

How do conventions emerge in group communication: Evidence from 2-4 player reference games

Veronica Boyce¹

¹ Stanford University

Author Note

First Year Project for Psych Department

The authors made the following contributions. Veronica Boyce: .

Correspondence concerning this article should be addressed to Veronica Boyce, .
E-mail: vboyce@stanford.edu

Abstract

In repeated reference games where a speaker describes the same set of images to their listener partner over a series of rounds, the number of words used for each reference decreases as the pair converge on ad-hoc names for the images. The dynamics of this efficient reference formation is well-studied in two-person contexts; however much communication takes place in non-dyadic contexts, 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 words total, we find high accuracy and patterns of reduction regardless of group size.

Keywords:

Word count:

How do conventions emerge in group communication: Evidence from 2-4 player reference games

Intro

Verbal communication is an integral part of our daily lives. We coordinate schedules with individuals, socialize with friends over board games, learn and teach in seminar classes, and listen to podcasts. These different communicative environments range in size from one-on-one to small group to larger groups to broadcast communication, 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 need to be created to refer to new or specific things.

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. While the range of communicative situations is wide, 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), the target was a specific order of 12 tangram images which the speaker described to the listener so they could correctly order their copies. After receiving feedback, the pair did the task again with the same images but a new order, for a total of 6 repeats. Crucially, reference expressions condense over the course of repeated reference to the same image (Clark & Wilkes-Gibbs, 1986). Early descriptions are long, hesitant and make reference to multiple features or concepts in the image, but by later rounds the figures are often referred to by definite shorthand names. Recently, online participant recruitment and web-based experiments have made it possible to study this convergence to shorthand reference experiments 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 on how they did and which image was correct 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, confirming the common finding.

While this reduction pattern is robust for dyads, less is known about 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 and more disfluent utterances when any new listeners were present than with only experienced listeners, but there were also graded effects. This suggests that targeting utterances at an audience of mixed knowledge level is difficult. 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 that differ. 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 trying to communicate to both their original listener and the new listener. While over the initial rounds, the time taken to indentify the images declined, 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 this study suggests 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, p. @ibarraFlexibilityConceptualPacts2016).

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 recruited participants to play a repeated reference game in groups of 2-4. Participants viewed an array of 12 tangrams (Fig 1). One person was assigned the speaker role and say one of the images highlighted; the goal was for them to communicate the

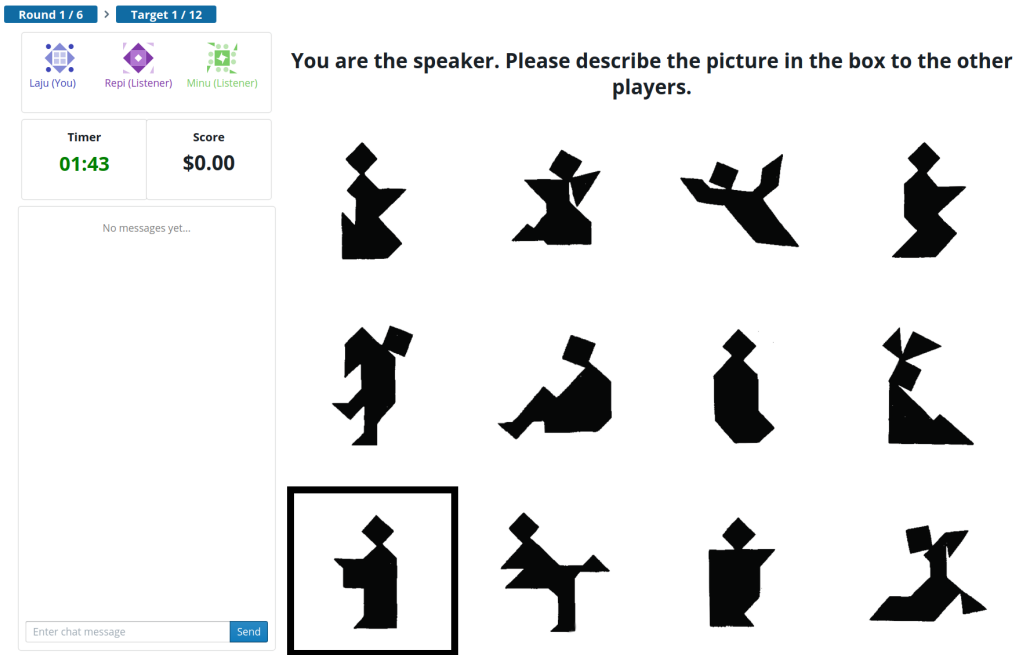


Figure 1. Screenshot of the speaker’s view. Participants see all 12 tangram images.

identity of this image to their partners who would then click on it. All participants were free to use the chat box to communicate. The speaker identified each of the 12 tangrams during a block; then the speaker role rotated to a different participant for the next block. Each group completed a total of 6 blocks, all with the same 12 images. We recorded what participants said in the chat, as well as who selected what image and how long they took to select.

