False Positives in Bayesian Analyses of Eye Movements in Reading: The Impact of Regularizing Priors and Credible Interval Width

Titus von der Malsburg[1], Bernhard Angele[2]

[1]Institute of Linguistics, University of Stuttgart, [2]Nebrija Research Center in Cognition, Faculty of Language and Education, Nebrija University

Recent years have seen a shift to Bayesian statistical methods in psycholinguistic research, driven by computational advances and new software, particularly Stan [5] and `brms` [1]. These tools offer increased flexibility in tailoring statistical models to research problems, the ability to fit more complex models, and the capacity to quantify uncertainty and integrate prior knowledge [3, 4]. Nonetheless, when used to test hypotheses, Bayesian models remain susceptible to fundamental problems like spurious effects (Type I errors) and the failure to detect true effects (Type II errors). For the frequentist framework, von der Malsburg and Angele [6] showed that false positive rates are especially inflated when a single hypothesis is tested by evaluating multiple dependent variables, as is common in reading research. The present work investigates how multiple comparisons [2] affect Bayesian inferences and evaluates potential mitigation strategies.

Method: Following von der Malsburg and Angele [6], we simulated synthetic eye movement data sets with four dependent measures (first fixation duration, gaze duration, go-past time, total reading time, all log-transformed) containing either no effect of a hypothetical manipulation or effects of varying sizes (2.5, 5, 10, 20, 40, 80 ms). We then examined Type I and Type II error rates under multiple decision criteria; here we focus on credible interval width and the strength of regularizing priors. In the primary scenario of interest, an effect was claimed if at least one of the four eye-tracking measures showed a robust difference – a common practice in reading research.

Results: 1. When using 95% credible intervals as the decision criterion and requiring at least one measure to show a robust effect, the Type I error rate was 17.1% (Table 1, Fig. 1), slightly higher than the rate reported by von der Malsburg and Angele [6] for analogous frequentist analyses using `lme4`. 2. Increasing the credible interval width to 97.5% reduced the Type I error rate to 8.6%, and 99% reduced it to 2.5%. 3. Weaker regularizing priors (flat, N(0,1), N(0,0.1)) had negligible effects on Type I error rates, whereas stronger priors (N(0,0.01) and N(0,0.001)) reduced the detection rate to near zero regardless of the decision criterion (Fig. 2). 4. In data sets with true effects, weaker priors did not affect true positive rates, but stronger priors severely compromised statistical power, reducing true positive rates to near zero for all effect sizes and decision criteria.

Discussion: False positives are as problematic in Bayesian analyses as in frequentist analyses; however, increasing the credible interval width is an effective strategy to mitigate false positives in the tested scenario. While priors can theoretically also reduce Type I errors to the desired rate, calibrating these priors sufficiently is near impossible; in practice, these priors either fail to control the Type I error rate or they suppress it at the cost of severely compromising the test's ability to detect true effects.

Conclusion: While Bayesian statistical methods offer substantial benefits, our results demonstrate that they do not automatically resolve foundational problems like Type I and Type II errors. Researchers analyzing multiple dependent measures must implement safeguards to address these issues. In addition, pre-registration and more specific hypotheses can mitigate the multiple comparisons problem regardless of the statistical framework.

