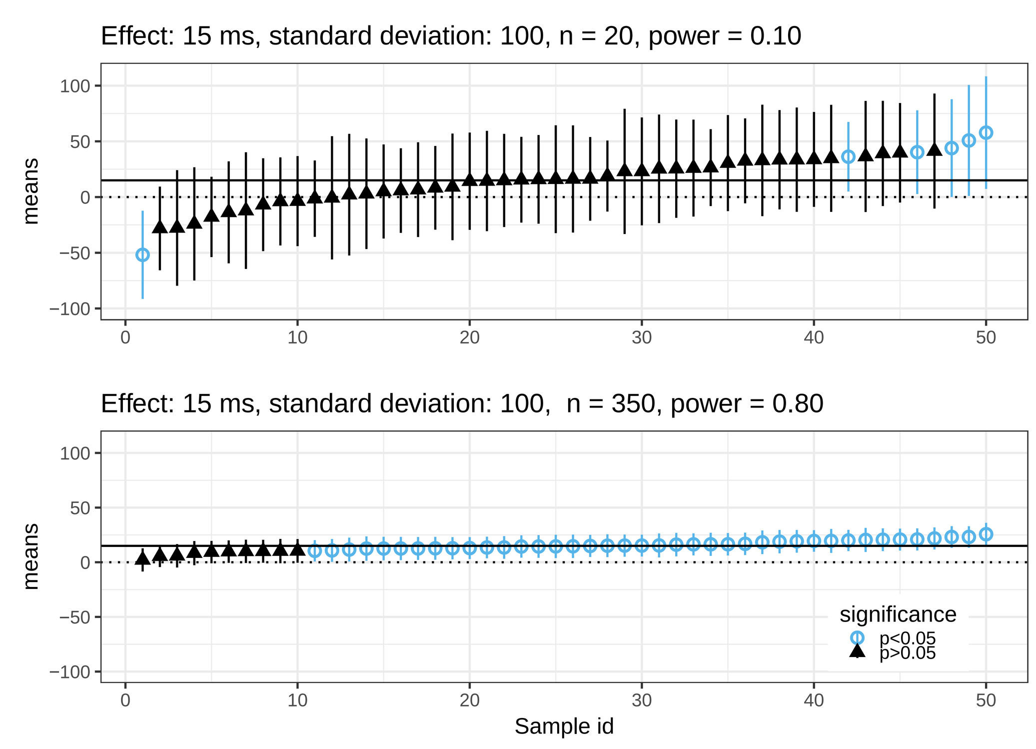
TYPE M ERRORS IN PRACTICE: A CASE STUDY

Shravan Vasishth[1], Daniela Mertzen[1], Lena A. Jäger[1] & Andrew Gelman[2]
[1]University of Potsdam, [2]Columbia University
vasishth@uni-potsdam.de

1. Motivation

- Power is relatively low in psycholinguistic studies.
  - E.g., in Jäger et al., 2017 (Appendix B): for effects (reading studies) ranging from -16 to -41 ms (sd = 150 ms, N = 40), power estimates ranged from 15% to 45%.
- Low power leads to exaggerated estimates
- Published claims will not be replicable
- Our paper (*Journal of Memory & Language*, in press) demonstrates this through direct replication of a published result (Levy & Keller, 2013).

2. The Problem: Demonstration of Type M error (simulated data)



If the estimated effect is statistically significant given that the true effect is not 0, under repeated sampling, low power leads to:

(i) **Type M** (= *magnitude*) error, i.e. an *overestimation* of the effect

(ii) **Type S** (= *sign*) error, i.e. effect in the *wrong direction*

(Gelman & Carlin, 2014)

When power is high, significant and non-significant effects will be tightly clustered near the true mean.