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 intentional of a \$10 hourly rate), in addition to 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 included 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 for each participant with order randomized for each participant. The same images were used every round.

Procedure

We implemented the experiment using Empirica, a Javascript-based platform for running real-time interactive experiments online (Almaatouq et al., 2020). 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 that explained 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 it and click it. All participants were free to use the chat box to communicate. Listeners could only click once the speaker had sent a message. Once all listeners has 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.

We used R (Version 4.0.3; R Core Team, 2020)¹,
for all our analyses.

¹ We, furthermore, used 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

In general groups showed expected patterns. They had high and increasing accuracy, coupled with faster response times, and decreases in utterance length showing the classic reduction pattern.

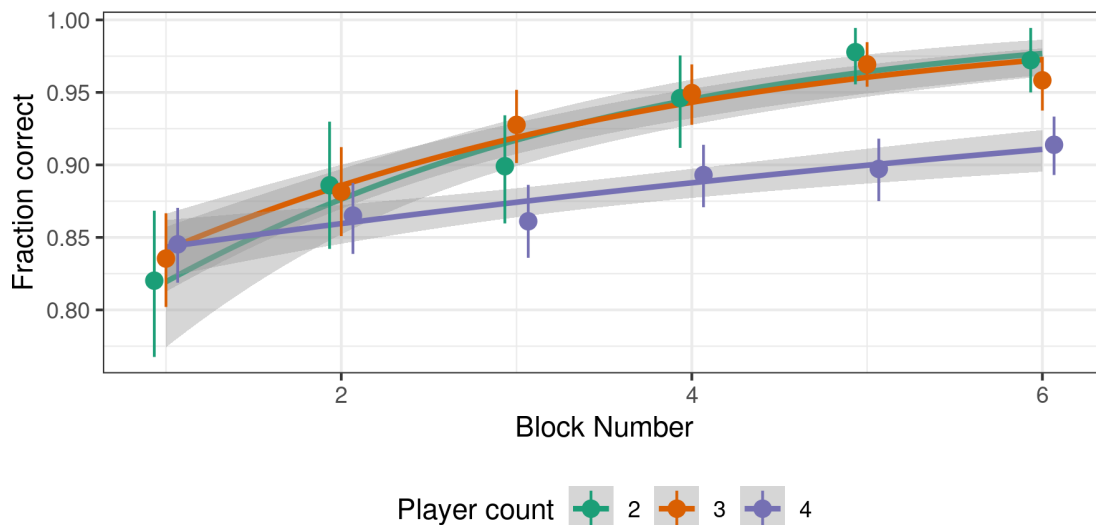


Figure 2. Listeners have high accuracy which increases of the course of the game, although accuracy increases less in 4-player games than smaller games. Accuracy rates are shown for each block, error bars are bootstrapped 95% CIs.

Most groups were accurate in their selections, with accuracy rising over rounds (Fig 2). This indicates that speakers were usually successful at conveying the intended referents. Participants are more accurate in later rounds 2 [1, 2].² 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.

Participants selected images faster in later rounds (Fig 3). There is wide variability, but appears to be an unintuitive effect of group size with fastest selection in 2-player games, but 4-player games being faster than 3-player games. This speed up is consistent with prior work by Weber and Camerer (2003) which used speed as the dependent measure.

The main effect of interest is whether speakers and listeners reduce in the number of words they say over the course of rounds. 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 rounds. 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

² Estimates from a Bayesian bernouilli model with formula: $\text{correct} \sim \text{block} * \text{numPlayers} + (\text{block} | \text{togram}) + (\text{block} | \text{playerId/gameId})$.

