The importance of replication in psycholinguistics

Daniela Mertzen, Lena Jäger & Shravan Vasishth (University of Potsdam)
mertzen@uni-potsdam.de

We report a project that began after two published studies from our own lab [1, 2] could not be replicated. Our aim was to investigate whether the relatively small sample sizes in psycholinguistic studies lead to statistically significant results that subsequently cannot be reproduced. Statistical theory [3] states that, if an effect is indeed present in reality, repeatedly conducting an experiment with low power will lead to a high proportion of (a) null results, and (b) exaggerated effect sizes which could be in the wrong direction (as illustrated in Figure 1). Given that many studies in psycholinguistics are underpowered (see, e.g. [4], Appendix B), we should see numerous null results and, in a large number of psycholinguistic publications, exaggerated effect sizes [5]. To investigate this, we tried to replicate a published eyetracking (reading) study from another lab [6]. We chose [6] as a test case because it had a sample size typical for psycholinguistics (28 subjects in each of their two studies), the results are plausible and supported by theory [7,8; 9,10], and there is independent evidence in favor of the two phenomena that [6] investigated: locality [English:11,12; Hindi:2,13; Persian:14] and anti-locality effects [English: 15–17; German: 18–20]. Surprisal predicts that intervening material between a verb and its subject facilitates the parser's prediction of the verb and, thus, its processing (anti-locality effect), while memory-based theories [9,10] predict more processing difficulty at the verb due to intervening material (locality effect). This was tested in [6] using complex syntactic structures where the subject and the critical matrix verb had four configurations in which the position of a *dative noun phrase (*DAT) and of a *prepositional adjunct* (ADJ) were manipulated.

a.	[Nachdem	der	Lehrer...],	hat	Hans					[ACC den	Fußball] **versteckt**.
	[After	*the*	*teacher..],*	*has*	*Hans*					*[ACC the*	*football]* *hid.*
b.	[Nachdem	der	Lehrer...],	hat	Hans	[ADJ zur	zusätzlichen	Ahndung]		[ACC den	Fußball] **versteckt**.
	[After	*the*	*teacher..],*	*has*	*Hans*	*[ADJ as*	*additional*	*payback]*		*[ACC the*	*football]* *hid.*
c.	[Nachdem	der	Lehrer...],	hat	Hans	[DAT dem	ungezogenen	Sohn]		[ACC den	Fußball] **versteckt**.
	[After	*the*	*teacher..],*	*has*	*Hans*	*[DAT the*	*naughty*	*son]*		*[ACC the*	*football]* *hid.*
d.	[Nachdem	der	Lehrer...],	hat	Hans	[ADJ zur	... Ahndung]	[DAT dem	... Sohn]	[ACC den	Fußball] **versteckt**.
	[After	*the*	*teacher..],*	*has*	*Hans*	*[ADJ as*	*... payback]*	*[DAT the*	*... son]*	*[ACC the*	*football]* *hid.*

'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.'

According to surprisal, the verb *versteckt* in condition (*a)* should be read slower than in (*b)*, and (*c)* slower than (*d)*; hence, (*d)* should show the most facilitation (anti-locality). Exp.1 had the target construction in a main clause; in Exp. 2 the same clauses as in Exp.1 were embedded in a relative clause, making the sentence syntactically much more complex. Results for Exp.1 showed a speed-up for (*d)* over (*c)* consistent with surprisal. However, in Exp. 2, the opposite pattern was observed: (*d)* was read slower than (*c)* (locality). It is suggested by [6] that locality effects outweigh anti-locality effects when syntactic complexity is high, as in Exp. 2.

We conducted a total of six attempted replications of Exp.1 and 2 reported by [6] using the same 24 items and participant sample size of 28 as [6]. These were tested using self-paced reading (SPR) (our Exp.1, 2) and eyetracking (ET, our Exp. 3, 4). Our Exp. 5 (SPR) and 6 (ET) tested only conditions *(c)* and *(d)* from the two studies by [6] as only effects in (*c)* and (*d)* of the original study were statistically significant. All six attempts to reproduce the originally significant results were unsuccessful (see Figure 2); no dependent measure showed any effect.

Our failure to replicate the above studies [1, 2, 6] is indicative of low statistical power. Given that it is common to run reading time experiments with 24–36 participants in psycholinguistics, it is very unlikely that we can obtain accurate estimates of the true parameters. We conclude that a significant number of published findings—including our own—may be reporting exaggerated estimates [3, 5]. We suggest that future studies in psycholinguistics determine sample size based on power calculations before running experiments, and establish robustness of their results by replicating the effect.

Figure 1

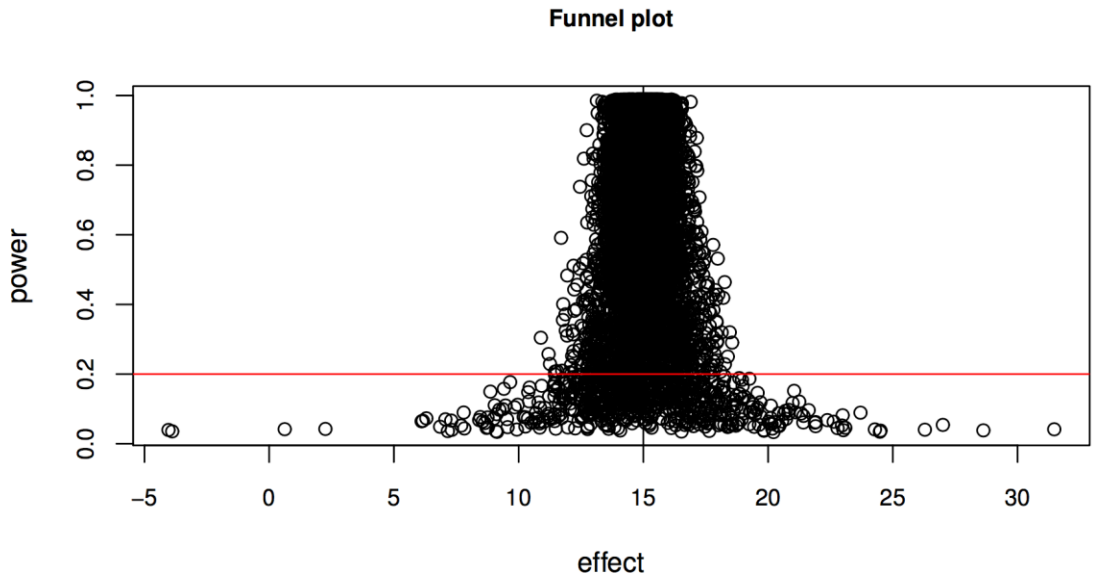
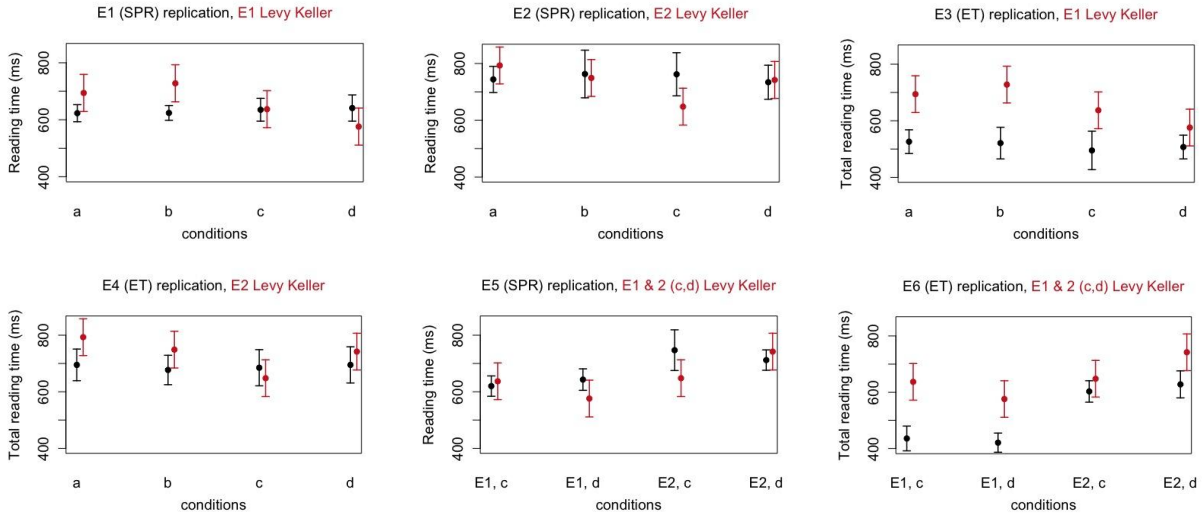


Figure 2



References

[1] Paape & Vasishth (2016). *Lang Speech.* **[2]** Husain et al. (2014), *PLoS ONE*. **[3]** Gelman & Carlin (2014). *Perspect Psychol Sci.* **[4]** Jäger et al. (2017). *JML*. **[5]** Vasishth & Gelman (2017). (submitted). **[6]** Levy & Keller (2013). *JML*. **[7]** Hale (2001). *Proceedings NAACL.* **[8]** Levy (2008a). *Cognition*. **[9]** Gibson (2000). In *Image, Language, Brain*. **[10]** Lewis & Vasishth (2005). *Cognitive Sci*. **[11]** Grodner & Gibson (2005). *Cognitive Sci.* **[12]** Bartek et al. (2011). *JEP:LMC.* **[13]** Husain et al. (2014). *Journal of Eye Movement Research*. **[14]** Safavi et al. (2016), *Frontiers*. **[15]** Vasishth et al. (2010). *Lang and Cogn Process*. **[16]** Linzen & Jaeger (2015). *Cognitive Sci*. **[17]** Boston et al. (2008). *Journal of Eye Movement Research.* **[18]** Konieczny (2000). *J Psycholinguist Res.* **[19]** Boston et al. (2011). *Lang and Cogn Process*. **[20]** Frank et al. (2015). *Cognitive Sci.*