The Statistical Significance Filter leads to overoptimistic expectations of replicability

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

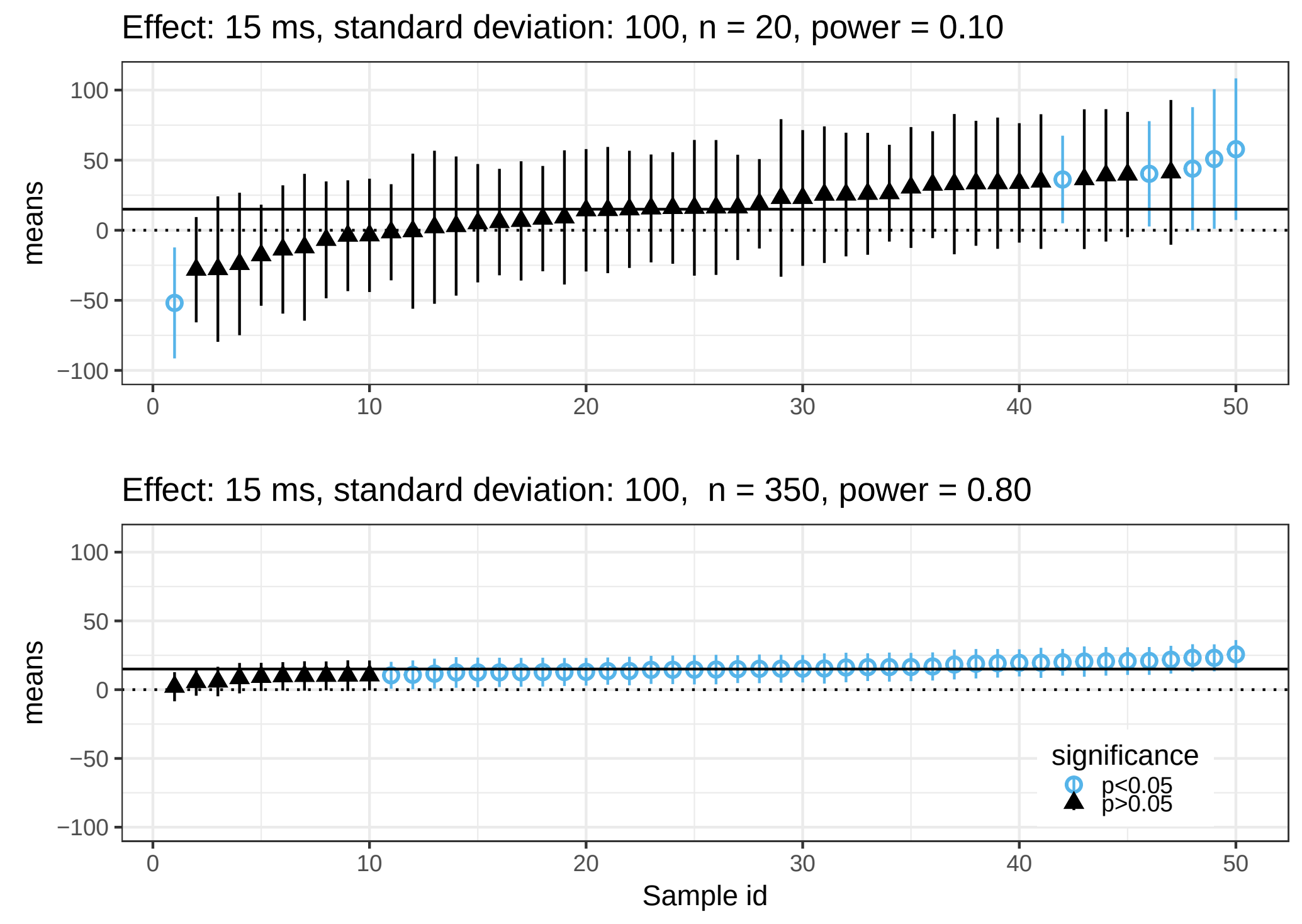
1. Motivation

Statistical significance filter: p<0.05 decision criterion for publication-worthiness.

Failure to replicate published work, e.g.,
Nieuwland et al. (2018)
Kochari and Flecken (2018)

GOAL OF OUR PAPER:
(*Journal of Memory & Language*, in press)
Demonstrate through direct replication of a published, plausible result (Levy & Keller, 2013) that the statistical significance filter leads to findings that are positively biased.

