**Transitioning to online language production: a direct comparison of in-lab and web-based experiments**

Margaret Kandel (Harvard), Cassidy Wyatt (UMD), Colin Phillips (UMD)

At a time when much in-person human subjects experimentation has been halted, the ability to collect data from web-based sources is increasingly valuable to language scientists. While some language tasks are already frequently executed online (e.g. self-paced reading, surveys, typed sentence completion; [e.g. 1-3]), there have been fewer web-based studies eliciting recorded speech. The collection and quality of production data may be more susceptible to limitations of online research [cf. 3] than other linguistic data. Variations in internet connections, software, and hardware may make it difficult to collect consistent data or obtain representative participant samples, and recorded speech may be more variable or noisier when elicited and recorded outside of a controlled lab environment. To assess the quality of web-collected production data and how well it can detect phenomena and measure variables of interest to production research, we performed a direct comparison of in-lab and web-based experiments analyzing speech errors and the production time-course of responses. The experiments investigated a robust language production phenomenon: verb agreement attraction (1) [e.g. 4-6].

**Method:** We used a speeded scene-description task to elicit responses. This task elicits speech through a process that more closely resembles natural production than the traditional preamble paradigm [e.g. 4-6]. Participants were introduced to three aliens (blueys, greenies, pinkies) and described scenes of these aliens *mimming* (lighting their antennae) (Fig 1). Each scene contained two groups of aliens to encourage participants to disambiguate the subject using spatial prepositions (e.g. “the pinky above the greenies”). We manipulated the number of aliens in the scenes so that the NPs in the target SubjPs either matched or mismatched in number (Table 1). 1s was added to the response window of the web experiment to accommodate the online setting and more diverse subject pool. We looked for evidence of attraction in both the distribution of errors and the time-course of error-free sentences (using a forced-aligner; [7]).

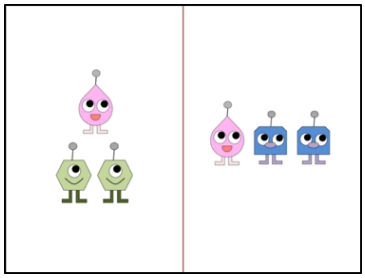
**Exp 1:** The in-lab experiment had 45 participants (34F; $M_{age}$ = 21, SD = 4.5). We found standard agreement attraction effects, reflected in higher error rates (Fig 2a) and greater probability of producing errors in the mismatch conditions ($p < 0.0001$). Sentences with no errors displayed slowdowns prior to verb articulation in these same environments (Table 2): participants were more likely to pause before the verb ($p < 0.0001$), and these pauses tended to be longer ($p = 0.058$). We saw a plural markedness effect [e.g. 4] on error likelihood ($p < 0.0001$). Singular attraction errors (PS condition) were more common than typically observed in preamble studies [cf. 8], though elevated PS error rates have been seen in other elicitation paradigms [e.g. 9, 10].

**Exp 2:** The online experiment had 37 participants (26F; $M_{age}$ = 41, SD = 9.97) recruited from Amazon Mechanical Turk. The experiment was conducted on PCIbex Farm [11] and was unsupervised. The audio quality of the responses was sufficient to identify agreement errors and to forced-align. We again found evidence of agreement attraction in error rates (Fig 2b) and probabilities ($p < 0.0001$) in addition to corresponding slowdowns in production time-course ($p$’s $< 0.0001$) (Table 2). The distributions of errors and pre-VP delays were comparable to Exp 1, though with fewer errors and more pauses, suggesting a tradeoff between errors and delays in articulation, perhaps due the longer response window. We again observed high PS error rates.

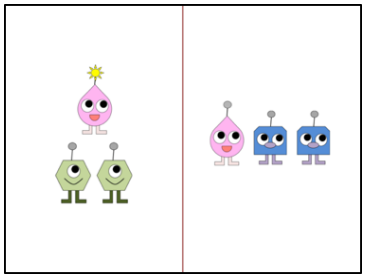
**Discussion:** The similarities in the results of our experiments indicate that web-based experimentation is a viable and attractive avenue for language production research. Data collection for the in-lab experiment took 3 months to complete, whereas the online experiment took only 9 days of data collection. Using a web-based platform allowed us to recruit a more geographically and age diverse subject sample. We employed several successful measures to minimize drop-out and trial loss and to reduce effects of equipment variation. Nevertheless, there were some differences in our online experiment, with slightly higher participant omission rates and effects of context variability on the forced-aligner’s ability to detect utterance onset. We believe that web-based experimentation will allow production research to proceed more flexibily and efficiently and provide easier access to the global population than ever before.