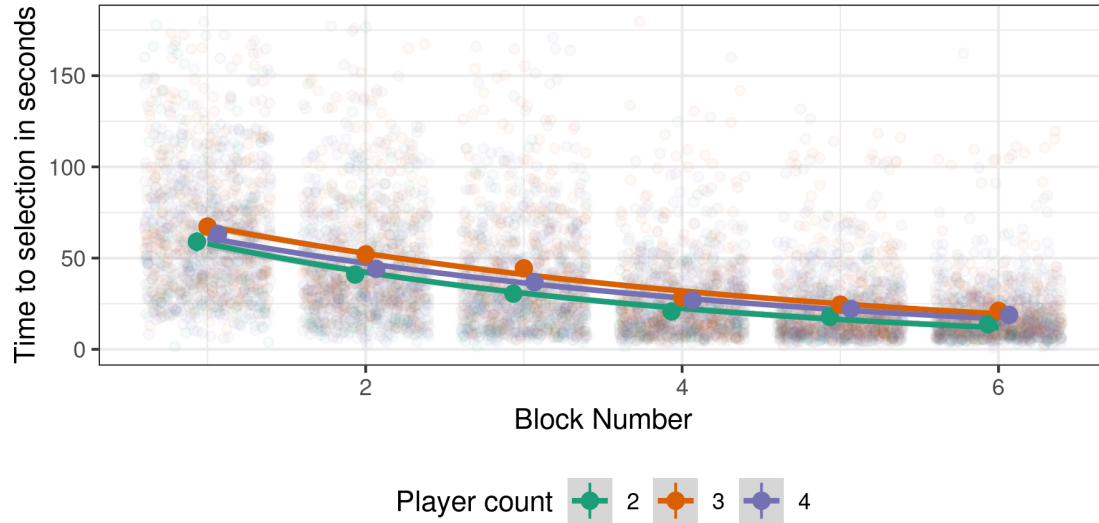


Figure 3. Listeners selected images faster in later blocks. Only times to correct responses are shown.

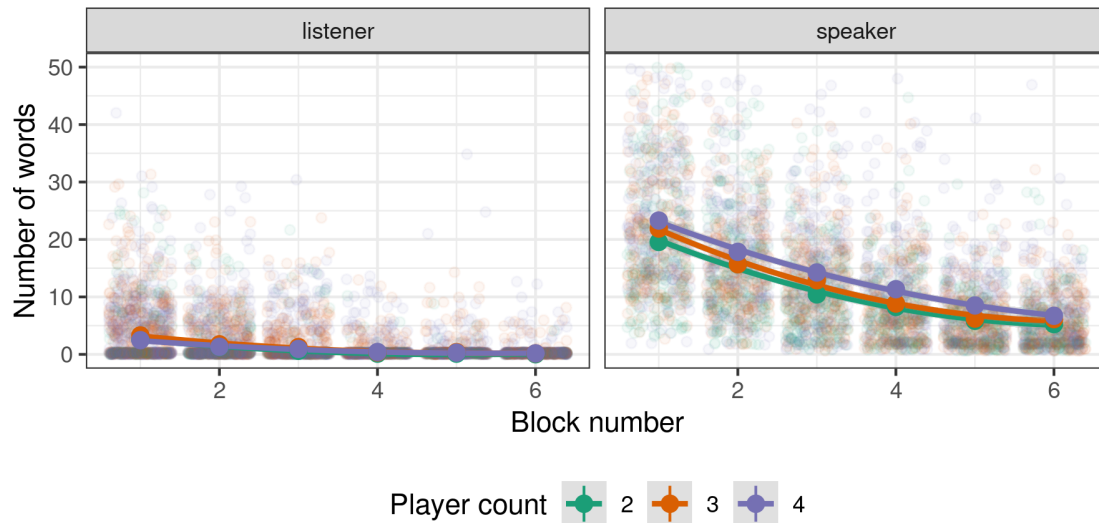


Figure 4. Speaker and listeners say fewer words in later blocks. Note: y-axis clipped at 50 which hides a few speaker outliers.

overall effect of having more players in a group is 1.93 [-0.15, 4.02] per additional player.³ 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

³ Estimates from a Bayesian linear model with formula: $\text{words} \sim \text{block} * \text{numPlayers} + (\text{block} \mid \text{tangram}) + (1 \mid \text{playerId}) + (1 \mid \text{tangram_group}) + (\text{block} \mid \text{gameId})$.

