Emergence of conventions in group communication: Evidence from 2-4 player reference games

Veronica Boyce[1]

[1] Stanford University

Author Note

Email: vboyce@stanford.edu; This paper was written to fulfill the First Year Project requirement for the Stanford Psych Department.

Abstract

In repeated reference games where a speaker describes the same set of images to a listener over a series of rounds, the number of words used decreases as the pair converge on ad-hoc names for the images. The dynamics of this efficient reference formation is well-studied in dyads; however much communication takes place in larger groups, which are rarely studied in this paradigm. The current work extends iterated reference games to groups of 2-4 people who rotate between speaker and listener roles in an online game with text-based communication. Across 53 games and more than 50K total words, we find high accuracy and patterns of reduction regardless of group size.

Emergence of conventions in group communication: Evidence from 2-4 player reference games

Verbal communication is an integral part of our daily lives. We coordinate schedules with partners, socialize with friends over board games, learn and teach in seminar classes, and listen to podcasts. Our communicative environments range in size from one-on-one to small group to large group to broadcast, but the goal of efficient communication is held in common. One necessity for efficient communication is shared reference expressions; when referring to a thing or an idea, it needs some sort of name that the interlocutors will jointly understand. In many cases, there are widely shared conventionalized expressions, but in other cases, spontaneous ad-hoc expressions are needed.

The formation of these new reference expressions is well-studied in dyadic contexts; however, dynamics may be different in larger groups, which are less studied. Our current work builds on the dyadic reference game tradition by extending it to small groups.

The typical paradigm for studying partner-specific referring expressions is an iterated reference game with asymmetric knowledge. That is, each round there is a speaker who knows what the target is and a listener who does not have this information. In Clark and Wilkes-Gibbs (1986), each speaker described 12 tangrams in order, so their listener could correctly order the images. After receiving feedback, the pair repeated the task with the same images but a new order, for a total of 6 repeats. Crucially, Clark and Wilkes-Gibbs (1986) found that reference expressions condense over the course of repeated reference to the same image. Early descriptions are long and make reference to multiple features in the image, but by later rounds definite shorthand names are used.

Recently, online participant recruitment and web-based experiments have made it possible to study this convergence in larger populations using a text-based communication interface. In Hawkins, Frank, and Goodman (2020), 83 pairs completed a cued version of the iterated reference experiment. On each trial, the speaker saw one image highlighted and described it to the listener who clicked on what they thought the target was. Both players received feedback before moving on to the next target image. All images were highlighted each block, for a total of 6 blocks. Speakers produced fewer words per image in later blocks than in earlier blocks, in line with results from face-to-face, oral paradigms.

While this reduction pattern is robust for dyads, less is known about the how utterances are adapted in larger groups. A couple of studies point to some potential difficulties in trying to communicate with multiple people at once.

Yoon and Brown-Schmidt (2019) had speakers complete a sorting task with some listeners, so that they would have a common ground of shared names for the images. Then in a test phase, the speaker described these images to a group of either all knowledgeable listeners from the sorting task, new listeners who had not done the sorting task, or a mix of knowledgeable and new listeners. Speakers produced longer utterances when any new listeners were present than with only experienced listeners. This might predict slower reduction in larger groups where there will inevitably be some variability in how people understand reference expressions. These studies included 3-hour experiments that were very time and labor intensive, but some of the questions about group dynamics may be addressable in online experiments taking advantage of natural variation in understanding without artificially inducing large knowledge differences.

It's difficult to communicate with naive listeners, but it can be even harder to communicate with someone with entrenched preconceptions. Weber and Camerer (2003) induced these conceptual differences by having two pairs of people (each pair representing a "firm") do an iterated reference game with the same set of pictures. After 20 rounds, there was a "merger" where the listener from one group joined the other group. The reference game continued with the speaker communicating to both their original listener and the new listener. After the merger, there was a jump in how long it took either listener to make a selection. Even after several more rounds, listeners were still not as fast as before the merger. With larger groups of people all speaking together, there's a greater chance for different people to independently develop different conceptualizations of an image, and it may be difficult for them to understand each other or agree on a common term of reference.