(1) Verb agreement attraction errors occur when nearby material interferes with normal agreement processes, as in the sentence **The key to the cabinets are on the table* [4]

**Figure 1:** Example scene with target sentence “the pinky above the greenies is mimming”



1s of preview



3s of mimming (4s in Exp 2)

**Table 1:** Experiment conditions

| Condition | Sub-Condition | Sample Sentence |
|---|---|---|
| Match | SS | the pinky above the greeny is mimming |
| Match | PP | the pinkies above the greenies are mimming |
| Mismatch | SP | the pinky above the greenies is mimming |
| Mismatch | PS | the pinkies above the greeny are mimming |

**Figure 2:** Participant agreement error rates by sub-condition

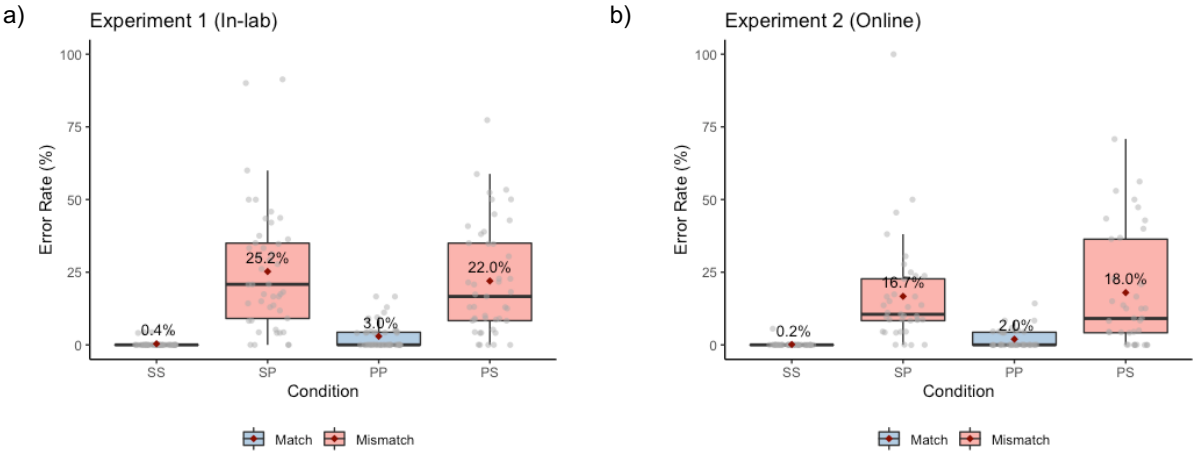



**Table 2**: Proportion of responses with pre-VP pauses & estimated pause durations

| Exp | Condition | Proportion | Duration | Sub-Condition | Proportion |
|---|---|---|---|---|---|
| 1 | Match | 0.05 | 73ms | SS | 0.05 |
| | | | | PP | 0.06 |
| | Mismatch | 0.15 | 98ms | SP | 0.15 |
| | | | | PS | 0.18 |
| 2 | Match | 0.12 | 57ms | SS | 0.11 |
| | | | | PP | 0.13 |
| | Mismatch | 0.31 | 85ms | SP | 0.29 |
| | | | | PS | 0.32 |

**References:** [1] Corley & Scheepers, 2002; [2] Enochson & Culbertson, 2015, [3] Sitka & Sargis, 2005, [4] Bock & Miller, 1991; [5] Bock & Cutting, 1992; [6] Bock & Eberhard, 1993; [7] McAuliffe et al., 2017; [8] Eberhardt et al., 2005; [9] Staub, 2009; [10] Veenstra et al., 2014; [11] Zehr & Schwartz, 2018