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. Design & Materials of Levy & Keller, 2013 (LK13)

LK13 study: Two eye-tracking experiments (28 subjects, 24 items each) investigating locality & anti-locality effects in German

Design: 2 × 2 fully-crossed factorial design

- Factor 1: Position of Dative NP (**DAT**) (main- vs. subordinate clause)
- Factor 2: Position of PP Adjunct (**ADJ**) (main- vs. subordinate clause)

LK Expt 1: target construction is in main clause
LK Expt 2: target construction is embedded in a relative clause → higher syntactic complexity

Example item:

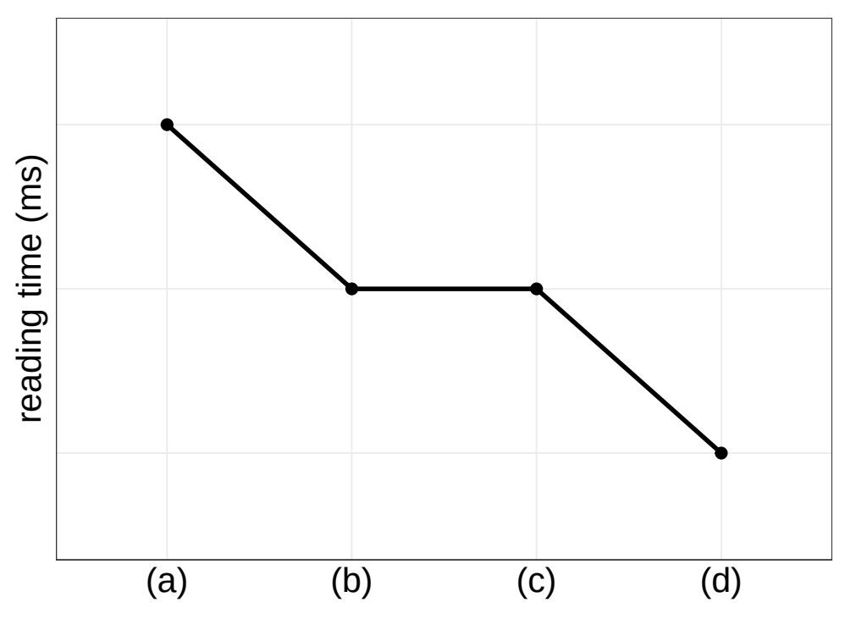
									critical	post-critical
a. PP Adjunct in subordinate clause, Dative NP in subordinate clause										
Nachdem	der	Lehrer	[**ADJ** zur Ahndung]	[**DAT** dem Sohn]	..., hat Hans Gerstner ...			den Fußball	**versteckt**,	und somit...
After	*the*	*teacher*	[**ADJ** *as payback*]	[**DAT** *the son*]	..., *has Hans Gerstner ...*			*the football*	*hid*,	*and thus...*
b. PP Adjunct in main clause, Dative NP in subordinate clause										
Nachdem	der	Lehrer		[**DAT** dem Sohn]	..., hat Hans Gerstner ...	[**ADJ** zur Ahndung]		den Fußball	**versteckt**,	und somit...
After	*the*	*teacher*		[**DAT** *the son*]	..., *has Hans Gerstner ...*	[**ADJ** *as payback*]		*the football*	*hid*,	*and thus...*
c. PP Adjunct in subordinate clause, Dative NP in main clause										
Nachdem	der	Lehrer	[**ADJ** zur Ahndung]		..., hat Hans Gerstner ...		[**DAT** dem Sohn]	den Fußball	**versteckt**,	und somit...
After	*the*	*teacher*	[**ADJ** *as payback*]		..., *has Hans Gerstner ...*		[**DAT** *the son*]	*the football*	*hid*,	*and thus...*
d. PP Adjunct in main clause, Dative NP in main clause										
Nachdem	der	Lehrer			..., hat Hans Gerstner ...	[**ADJ** zur Ahndung]	[**DAT** dem Sohn]	den Fußball	**versteckt**,	und somit...
After	*the*	*teacher*			..., *has Hans Gerstner ...*	[**ADJ** *as payback*]	[**DAT** *the son*]	*the football*	*hid*,	*and thus...*

'After the teacher imposed detention classes, Hans Gerstner hid the football from the naughty son of the industrious janitor as additional payback for the multiple wrongdoings, and thus corrected the affair.'

4. LK13 Predictions

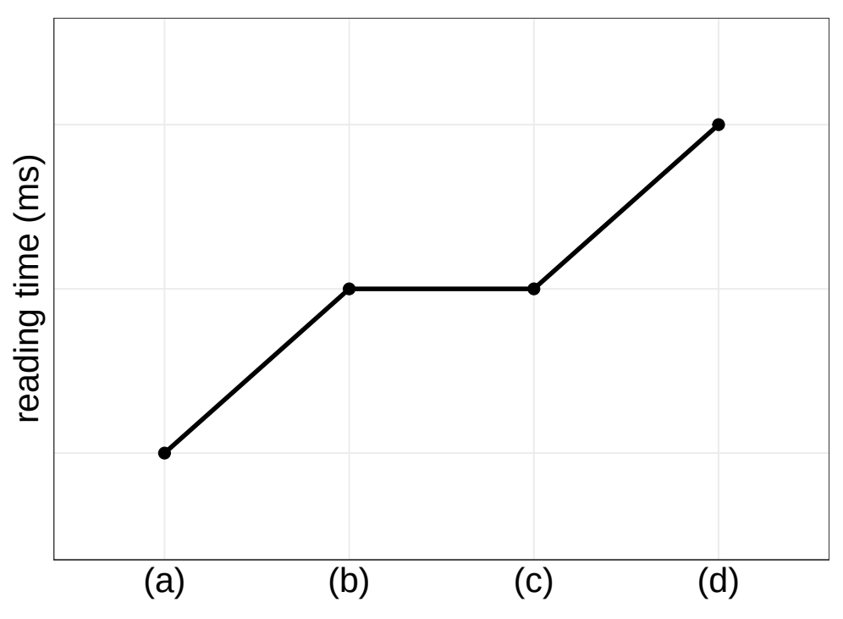
Expectation accounts

e.g., Hale (2001), Levy (2008)