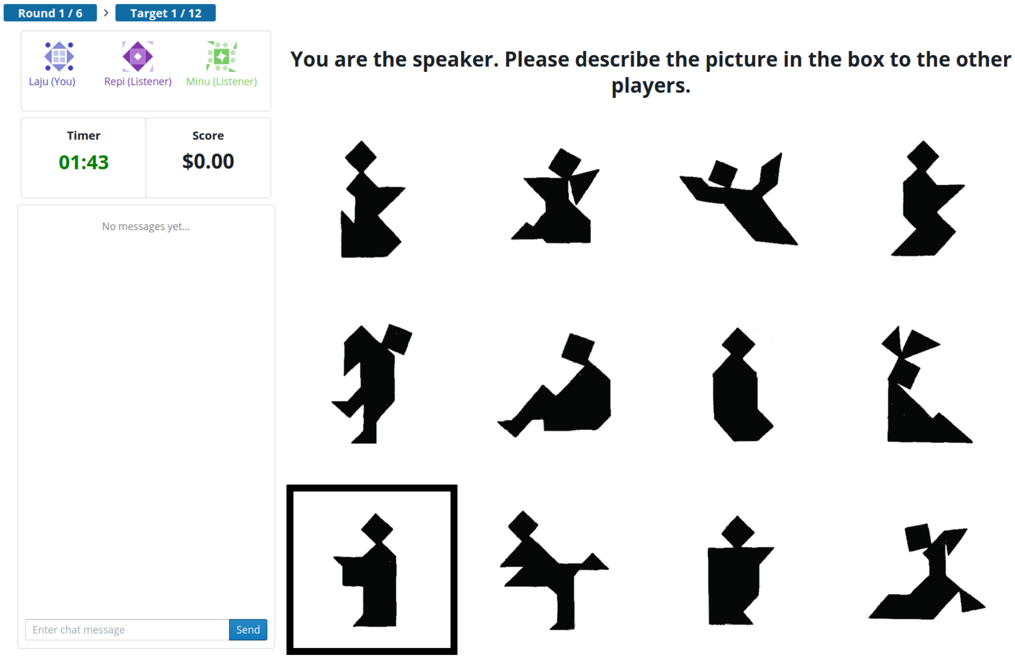
Studies vary in whether the same participant keeps the speaker role the entire game (ex. Clark & Wilkes-Gibbs, 1986) or whether the roles rotate (ex. pre-merger rounds of Weber & Camerer, 2003). Role rotation makes the paradigm more similar to collaborative puzzle-solving exercises also used to study conventions (Garrod & Doherty, 1994; Ibarra & Tanenhaus, 2016).

In general, listeners expect conventions to be maintained, but they are not surprised to new descriptions of a familiar object if it comes from a new speaker (Metzing & Brennan, 2003) or if a new listener is present (Yoon & Brown-Schmidt, 2014). It's unclear how these expectations map onto a group of people rotating roles in the task who are all present the entire time. Do later speakers count as new, or are they expected to follow conventions they've already heard? Do additional non-new listeners license changes in descriptions?

In this work, we extend the dyadic repeated reference game paradigm of Hawkins et al. (2020) to games for 2-4 players who rotate between speaker and listener roles. We compare accuracy and reduction rates in groups of different sizes.

Methods

We adapted the methods of Hawkins et al. (2020), adjusting them to work for multi-player games with rotating speakers. Participants played a repeated reference game where a speaker saw an array of tangrams with one indicated (Fig 1) and had to communicate which figure to click to the listeners using the chat box. Within each block, each of the 12 tangrams was the target once, and the speaker role rotated each block, so all participants were the speaker at least once. Games ran for a total of 6 blocks. We recorded what participants said in the chat, as well as who selected what image and how long they took to make their selections. The experiment was implemented in Empirica (Almaatouq et al., 2021); materials to run the experiment, as well as data and code are available at at https://github.com/vboyce/FYP.



You are the speaker. Please describe the picture in the box to the other players.

Figure 1. Screenshot of the speaker's view. Participants see all 12 tangram images.

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. Our preregistration is at https://osf.io/cn9f4.

Participants

Participants were recruited using the Prolific platform between 4th and 10th of May 2021. We screened for participants who were fluent, native English speakers. Participants were paid $7 for 2-player games, $8.50 for 3-player games, and $10 for 4-player games (with the intention of a $10 hourly rate using pilot studies to estimate median game lengths), in addition to up to $2.88 in performance bonuses.