3. Investigating Type M error in published data: Levy & Keller, 2013

Levy & Keller (2013) study:
Two eye-tracking reading experiments: each had 28 subjects and 24 items presented in a Latin Square

Design: 2 × 2 repeated measures fully-crossed factorial design

- Two main effects and one interaction

Dependent measure:

- reading time in milliseconds (*rt*)

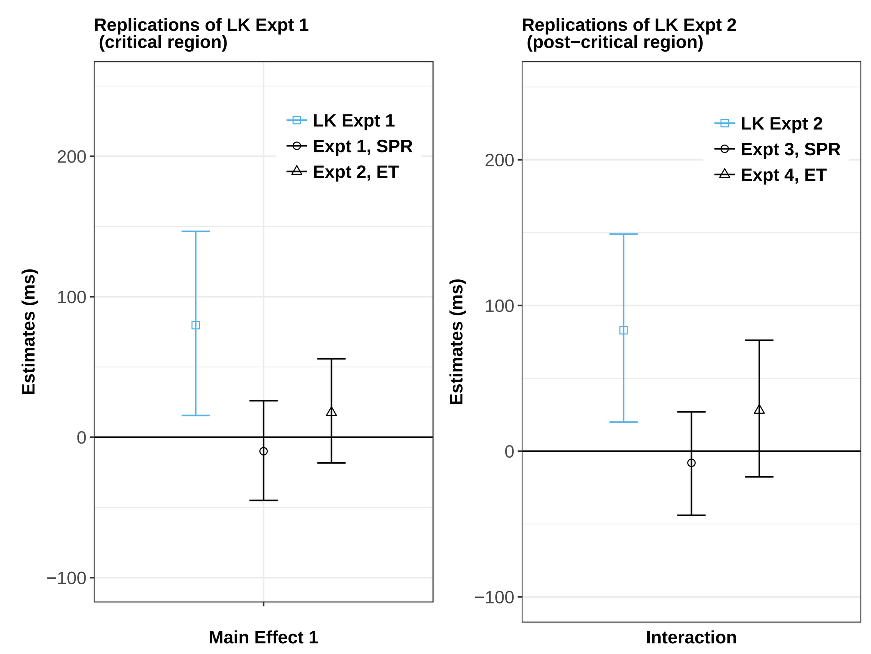
We conducted seven replication attempts of Levy & Keller, 2013

Our Replication Attempts:

Our Expt	Original Expt	Subj	Items
Expt 1 (SPR)	LK 1	28	24
Expt 2 (ET)	LK 1	28	24
Expt 3 (SPR)	LK 2	28	24
Expt 4 (ET)	LK 2	28	24
Expt 5 (SPR)	LK 1, 2 (c,d)	28	24
Expt 6 (ET)	LK 1, 2 (c,d)	28	24
Expt 7 (ET)	LK 1, 2 (c,d)	100	24

ET: eye-tracking while reading; SPR: self-paced reading

4. Results of our Expts 1–4



Posterior means with 95% credible intervals computed from a Bayesian maximal linear mixed model using Stan. Shown are mean reading time at the critical or at the post-critical region of the original studies vs. our replication attempts.

5. Hierarchical linear mixed models in Stan

$i = 1,\ldots,I$ subjects; $j = 1,\ldots,J$ items; n data points; p predictors

$$log(rt) = \underbrace{X\beta}_{\text{fixed effects}} + \underbrace{Z_u b_u}_{\text{subjects random effects}} + \underbrace{Z_w b_w}_{\text{items random effects}} + \epsilon$$

$$X_{n\times p} = \begin{bmatrix} 1 & -1 & -1 & +1 \\ 1 & +1 & +1 & +1 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \qquad \beta_{p\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$= Z_u = Z_w$$

