Journal of Memory and Language 94 (2017) 119–133

Contents lists available at ScienceDirect

Journal of Memory and Language

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

False positives and other statistical errors in standard analyses of eye movements in reading ☆

Titus von der Malsburg [a,*], Bernhard Angele [b]

[a] Department of Linguistics, University of Potsdam, Germany
[b] Department of Psychology, Bournemouth University, United Kingdom

ARTICLE INFO

Article history:
Received 28 January 2016
revision received 10 October 2016

Keywords:
Statistics
False positives
Null-hypothesis testing
Eye-tracking
Reading
Sentence processing

ABSTRACT

In research on eye movements in reading, it is common to analyze a number of canonical dependent measures 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 appears to be 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 presented to justify this assumption. Here, we report a computational investigation of this issue using Monte Carlo simulations. Our results show that, contrary to conventional wisdom, false positives are increased to unacceptable levels when no corrections are applied. Our simulations also show that counter-measures like the Bonferroni correction keep false positives in check while reducing statistical power only moderately. Hence, there is little reason why such corrections should not be made a standard requirement. Further, we discuss three statistical illusions that can arise when statistical power is low, and we show how power can be improved to prevent these illusions. In sum, our work renders a detailed picture of the various types of statistical errors than can occur in studies of reading behavior and we provide concrete guidance about how these errors can be avoided.

Introduction

A key advantage of using eye-tracking 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

☆ We dedicate this article to the memory of Keith Rayner who commented on an early version of this work. We also thank Shravan Vasishth, the members of the Rayner Lab, and audience members at the CUNY, AMLaP, and ECEM conferences for insightful comments. The code and simulation results are available at https://github.com/tmalsburg/MalsburgAngeleJML. This research was conducted using free software published under the GNU General Public License, particularly the R system for statistical computing, the Slurm Workload Manager, GNU Emacs, the TeX Live typesetting system, the knitr package for literate programming, and GNU/Linux. 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.

* Corresponding author at: Department of Linguistics, University of Potsdam, Germany.
E-mail address: malsburg@uni-potsdam.de (T. von der Malsburg).

http://dx.doi.org/10.1016/j.jml.2016.10.003
0749-596X/

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, von der Malsburg, & Engelmann, 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 durations 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 fixations 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 varying degrees. These correlations, combined with the fact that several of these measures are potentially informative, lead to a problematic situation.

Let's assume that a researcher wants to test whether a certain experimental manipulation has an impact on reading behavior. They 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 additional information about the time course of the affected 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 positive result with a probability of 5% even if there was no true effect. Assuming further that four statistically independent eye-tracking measures are tested, the probability that at least one of these tests produces a false positive result increases to $1 - 0.95^4 = 0.185$. This means that the probability of finding a false positive result would be 18.5% instead of the conventionally accepted 5%.

Of course, in reality, eye-tracking measures are not quite independent as the different fixation time measures are typically highly correlated. For example, the correlation between first fixation duration (FFD) and gaze duration (GZD) is determined by the rate at which readers refixate. If readers do not refixate at all, FFD and GZD are identical, in other words, their correlation is 1. The more readers refixate, the lower becomes 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 18.5% calculated above. We suspect that this is one reason for 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 four dependent measures lies somewhere between 5% and 18.5%.

If the rate of false positives is inflated to unacceptable levels due to multiple comparisons, adjustments to the decision making process are needed. The textbook solution for that is the Bonferroni correction (Bonferroni, 1936). All a researcher has to do in order to apply this correction is to divide the α 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 $\alpha = 0.05/4 = 0.0125$. 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%. We then say that the family-wise error rate is controlled, the family being the set of tested hypotheses: there is an effect in FFD, there is an effect in GZD, and so on.

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, they are typically not. The Bonferroni correction may therefore be too conservative, which means that the false positive rate will be even lower than 5%. While this is not in itself a problem, any reduction of false positives also comes at the price of reduced statistical power, i.e. a reduced ability to detect true effects. Hence, it is desirable to reduce false positives to the conventional 5% but not more because we would otherwise sacrifice more statistical power than necessary in order to control the family-wise error rate.

Concerns about loss of statistical power may therefore be another reason for why reading researchers do not apply corrections for multiple comparisons such as the Bonferroni