Our intended sample size was 20 complete games in each group size, but we ended up with 15 complete 2-player games (4 partial), 18 complete 3-player games (2 partial), and 20 complete 4-player games (1 partial). We excluded incomplete blocks from analyses, but included complete blocks from partial games. (Partial games occurred when a participant disconnected early, for example due to internet trouble.)

Materials

We used the 12 tangram images used by Hawkins et al. (2020) and Clark and Wilkes-Gibbs (1986) (see Fig 1). These images were displayed in a grid with order randomized for each participant. The same images were used every block.

Procedure

We implemented the experiment using Empirica, a Javascript-based platform for running real-time interactive experiments online (Almaatouq et al., 2021). Code for running this experiment is available at https://github.com/vboyce/FYP. From Prolific, participants were directed to our website where they navigated through a self-paced series of instruction pages explaining the game. Participants had to pass a quiz to be able to play the game. They were then directed to a “waiting room” screen until their partners were ready.

Once the game started, participants saw screens like Fig 1. Each trial, the speaker had to describe the highlighted tangram image so that the listeners could identify and click it. All participants were free to use the chat box to communicate, but listeners could only click once the speaker had sent a message. Once a listener clicked, they could not change their selection. Once all listeners had selected (or a 3-minute timer had run out), participants were given feedback. Listeners only learned whether they individually had chosen correctly or not; listeners who were incorrect were not told the correct answer. The speaker saw which tangram each listener had selected. Listeners got 4 points for each correct answer; the speaker got points equal to the average of the listeners’ points. These points translated into cents of performance bonus at the end of the experiment.

In each block, each of the 12 tangrams was indicated to the speaker once. The same person was the speaker for an entire block, but participants rotated roles between blocks. Thus, over the course of the 6 blocks, participants were speakers 3 times in 2-player games, 2 times in 3-player games and once or twice in 4-player games.

After the game finished, participants were given a survey asking for optional demographic information and feedback on their experience with the game.

Data analysis

I skimmed through the chat transcripts, tagging utterances that did not refer to the current tangram. These were primarily pleasantries (“Hello”), meta-commentary about well or fast the task was going and confirmations or denials (“ok”, “got it”, “yes”, “no”). We exclude these utterances from our analyses. Note that chat lines sometimes included non-referential words in addition to words referring to the tangrams (“ok, so it looks like a zombie”, “yes, the one with legs”); these lines were retained intact. We conducted data processing and analyses using R (Version 4.0.3; R Core Team, 2020) and the R-packages *brms* (Version 2.14.4; Bürkner, 2017, 2018), *here* (Version 1.0.1; Müller, 2020), *jsonlite* (Version 1.7.2; Ooms, 2014), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *rlang* (Version 0.4.10; Henry & Wickham, 2020), *rstan* (Version 2.21.2; Stan Development Team, 2020), and *tidyverse* (Version 1.3.0; Wickham et al., 2019).

Results

We find results generally in line with previous work on dyads. Overall, participants had high and increasing accuracy, coupled with faster response times, and decreases in utterance length showing the classic reduction pattern.

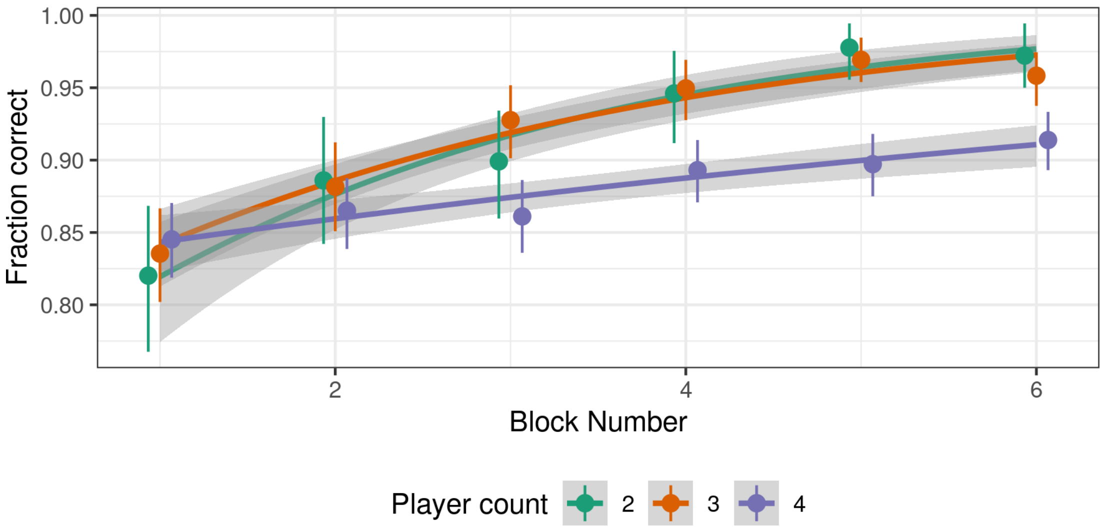


Figure 2. Listeners have high accuracy which increases of the course of the game. Accuracy rates are shown for each block, error bars are bootstrapped 95% CIs.

Most groups were accurate in their selections, with accuracy rising over blocks (Fig 2). This indicates that speakers were usually successful at conveying the intended referents. Participants are more accurate in later blocks 1.66 [1.15, 2.21].[1] 4-player games show lower gains in accuracy than smaller games, but neither the number of players nor the interaction of players and block are reliably different from 0. We do not have a clear explanation for this possible difference or what pattern to expect for even larger (ex. 5 person) games.

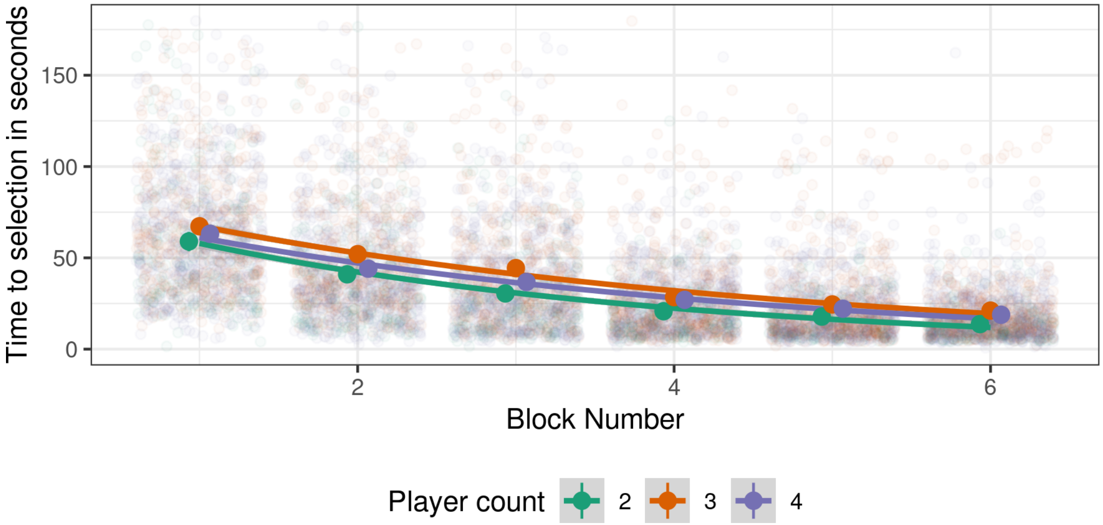


Figure 3. Listeners selected images faster in later blocks. Only times to correct responses are shown.

Participants selected images faster in later blocks (Fig 3), although there is wide

[1] Estimate and 95% credible intervals from a Bayesian Bernoulli model with formula: correct ~ block * numPlayers + (block | tangram) + (block | playerId/gameId).

variability. This speed up is consistent with prior work by Weber and Camerer (2003) which used speed as the dependent measure.

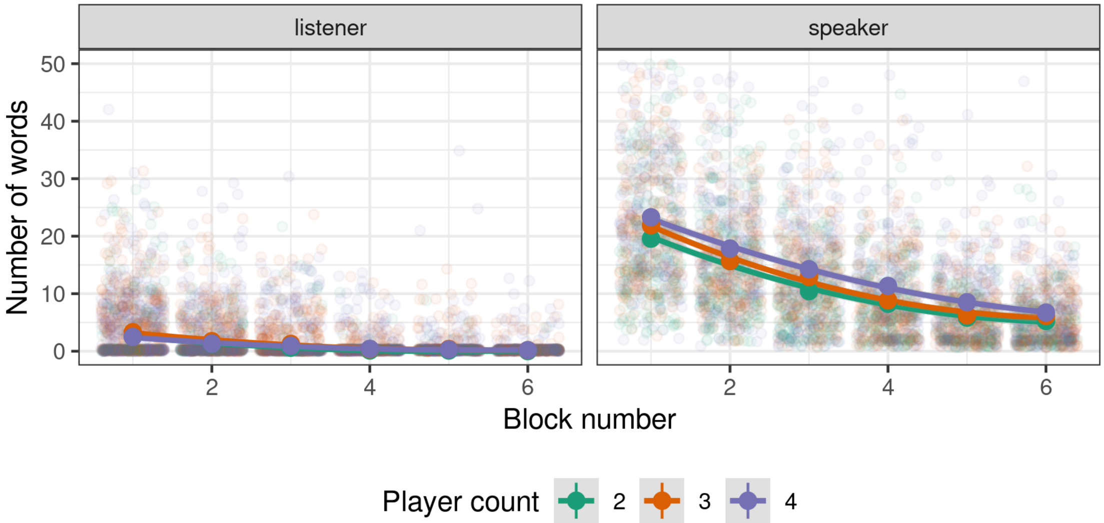


Figure 4. Speaker and listeners say fewer words in later blocks. Note: y-axis clipped at 50 which hides a few speaker outliers.

The main effect of interest is whether speakers and listeners reduce in the number of words they say over the course of repeated reference. As shown in Fig 4, the number of words produced does decrease. Listeners often don't talk much, but are more likely to ask questions or make clarification in early blocks. Speakers make longer utterances in early blocks and reduce to shorter utterances in later blocks. Notably, this shortening pattern occurs even as speakers rotate. In aggregate, the effect of being one block later is -3.22 [-4.95, -1.55] words. The overall effect of having more players in a group is 1.93 [-0.15, 4.02] words per additional player.[2] This estimate is uncertain because of a relatively small number of groups and wide group-level variability.

This variability can be seen in Fig 5. While the averaged data shows a smooth reduction in the number of words, individual trajectories for specific tangrams in specific groups are more varied. Reduction is not monotonic, as some later speakers use more words than were used in earlier blocks.

Because the ground truth answers are not provided to listeners who make mistakes, they may not learn what an utterance referred to (unless they ask in the chat). What happens if a listener gets a tangram wrong and then is the speaker on the next block? For that tangram, they are unlikely to build off the previous description they didn't understand. In contrast, a speaker who previously got the tangram right is likely to continue the conceptualization used so far and conventionalize it more, such as by reducing unneeded details. Taken together, this leads to the hypothesis that speakers should say more words

[2] Estimates and 95% credible intervals from a Bayesian linear model with formula: words ~ block * numPlayers + (block | tangram) + (1 | playerId) + (1 | tangram_group) + (block | gameId).

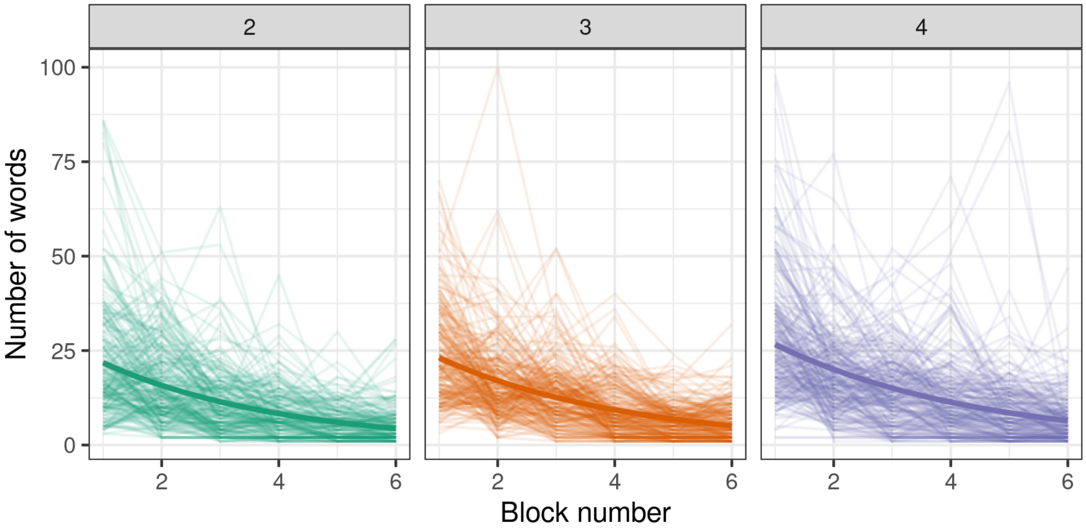


Figure 5. Words said by the speaker for each tangram in each group. Each referent/group trajectory is a thin line; aggregate curve is bolded. No outliers were omitted.

when they got the tangram wrong the previous block, after controlling for other effects. This is borne out; speakers say 4.15 [2.54, 5.79] more words when previously wrong.[3]

Discussion

The overall pattern of utterances shortening over repeated reference extends to groups of 3 or 4 people talking together and rotating between speaker and listener roles. Rotating speakers gives a stronger interpretation of reduction as conceptual agreement because more people have to produce the shorthand names.

We provided less feedback than previous studies such as Hawkins et al. (2020). This low level of feedback means that there isn't a way for people to find out what was meant for utterances they initially did not understand outside of the verbal communication channel (or process of elimination). Similarly, speakers don't have direct access to how well their partners did in the previous block. Real-life communicative situations vary in what extra-textual feedback exists, but we do show that people can work around their initial confusion to eventually understand utterances, rather than just memorizing pairings after the first occurrence.

This is a rich data set consisting of 50000 words across 4000 referring expressions by 176 speakers, in addition to clarifications questions and comments from listeners. In this set of analyses, we rely on the easy to calculate measures of accuracy, speed, and word counts as proxies for the content of the utterances. In future analyses, it would be useful to do content analysis to understand how and when concepts are introduce and conventionalized and how

[3] Estimates and 95% credible intervals from a Bayesian linear model with formula: words ~ block * numPlayers + block * was_INcorrect + (block | tangram) + (1 | playerId) + (1 | tangram_group) + (block | gameId).

much the semantics of utterances varies block to block (and speaker to speaker) depending on group size.

We demonstrate that it is feasible to extend iterated reference game paradigms to small groups of participants using an online platform, and thus rapidly gather high quality utterance data from a number of games. We found that the widely observed pattern of partner specific adaptation and reduction extends to 3 and 4 person games. Inter-group variability suggests that a closer look at interpersonal communication dynamics, for example, comparing the semantic content of utterances of players in the same or different games is warranted. A closer analysis of the utterances may yield information about how humans adapt language quickly, and the dataset may be useful for training artificial agents to use and understand language more dynamically.

References

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: A virtual lab for high-throughput macro-level experiments. *Behav Res*. https://doi.org/10.3758/s13428-020-01535-9

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*.

Garrod, S., & Doherty, A. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions, 12.

Hawkins, R. D., Frank, M. C., & Goodman, N. D. (2020). Characterizing the dynamics of learning in repeated reference games. *arXiv:1912.07199 [Cs]*. Retrieved from http://arxiv.org/abs/1912.07199

Henry, L., & Wickham, H. (2020). *Rlang: Functions for base types and core r and 'tidyverse' features*. Retrieved from https://CRAN.R-project.org/package=rlang

Ibarra, A., & Tanenhaus, M. K. (2016). The Flexibility of Conceptual Pacts: Referring Expressions Dynamically Shift to Accommodate New Conceptualizations. *Front. Psychol.*, *7*. https://doi.org/10.3389/fpsyg.2016.00561

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*(2), 201–213. https://doi.org/10.1016/S0749-596X(03)00028-7

Müller, K. (2020). *Here: A simpler way to find your files*. Retrieved from https://CRAN.R-project.org/package=here

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv:1403.2805 [stat.CO]*. Retrieved from https://arxiv.org/abs/1403.2805

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Stan Development Team. (2020). RStan: The R interface to Stan. Retrieved from http://mc-stan.org/

Weber, R. A., & Camerer, C. F. (2003). Cultural Conflict and Merger Failure: An Experimental Approach. *Management Science*, *49*(4), 16.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Yoon, S. O., & Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 919–937. https://doi.org/10.1037/a0036161

Yoon, S. O., & Brown-Schmidt, S. (2019). Audience Design in Multiparty Conversation. *Cognitive Science*, *43*(8), e12774. https://doi.org/10.1111/cogs.12774