The Importance of Replication in Psycholinguistics  
Daniela Mertzen, Lena A. Jäger & Shravan Vasishth  
University of Potsdam  
mertzen@uni-potsdam.de

1. Motivation
Failure to replicate published work
- Paape & Vasishth (2016): local coherence in German; self-paced reading (SPR), N = 40
- Husain, Vasishth, & Srinivasan (2014): expectation vs locality effect in Hindi; SPR, N = 60

2. The Problem
Low power due to small sample sizes (Gelman & Carlin, 2014) leads to:
(i) high proportion of null results
If power is ≤ 20% (not uncommon in psycholinguistic studies) \( \Rightarrow \) probability of finding a true effect only 20% or less
For example, in L.A. Jäger et al., 2017 (Appendix B): effects (range: -16 to -33 ms, sd = 150 ms, N = 40, SPR studies) had power estimates ranging from 10% to 30%

If the estimated effect is statistically significant given that the effect is not 0, under repeated sampling, low power leads to:
(ii) Type M (= magnitude) error, i.e. an overestimation of the effect
(iii) Type S (= sign) error, i.e. effect in the wrong direction

3. Investigating Replicability
Six replication attempts of Levy & Keller (2013): locality & anti-locality effects in German, eye-tracking, Experiments E1 and E2, N = 28 each
Why replicate Levy & Keller (2013)?
- typical participant sample size
- theoretically highly plausible results
  - support surprisal e.g. Hale (2001), Levy (2008)
  - support memory-based theories e.g. Lewis & Vasishth (2005)
  - existing empirical evidence
    * anti-locality effect e.g. Linzen & F. Jaeger (2015)
    * locality effect e.g. Bartek et al. (2011)

- Results E1: anti-locality effect (cond. d < c)
- Results E2: locality effect (d > c) \( \Rightarrow \) locality outweighs anti-locality when syntactic complexity is high
Seemingly robust results \( \Rightarrow \) effect should be replicable

4. Design & Materials
2 × 2 fully-crossed factorial design
- Factor 1: Position of dative NP (NP) (main- vs subordinate clause)
- Factor 2: Position of PP adjunct (PP) (main- vs subordinate clause)
E1: target construction in main clause
E2: same construction embedded in relative clause \( \Rightarrow \) higher syntactic complexity
Critical region: matrix clause verb (versteckt, below) referring back to subject (Hans, below)

5. Levy & Keller (2013) Predictions
Surprisal theory (anti-locality) a > b; c > d
Memory accounts (locality) a < b; c < d

6. Replication results (N = 28 each)

7. Conclusion
Replication failure:
Even seemingly robust results should be scrutinized
Low sample size
\( \Rightarrow \) low statistical power
\( \Rightarrow \) low probability of obtaining accurate estimates of true parameters (Type M error)

- Prior to running an experiment compute sample size based on power calculations
- Replicate the effect to establish robustness (see Nicenboim et al. (under revision), Safavi et al., 2016)

8. Future directions
We are currently planning a relatively high power large scale replication attempt of our eye-tracking study E6 (cond. c and d of the original E1 and E2 by Levy & Keller, 2013)

Mean reading time (total reading time for eye-tracking) and 95% confidence intervals at the critical verb (versteckt) of original studies vs our replication attempts (our E5 and 6 combine cond. c, d of E1 and E2 by Levy & Keller as only these showed a statistically significant effect)