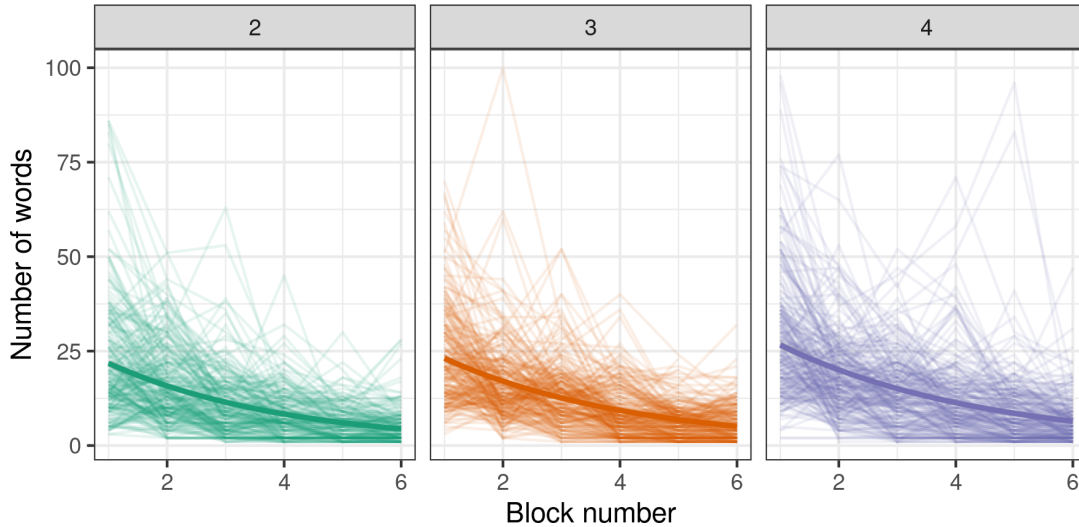


Figure 5. Words said by the speaker for each tangram in each group. Each referent/group trajectory is a thin line; aggregate curve is bolded. No outliers were omitted.

than were used in earlier rounds.

Because the ground truth answers are not provided to listeners who make mistakes, they may not learn what utterance was referring 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 descriptions and conventions that they don't understand. In contrast, a speaker who previously got the tangram right is reasonably likely to continue the conceptualization used so far and conventionalize it more, such as by reducing unneeded details. In aggregate hypothesizes that speakers should say more words if they got that tangram wrong the previous round than not, after controlling for other effects. This is borne out; speakers say 4.15 [2.54, 5.79] more words when previously wrong.⁴

Discussion

TODO effects/interpretation of speaker rotation and feedback regime

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. 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, p. @ibarraFlexibilityConceptualPacts2016). In rotating conditions, the interpretation of reduction is stronger evidence of conceptual agreement because more people have to agree to

⁴ Estimates from a Bayesian linear model with formula: words ~ block * numPlayers + block * was_INCORRECT + (block | tangram) + (1 | playerId) + (1 | tangram_group) + (block | gameId).

use the shorthand names, rather than just interpret them.

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 outside of the communication channel (and process of elimination) for people to find out what was meant for utterances they initially did not understand. Similarly, speakers don't have access to how well their partners did in the previous round (again, outside of explicit chat comments, or implicit assumptions drawn on the number of questions asked).

Real-life communicative situations vary in what extra-textual feedback exists, but we do show that people can work around their initial confusion rather than just memorizing pairings after the first round.

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 introduced and conventionalized and how 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 online platforms, and thus rapidly gather high quality utterance data from a number of games. We further found that the widely observed pattern of partner specific adaptation and reduction extends to 3 and 4 person games. The 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. (2020). Empirica: A virtual lab for high-throughput macro-level experiments. *arXiv:2006.11398 [Cs]*. Retrieved from <http://arxiv.org/abs/2006.11398>
- 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=rclang>
- 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](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/>

- 256 Stan Development Team. (2020). RStan: The R interface to Stan. Retrieved from
257 <http://mc-stan.org/>
- 258 Weber, R. A., & Camerer, C. F. (2003). Cultural Conflict and Merger Failure: An
259 Experimental Approach. *Management Science*, 49(4), 16.
- 260 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., . . . Yutani,
261 H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.
262 <https://doi.org/10.21105/joss.01686>
- 263 Yoon, S. O., & Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party
264 conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,
265 40(4), 919–937. <https://doi.org/10.1037/a0036161>
- 266 Yoon, S. O., & Brown-Schmidt, S. (2019). Audience Design in Multiparty Conversation.
267 *Cognitive Science*, 43(8), e12774. <https://doi.org/10.1111/cogs.12774>