Rule-of-thumb criterion

The rule-of-thumb criterion evaluated in our study required that at least two dependent variables show an effect with a p-value below 0.05. This criterion has no rigorous mathematical foundation. It is rather a simple ad-hoc criterion which is intuitively and informally used by some researchers. The results of our simulation suggest, that this criterion performs fairly well: its false positive rate was closer to the desired 5% than that of obtained with the Bonferroni correction and its power was only slightly less than that of the Bonferroni criterion. Despite these encouraging results, there are a number of caveats suggesting that the more principled Bonferroni correction is clearly preferable.

One problem is that the rule-of-thumb criterion is not sensitive to the number of dependent variables that are tested. Requiring at least two effects with $p \leq 0.05$ is a strict criterion when only two dependent variables are tested but it is not sufficient when twenty dependent variables are tested. The latter situation is not uncommon in reading research, because researchers often evaluate a number of common eyetracking measures calculated for several regions of interest, e.g., the pre-target region, target, and post-target region, or, in experiments using the boundary paradigm, the pre-boundary region and the post-boundary region. Thus, it was a mere coincidence that the parameters of our study were such that the rule-of-thumb criterion performed so well.

Another problem with the rule-of-thumb criterion is that it is inappropriate in situations where an effect can only appear in one measure. For example, if a manipulation leads to measurable differences in processing only in very late stages, its effects may appear only in total reading times. Requiring this effect to also emerge in an early measure would make it impossible to detect it.

Finally, the suitability of the rule-of-thumb criterion also depends on statistical properties of the dependent measures such as their mutual correlations and these properties change from experiment to experiment. Although the rule-of-thumb criterion performed reasonably well in the present study, it is clear that it will be much less appropriate in other settings. Different situations therefore require different rules of thumb, and determining in which situation which rule should be applied is not trivial. While some experienced researchers might be capable of applying it intuitively based on their experience with eye movement data, junior researchers are almost certainly not. For these reasons, we advise against using the rule-of-thumb criterion. The Bonferroni correction is preferable because it is more principled and easy to use.

The role of replication

One possible objection is that replication is a better method to establish whether an effect is true or not than applying corrections for multiple comparisons. We fully agree that the replication of initial positive findings is extremely important. McElreath and Smaldino (2015) show that the veracity of many true hypotheses can only be determined through several iterations of replications and that replications also play a crucial role in weeding out false positives which otherwise would clog up the scientific discourse. However, correction for

multiple comparisons and replication are not mutually exclusive. In fact, the opposite is the case. A researcher considering to attempt the replication of a result should not base this decision on an uncorrected analysis because the results of that analysis are misleading. Consider a hypothetical study in which 5 measures were tested in three regions of interest. This means that overall 15 statistical tests have to be conducted to test for effects of a manipulation. Further assume that one of these tests shows a significant effect with $p = 0.04$ while the other tests produce p-values higher than 0.05. In this situation, there is a relatively high chance that the effect will not replicate. However, the very same effect would be more likely to replicate if it had been obtained in a study that tested only two dependent measures in one region of interest. This shows that the p-value of an effect alone is a poor foundation for deciding whether a replication study is worth the investment of time and money. A decision based on the p-value in relation to the corrected alpha threshold is going to be more informative about the expected utility of a replication study. Thus, corrections for multiple comparisons should not be seen as an alternative to replication, they should rather be used to amplify the effectiveness of replication studies.

Conclusions

The problem of false positives due to multiple comparisons may seem like an old chestnut and should be known to anyone who has taken an introductory class in statistics. However, the fact that there are virtually no eyetracking studies of reading behavior (including those by the authors) formally addressing this problem shows that established real-world research practices can easily override our basic statistical knowledge. Our simulation results challenge the conventional wisdom by demonstrating that false positive rates are a serious problem: Even when only four dependent measures are tested, a failure to deal with multiple testing can lead a researcher to declare effects as significant that should not even qualify as marginally significant. The problem is even worse when eyetracking measures in several regions of interest are tested because then the dependent variables and statistical tests multiply. Hence, the failure to address multiple testing may considerably compromise the reproducibility of published results. Experienced researchers may deal with this issue by applying rules of thumb, but such informal criteria are not consistent enough to protect the scientific record from excessive rates of false positive results. Similarly, while posing strong hypotheses as to which measure an effect should appear in and testing only those dependent variables for which a strong prediction exists would be a perfectly acceptable alternative to using a Bonferroni correction, but few studies have hypotheses that are explicit enough for this. Additionally, if a certain test is not reported, peer reviewers might request it. In summary, our findings result in a clear recommendation because they show that the Bonferroni correction is an effective remedy while only moderately reducing statistical power and requiring little subjective judgment on the part of the researcher. Authors, reviewers, and journal editors should be aware of the issues associated with multiple dependent variables and act accordingly.

Acknowledgments: We dedicate this article to the memory of Keith Rayner who commented on an early version of this work. We also thank Shravan Vasishth and the members of the Rayner Lab for insightful comments on an earlier version of this paper. Titus von der

Malsburg was funded by a Feodor Lynen Research Fellowship awarded by the Alexander von Humboldt Foundation and by NIH grant HD065829 awarded to Roger Levy and Keith Rayner.

References

Angele, B., Tran, R., and Rayner, K. (2013). Parafoveal-foveal overlap can facilitate ongoing word identification during reading: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 39(2):526–538.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Bates, D., Maechler, M., Bolker, B. M., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. ArXiv e-print; submitted to *Journal of Statistical Software*.

Engbert, R., Nuthmann, A., Richter, E. M., and Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777–813.

Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Unpublished manuscript, version of Nov 14th 2013.

McElreath, R. and Smaldino, P. E. (2015). Replication, communication, and the population dynamics of scientific discovery. ArXiv e-print.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

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

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366.

Vasishth, S., von der Malsburg, T., and Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(2):125–134.