Memory accounts

e.g., Lewis & Vasishth (2005)



5. LK13 Results

Results:
LK Expt 1: anti-locality effect (d < c)
LK Expt 2: locality effect (d > c)

Conclusion: Locality outweighs anti-locality when syntactic complexity is high.

Pattern seen across LK Expt 1 and LK Expt 2 suggests a cross-over interaction. We test the 'Load-Distance' interaction formally (see our Expts 5–7 below).

6. Investigating Replicability

Seven replication attempts of Levy & Keller (2013)

Why replicate Levy & Keller (2013)?

- typical participant sample size
- theoretically highly plausible results
  - support expectation-based accounts e.g., Hale (2001), Levy (2008)
  - support memory-based theories e.g., Lewis & Vasishth (2005)

7. Definitions: Replication Success

Definition 1: A statistically significant result in the original study is also found to be significant in the replication attempt.

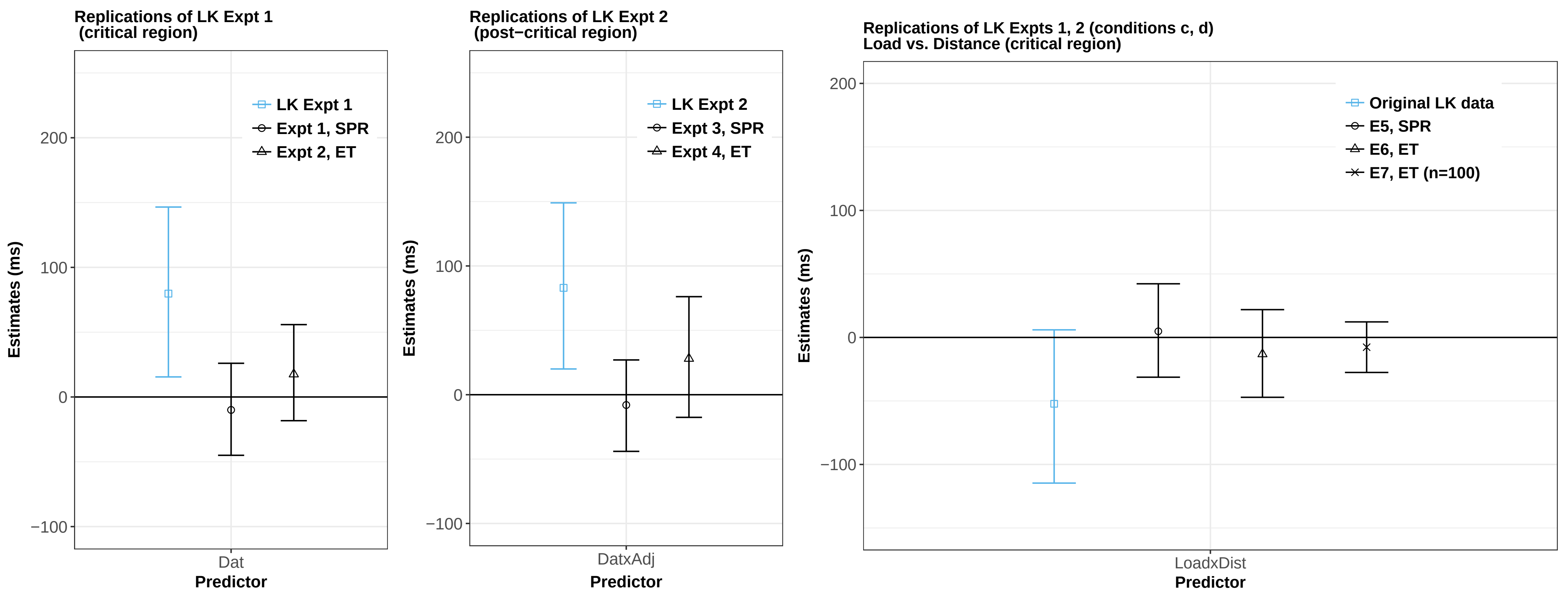
Definition 2: The estimated mean from a replication attempt falls within the 95% credible interval of the original estimate.

8. 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

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

9. Replication results: Expts 1–6 (N=28 each), Expt 7 (N=100)



Posterior means with 95% credible intervals computed from a Bayesian maximal linear mixed model using Stan. Shown are mean reading time (total reading time for eye-tracking) at the critical region (**versteckt**, *hid*) or at the post-critical region of the original studies vs. our replication attempts.

10. Conclusion

Seven replication attempts found no evidence of the effects found in the original study according to Definition 1 of Replication Success.

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

11. 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