False-positive rates in eyetracking studies with multiple dependent measures

Titus von der Malsburg (UC San Diego) & Bernhard Angele (Bournemouth University)
malsburg@ucsd.edu

A key advantage of using eyetracking in psycholinguistic research on reading behavior is the wealth of data that can be collected. This data is typically summarized using a range of conventional measures that furnish a detailed picture of how the reading process unfolds. Unfortunately, the use of these measures also confronts us with a non-trivial problem: Analyzing multiple measures requires running multiple statistical tests, which leads to an inflated rate of false positives, i.e., an increased risk that an effect is declared significant even though it was just a random fluke. Although this instance of the multiple-testing problem is widely acknowledged, surprisingly little is typically done to address it, which may in part be due to a perceived lack of a satisfactory correction method. The Bonferroni correction was designed to address this problem but it is believed to be too aggressive especially when the dependent variables are highly correlated, as is the case with eyetracking measures. In this situation many researchers resort to "soft" criteria for evaluating the robustness of an effect. For example, an effect may be declared robust if it is subjectively consistent with previous research, or if it reaches significance in more than one measure. Whether such rules of thumb are appropriate is far from clear and has, to our knowledge, not been tested.

Here, our goal was to investigate how bad the problem of inflated false positives really is and how it can be addressed. We answered these questions by empirically testing false-positive rates through Monte Carlo simulations. The general approach is similar to that taken by Barr, Levy, Scheepers, Tily (JML, 2013): We simulated a 100,000 artificial data sets with properties resembling those found in a real reading experiment and calculated four eyetracking measures: first fixation duration, gaze duration, go-past time, and total viewing time. For the present purpose, it is important that these measures exhibit the statistical dependencies found in real eyetracking data. Therefore, we generated the data using a simple generative model of eye movements. Although this model is much simpler than established models of eye movement control such as E-Z Reader and SWIFT, our evaluation shows that it faithfully reproduces crucial statistical properties. Since these data sets were generated to have no effect of the hypothetical manipulation, we know that any effects reaching significance must be false positives. Significance was determined using linear mixed models, likelihood ratio tests, and the conventional criterion $p \leq 0.05$.

Results: The chance that a significant effect was found in at least one of the four measures was 12%, far exceeding the desired 5%. This shows that the risk of falsely rejecting the null hypothesis is indeed greatly inflated when testing four eyetracking measures instead of just one. We also tested a simple rule of thumb that some researchers use: When we required significant effects in at least two measures in order to reject the null hypothesis, the rate of false positives was lowered to 4%. While this rule generated an acceptable false-positive rate, its appropriateness depends on statistical properties of the data that vary from study to study; it should therefore be used with caution. When we applied the Bonferroni correction, the overall false-positive rate was at 3%. As expected, this is too conservative, however, much less so than is often claimed. In sum, these results show that inflated false-positive rates are a serious concern even when only four dependent eyetracking measures are tested. This problem is exacerbated when eyetracking measures are tested in several regions of interest (e.g., the pre-target, target, and spill-over region) because then the tested dependent variables multiply. A failure to properly address multiple testing may therefore considerably compromise the reproducibility of a result obtained using eyetracking. Contrary to conventional wisdom, the Bonferroni correction seems to be an appropriate remedy for this problem.

False-Positive Rates in Eyetracking Studies With Multiple Dependent Measures

Titus von der Malsburg, UC San Diego
Bernhard Angele, Bournemouth University

Background:

- Standard measures analyzed in eyetracking experiments investigating reading behavior: Skipping rate, single-fixation duration, first fixation duration, gaze duration, regression rate, go-past time, re-reading time, total viewing time.
- It is common to calculate and analyze some or all of these measures for one or several regions of interest.

⇒ Increased rate of false positives due to multiple testing.

Correcting for multiple comparisons to control false positives is almost unheard of in the literature but no formal justification has ever been given.

Textbook solution: Bonferroni correction, but independence of tests is violated.

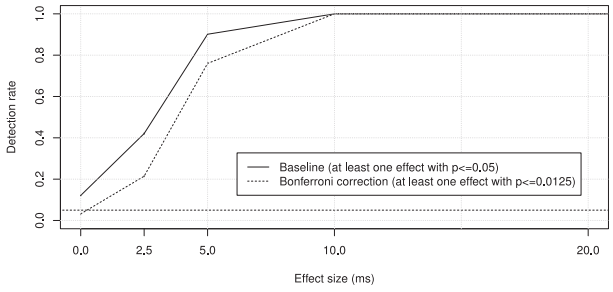
Approach: Monte Carlo Simulations

1. Generate 100K artificial data sets with properties resembling those of typical eyetracking data sets and a hypothetical experimental manipulation with two conditions.
2. Create versions of each data set with true effect sizes of the manipulation of: 0 ms, 2.5 ms, 5 ms, 10 ms, 20 ms, and 40 ms.
4. Analyze four standard measures in each data set: first fixation duration, gaze duration, go-past time, total viewing time.
5. For each effect size: Count data sets with
   - one or more effects with p≤0.05 (baseline, no correction),
   - one or more effects with p≤0.0125 (Bonferroni correction).

Any significant effect in a data set with an effect size of 0 ms is a false positive by definition. In data sets with true effect sizes > 0 ms, the rate of correctly detected effects is our measure of power.

Results:

Criterion	False positives	95% CIs
No correction (p≤0.05):	12.1%	11.9% – 12.3%
Bonferroni correction (p≤0.0125):	3.2%	3.1% – 3.3%



Discussion:

- Without correction and only four measures, false positives are increased to 12%, which is much higher than the accepted 5%.
- This rate will be even higher when more dependent measures and more regions of interest are tested.
- The Bonferroni correction is only slightly too conservative.
- The reduction in power due to the Bonferroni correction was less than 20%.

Bottom line: The widespread practice of not correcting for multiple comparisons in reading studies is unjustified. Despite the violation of the independence assumption, the Bonferroni correction is a suitable method for keeping false positives under control.

Questions:

- By how much are false positives increased?
- Is the Bonferroni correction appropriate or is it much too conservative as is often claimed?
- By how much is statistial power reduced by applying the Bonferroni correction?

Parameters Used for Generation of Artificial Data:

Name	Value	Name	Value
# subjects	40	σ subjects	22
# items	132	σ items	21
P(refix)	0.14		
P(regr)	0.07		
P(reread)	0.197		
μ ffd	221	σ ffd	1.3
μ gzd-ffd	194	σ gzd-ffd	1.4
μ gpd-gzd	238	σ gpd-gzd	1.8
μ tvt-gzd	236	σ tvt-gzd	1.5

Means (μ) and standard deviations (σ) are geometric.

Real vs. Artificial Eyetracking Measures: