Prolific vs. MTurk: Differences in overall reading times and accuracy, but not effect sizes.

Grusha Prasad (gprasad@colgate.edu) Omar Fargally (ofargally@colgate.edu)
Colgate University *Colgate University*

Introduction: Historically, psycholinguistics has focused on testing *qualitative* predictions. Consider the temporarily ambiguous sentence A and its unambiguous counterpart B:

A. The girl fed the lamb remained relatively calm despite wanting chicken. (RRC)
B. The girl who was fed the lamb remained relatively calm despite wanting chicken. (URC)

Most psycholinguistic work has focused on answering yes/no questions like whether the disambiguating region in A (underlined) is read more slowly than in B (garden path effect; GPE; [1]), or whether GPEs decrease with increased exposure (adaptation; [2]). With the advent of broad-coverage computational models (e.g., language models), more recent work has focused on generating predictions about the *magnitude* of these effects, and testing these predictions with data from large-scale web-based experiments [3,4]. In this work we ask whether the magnitudes estimated from web-based experiments are *reliable*. Prior work suggests that the estimated magnitudes might be impacted by factors like choice of the web-based platform: for a different type of GPE (see Tab 1 for example), the effect size was ~30ms in an experiment run on MTurk [5] but ~115ms in an experiment run on Prolific [3]. To study the potential impacts of platform choice, we use Prolific to replicate an adaptation experiment previously run on MTurk [6]. We chose [6] for two reasons: first, it was highly powered (N=828); second, it lets us study the magnitudes of both two-way (GPE x Platform) and three-way (GPE x Group x Platform) interactions.

Experimental design: As in [6], we used a between-group design with two phases: in the adaptation phase, participants were either exposed to 16 RRCs and 16 fillers (RRC-exposed) or 32 fillers (Filler-exposed). In the test phase, both groups were exposed to 12 RRCs, 12 URCs and 24 fillers. We define GPE as the difference in reading times between RRCs and URCs in the test phase, and the adaptation effect as the between group difference in GPE.

Methods: We recruited 670 US-based native English speakers from Prolific and included 641 participants in our analyses: were excluded 6 because they reported they were non-native speakers on the demographic survey; and 23 because their filler accuracy was lower than 80% (see Tab2 for exclusion rate comparison with [6]). Participants were compensated with 4 USD. Using the data from this experiment for Prolific, and data from [6] for MTurk, we fit a Bayesian mixed effects model with ambiguity, group, platform and the corresponding 2-way and 3-way interaction terms as fixed effects, and with the maximal random effect structure (see Tab 3)

Results: The proportion of participants excluded due to low filler accuracy was much higher in MTurk than Prolific (21.4% vs. 3.5%; Tab 2). Additionally, there was moderate evidence that Prolific participants read sentences more slowly than MTurk participants (BF = 9.72). These results are consistent with the observation that MTurk can yield some “bot-like” data [e.g., 7] that is less likely in Prolific. However, when comparing the effects of interest across the platforms, there was moderate evidence that there were no differences in the magnitude of GPEs (BF = 0.27) or adaptation effects (BF = 0.21); see Fig 2. Taken together, these results suggest that differences in effect sizes observed in [3] and [5] were a likely result of other factors (e.g., stimuli and # participants) and that garden path and adaptation effects observed in one web-based platform are likely to generalize to other web-based platforms both qualitatively and quantitatively. Future work can explore if this conclusion holds when comparing effect sizes in web-based and in-lab experiments, in other empirical phenomena (e.g., agreement attraction, priming), and in other experimental paradigms (e.g., MAZE, acceptability judgments).

References:
[1] MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4), 676–703.
[2] Fine, A. B., Jaeger, T. F., Farmer, T. A., and Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. *PLoS One,* 8(10):e77661.
[3] van Schijndel, M. & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*.
[4] Huang, K. J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2022). *S*PR mega-benchmark shows surprisal tracks construction-but not item-level difficulty. *In the 35th HSP*.
[5] Prasad, G., & Linzen, T. (2019). How much harder are hard garden-path sentences than easy ones? *CogSci conf.*
[6] Prasad, G. and Linzen, T., 2021. Rapid syntactic adaptation in self-paced reading: Detectable, but only with many participants. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 47(7), p.1156.
[7] Huang, K. J., & Staub, A. (2023). The transposed-word effect does not require parallel word processing: Failure to notice transpositions with serial presentation of words. *Psychonomic Bulletin & Review*, 30(1), 393-400.

NP/Z Ambiguous:	When the little girl attacked the lamb remained relatively calm.
NP/Z Unambiguous:	When the little girl attacked , the lamb remained relatively calm.

Tab 1: Example sentences for Transitive/ Intransitive (NP/Z) GPE used in [4] and [5].

	Prasad & Linzen (2021)	**Current study**
Platform (year collected)	MTurk (2019)	Prolific (2023)
# Participants recruited	828	670
# Non-native participants (% total)	11 (1%)	6 (0.8%)
# Participants filler accuracy < 80% (% native)	175 (21.4%)	23 (3.5%)
Compensation	2 USD	4 USD

Tab 2: Comparing exclusion criteria and compensation across the two studies

Model (LMER notation)

```
log(RT) ~
  ambiguity * group * platform  +
  (1 + ambiguity | part) +
  (1 + ambiguity*group*platform | item)
```

Priors
Intercept: Normal(5.70, 1)
Fixed effects: Normal(0, 0.05)
SD (random effect): Normal(0, 0.1)
SD (residuals): Normal(0, 0.5)

Tab 3: Statistical model

Coefficient	**Estimate**	**SE**	**BF**
Intercept	5.87	0.020	Inf
ambiguity (RRC 0; URC 1)	-0.05	0.006	7.55e+06
group (Filler-exposed 0; RRC-exposed 1)	-0.04	0.022	2.58
Platform (Prolific 0; MTurk 1)	-0.05	0.022	9.72
ambiguity : group	0.03	0.007	168.14
ambiguity : platform	0.008	0.007	0.27
group : platform	-0.04	0.026	2.14
ambiguity : group : platform	-0.004	0.010	0.21

Tab 4: Estimates from BRMS model

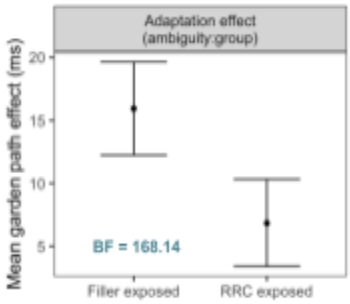

Fig 1: Adaptation effect across platforms

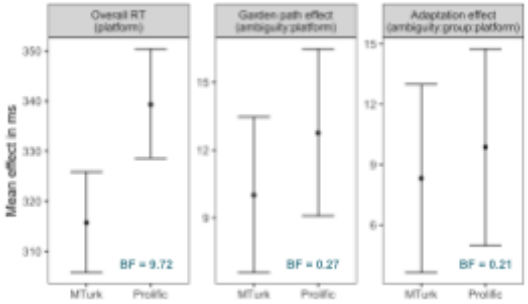

Fig 2: Main effect of platform, and interaction with GPE and adaptation

Figures were generated by sampling fixed effects from the posteriors, computing log RTs in each condition by summing together relevant effects and exponentiating them to raw RTs. The main effect of the platform reflects mean RTs, and all other interaction effects reflect the mean differences in RTs across the relevant conditions.