Gender stereotype subversion across English *wh* islands

Dennis Ryan Storoshenko, University of Calgary and Jesse Weir, Simon Fraser University

Issue   Kotek et al. (2020) raises important concerns regarding the content of example sentences in syntax. In this paper, we extend the discussion to the experimental setting. We report on a preliminary study testing whether upholding or subverting a gender stereotype impacts participant performance in a rating task of ungrammatical items. Our findings are that while sentence ratings are not significantly shifted by stereotype subversion, there is a significant impact on response times, and that including stereotypes may lower ratings overall.

Background   Often, stimulus sentences in experimental research can perpetuate stereotypes, even when the research question has no bearing on the content of stimuli. However, in processing studies where gender stereotypes are the object of examination, subverting stereotypes leads to slower reading times (Carreiras et al., 1996; Garnham et al., 2002). A question arises then of how stereotyped content influences studies where that content is unrelated to the research question.

Study Design   Participants were presented with English *wh*-questions where a gender stereotyped object was extracted from a *wh*-adjunct island. The subject was a gender stereotyped profession, determined from a previous pre-test, and the extracted object would either Match (uphold) or Mismatch (subvert) the stereotype. A gendered pronoun inside the island, referring to the subject, would also either Match or Mismatch the stereotype. This results in the following paradigm:

Which <u>tie</u> did the **general** know where <u>he</u> had left after the parade?	MatchMatch
Which <u>tie</u> did the **general** know where <u>she</u> had left after the parade?	MatchMismatch
Which <u>lipstick</u> did the **general** know where <u>he</u> had left after the parade?	MismatchMatch
Which <u>lipstick</u> did the **general** know where <u>she</u> had left after the parade?	MismatchMismatch

Participants rated the sentences on a seven point Likert scale for acceptability. Participants saw a total of four items from each condition (16 trial items) along with 58 distractor items. Results are reported from 64 monolinguial English-speaking participants who were recruited online using Prolific Academic; data was collected using PsychoPy3 (Pierce et al., 2019).

Results   All stereotype conditions have relatively consistent mean ratings from 3.08 to 3.24. Using lme4 and lmerTest (Bates et al., 2015; Kuznetsova et al., 2017), we determine that the best fitting model for sentence ratings includes the gender stereotype (conterbalanced "male" and "female" professions) as the only fixed effect, but the effect itself is not significant. Eight "control" distractor items have the same configuration as the trial items, but use unstereotyped nouns and semantically congruent extracted objects (e.g. *which helmet did the cyclist...*). These have a mean rating of 3.72, significantly higher (p=0.002) than even the MatchMatch mean rating of 3.24. Turning to the reaction time data, there is no significant difference between the mean reaction time of the control distractors (7.50s) and the MatchMatch condition (7.27s). However, in the lmer modelling of the stereotype effects, we find a significant (p=0.043) interaction effect based on the combinations of stereotype subversion. The doubly subverted MismatchMismatch trials are significantly faster (approximately 0.75s) than the MatchMismatch and the MismatchMatch trials.

Discussion   The results of this study suggest that participants are unlikely to change the rating of an island violating sentence based on gender stereotype subversion. Though, echoing the Kotek et al. concerns, there is a suggestion that including stereotypes at all depresses ratings. While a follow-up self-paced reading study will more clearly localize any processing effects of stereotype subversion interacting with encountering a *wh*-isalnd, the initial reaction time data suggests that online studies that are not directly examining stereotyped content may be better advised to avoid such content completely.

References

Bates, Douglas, Martin Maechler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software* 67:1–48.

Carreiras, Manuel, Alan Garnham, Jane Oakhill, and Kate Cain. 1996. The use of stereotypical gender information in constructing a mental model: Evidence from English and Spanish. *The Quarterly Journal of Experimental Psychology* 49A:639–663.

Garnham, Alan, Jane Oakhill, and David Reynolds. 2002. Are inferences from stereotyped role names to characters' gender made elaboratively? *Memory and Cognition* 30:439–446.

Kotek, Hadas, Rikker Dockum, Sarah Babinski, and Christopher Geissler. 2020. Gender bias and stereotype in linguistic example sentences. URL: lingbuzz/005367.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H.B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82:1–28.

Pierce, Jonathan, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyukli Sogo, Erik Kastman, and Jonas Kristoffer Lindelov. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* 51:195–203.