β_1 = Main Effect 1; β_2 = Main Effect 2; β_3 = Interaction

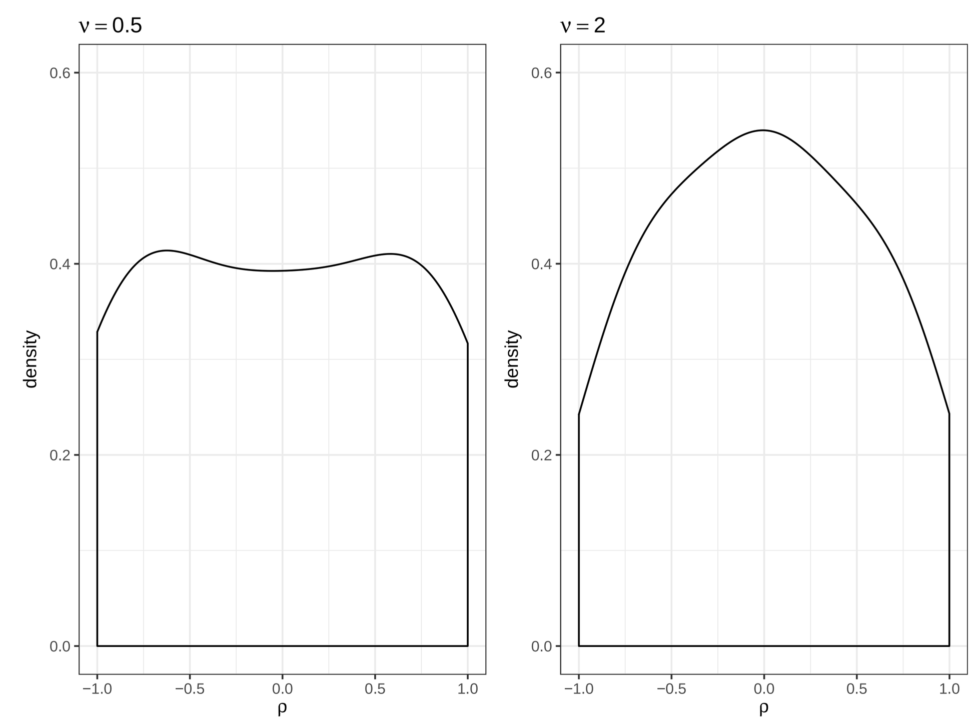
Priors

$$\beta_0 \sim Normal(0, 10)$$
$$\beta_{1,2,3} \sim Normal(0, 1)$$
$$\sigma \sim Normal_+(0, 1)$$
$$\rho \sim LKJ(\nu = 2)$$

$$b_u = MVN_4(0, \Sigma_u)$$
$$b_w = MVN_4(0, \Sigma_w)$$
$$\epsilon = Normal(0, \sigma)$$