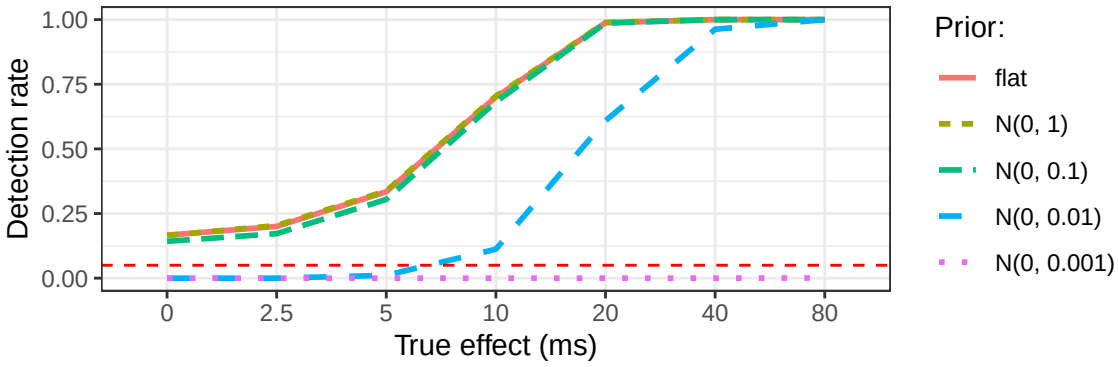

Figure 1: Detection rate (at least one measure, correct direction) as a function of effect size and prior. For effect size zero, the graph shows the detection rate for effects in either direction. Red dashed horizontal line shows the conventional α-level of 0.05.

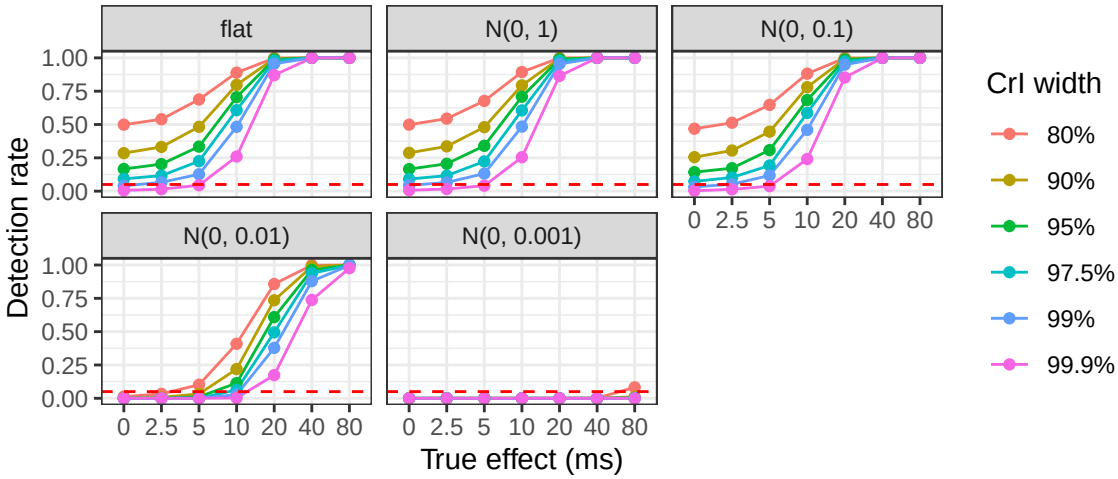

Figure 2: Effect of different credible interval sizes on family-wise detection rate (at least one measure, correct direction). For effect size zero, the graph shows the detection rate for effects in either direction. Red dashed horizontal line shows the conventional α-level of 0.05.

Table 1: Proportion of false positives (family-wise) for different prior distributions and Credible Interval widths at 0 ms effect size.

CrI width	Prior: flat	N(0, 1)	N(0, 0.1)	N(0, 0.01)	N(0, 0.001)
80%	49.9%	49.9%	46.8%	1.3%	0.0%
90%	28.5%	28.8%	25.5%	0.1%	0.0%
95%	16.6%	16.5%	14.2%	0.0%	0.0%
97.5%	9.2%	9.1%	7.3%	0.0%	0.0%
99%	4.0%	4.3%	3.2%	0.0%	0.0%
99.9%	0.6%	0.7%	0.3%	0.0%	0.0%

Reference: [1] Bürkner, 2017, JSS; [2] Gelman, Loken, 2013, unpubl. ms.; [3] Nicenboim, Vasishth, 2016, LLC; [4] Schad, Betancourt, Vasishth, 2021, Psychol. Methods; [5] Stan Development Team, 2017; [6] Malsburg, Angele, 2017, JML.