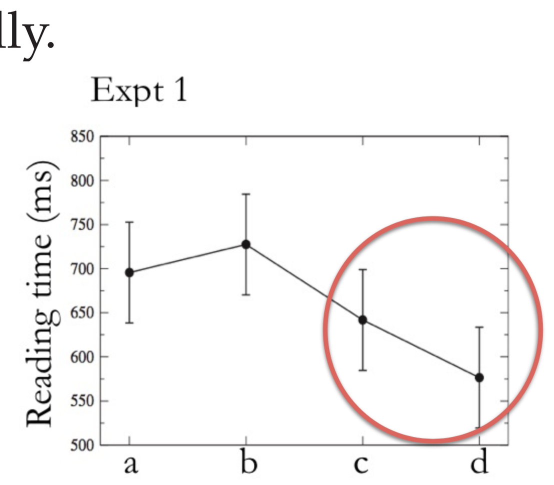
6. LKJ Prior

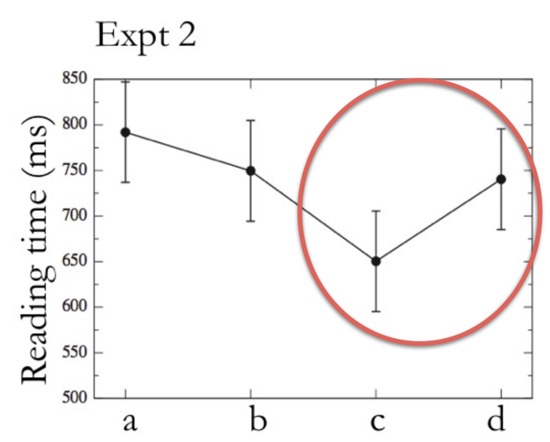
$$\rho \sim LKJ(\nu = 2)$$



7. Why run Expts 5–7?

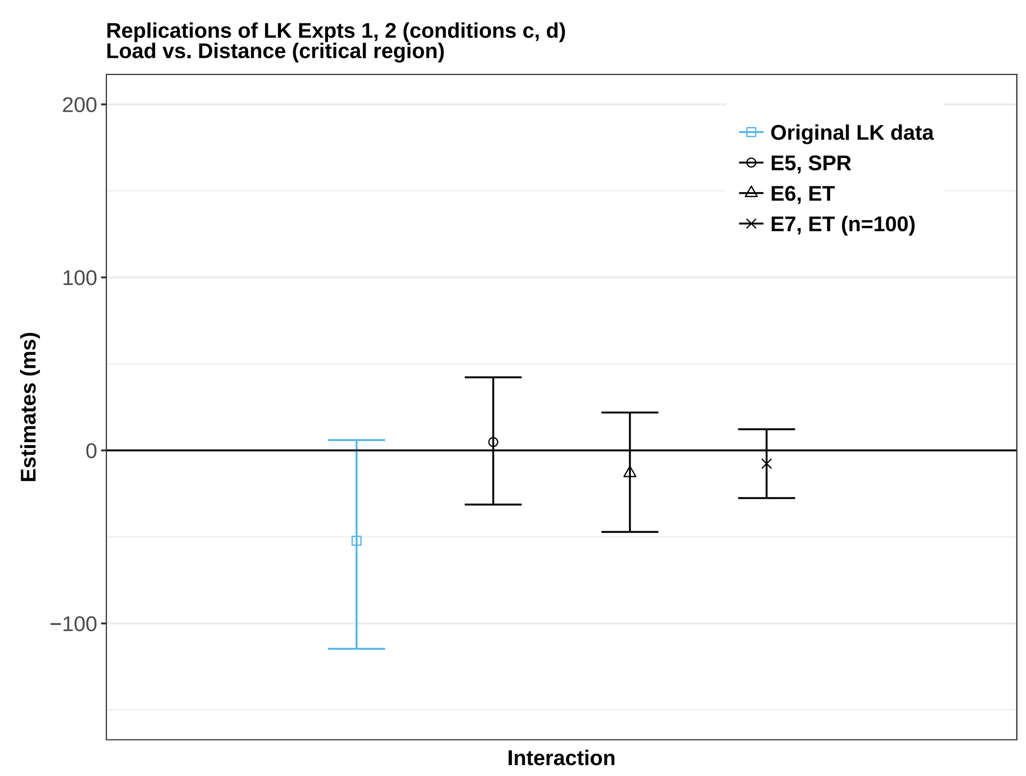
Levy & Keller (2013) claimed an interaction across their two experiments but never checked it statistically.





We tested this formally (see our Expts 5–7).

8. Results of our Expts 5–7



Posterior means with 95% credible intervals computed from a Bayesian maximal linear mixed model using Stan. Shown are mean reading time at the critical region of the original studies vs. our replication attempts.

9. Conclusion

Seven replication attempts found no evidence of the effects found in the original study.

Low statistical power + noisy estimates + flexible multiple comparisons $\Longrightarrow$ many published, 'significant' findings are the result of an overestimation (**Type M error**).

10. Improving current practices

OUR PROPOSAL:

- Move focus away from statistical significance
- Focus on estimation: run high-precision experiments
- Conduct direct replications to establish robustness of effect
- Pre-register hypotheses, design and analyses plan of study