Communicative reduction in referring expressions in a multi-player negotation game

Anonymous CogSci submission

Abstract

Reduction of referring expressions to short conventionalized terms is found robustly in dyadic repeated reference games when the target images are initially difficult to name. Establishing reference is a crucial part of many conversations with different goals and more complex motivations; thus, it is useful to look at how the findings from reference games, where participants share the explicit goal of establishing joint reference, carry over into situations where participants have different overarching goals. In the current work, we analyse a dataset where reference was embedded in strategic 3- player negotiation and coordination games. In these more complex games, we found that the patterns of reduction and convergence to conventions held across two different incentive conditions, with some suggestive differences between the conditions.

Keywords: Convention formation; reference games; reduction of referring expressions; multi-party communication;

Successful communication is grounded in a shared understanding of what utterances mean in the context they are produced in. In many cases conventional word meanings are enough, but there are also contexts where objects to be referred to are not easy to distinguish. In these situations, establishing reference is more difficult, but no less important.

The formation of these ad-hoc referring expressions has been studied extensively in the context of dyadic reference games (Clark & Wilkes-Gibbs, 1986, ADD MORE CITES). In these, two participants see a set of images (often abstract shapes) and one person is tasked with identifying an image so the other person can pick it out. People can be very successful at this task, achieving high levels of matches. Over repeated reference to the same images, pairs develop shorthand names for the images, leading to shorter descriptions in later repetitions. These nicknames conventionalize with a pair, as they tend to stick to the same description, but are idiosyncratic and differ between pairs.

This has been a well-studied microcosm for understanding reference which has proven useful for testing theories about how referring expressions originate, how expressions are designed, etc [TODO CITES]. The implicit theoretical claim is that these phenomena hold whenever people interact repeatedly in ways that require reference to some objects that do not have pre-conventionalized names that are adequate to distinguish them in context.

However, there are some ways in which this microcosm is not representative of typical language use. For one, in everyday communication, establishing reference rarely happens in isolation but is usually a subgoal in a larger conversational goal. The overall goal might be cooperative, such as in asking for a piece of kitchen equipment while cooking together, but it can also be adversarial, such as in a negotiation where the assets to be divied up must be referenced. In the reference game context, the goal is always cooperative, to match the images or get them in the same order.

TODO paragraph on on the issue of shared vs. privileged knowledge, which has been studied extensively in communication psycholinguistics (e.g., heller, keysar, barr) but IIRC not in the context of creating novel referring expressions in the tangram task. so we are blending in this other important element of naturalism, albeit only around the values and not the actual stuff?

Mankewitz et al. (2021) tried to bridge the gap between datasets of negotiation over easily nameable objects, and the reference game literature where there is reference (but no negotiation) over hard to name objects. They created a 3 person game where players each select what flowers to grow from a set of 6 flowers with variable and partially hidden value. Players got the value of the flower only if they were the only one to select it, which incentivized coordination and negotiation between players. However, all the flowers shown were the same color, so they were not easily nameable, and the flowers repeated, leading to opportunities for ad-hoc conventionalization. This game structure encouraged players to describe flowers while never explicitly saying that a player should refer to any specific flower. They also had two conditions, one where players within a group won points together (and thus had fully aligned incentives) and another where they won points as individuals (and didn't). Mankewitz et al. (2021) found a slight decrease in language over the course of the game, but they did not conduct a detailed analysis to assess whether the decreases in words were associated with shorter and more conventionalized referring expressions.

Historically, studies of convention formation have had to rely on proxies such as reduction and fluency, qualitatively assess similarity of utterances, or manually measure lexical overlap in small datasets to get at the ideas of convergence and divergence. More recent work has taken advantage of natural language processing tools for mapping words and sentences to semantic vector spaces in order to quantify semantic similarities between sentences Boyce, Hawkins, Goodman, & Frank (2022). We follow this recent work in using vector

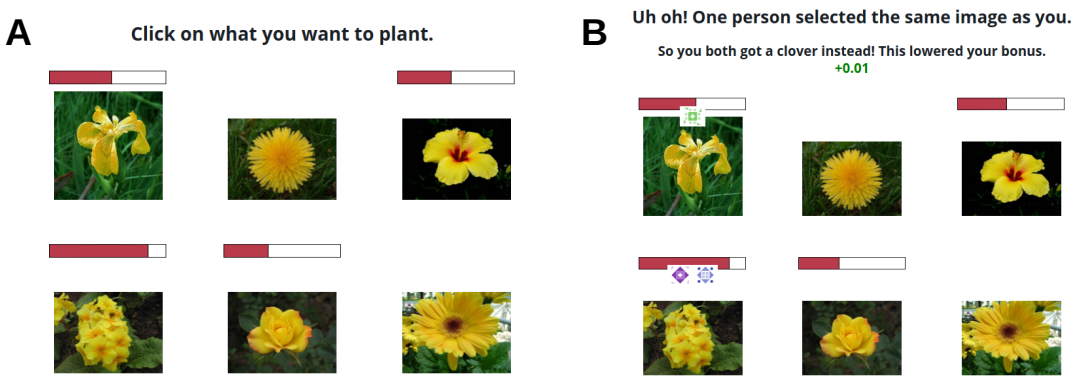

Figure 1: Player interface. Panel A shows the selection phase, where each participant sees 6 flowers, 4 with value bars. Panel B shows the feedback stage, after all players have selected. When multiple players select the same flower, they recieve a lower value rather than what is shown.

embeddings to measure semantic distances between referring expressions in the Mankewitz et al. (2021) dataset.

Current work

Here we ask whether participants in a complex, multi-player strategic game with differing incentives and shared knoweldge were able to develop conventionalized reference expressions. We explored whether the phenomena found in classic dyadic iterated reference games extended to this more naturalistic domain, and whether they differed across the individual and shared incentive conditions. The key questions we addressed were:

1. Do the referring expressions get shorter over time?
2. How do referring expressions change within and between groups over the course of the games? Do descriptions converge within groups and diverge between groups?
3. How do the different incentive conditions influence the referring expressions?

Methods

This paper presents a reanalysis of data from Mankewitz et al. (2021). We describe both the original data collection and the additional data processing done for reanalysis.

Procedure

As described in Mankewitz et al. (2021), participants played a real-time coordination game in groups of three, implemented using Empirica (Almaatouq et al., 2020). On each trial, each group saw a set of 6 flower images (Figure 1A) . Each participant saw the values for four of the flowers (represented as a colored bar), such that each flower's value was hidden from one participant. Players could coordinate and discuss using a chat box before each selected a flower. If one player selected a flower, it was worth the shown reward; if multiple players collided and selected the same flower, they each got a lower reward instead (Figure 1B). This incentivized participants to communicate about the flowers in order to coordinate on selecting differenct flowers. The rewards earned over the course of game translated in a monetary bonus for the participants at the end of the game.

In *individual utility* games, each player earned points for the flowers they selected; in the *shared utility* games, the points earned were averaged together, and all players in a game got the same reward. This made for slightly different incentives; in an individual game, players wanted to maximize the rewards of flowers they selected, and only cared about avoiding collisions with other players' selections; in a shared game, players wanted their teammates to select different high reward flowers, and were indifferent on who selected the highest one.

Each game was assigned a color of flower (white, red, yellow, purple) and the flower images were drawn from a set of 12 for that color, so players saw the same flowers repeatedly across the game, in different combinations. Each game consisted of 24 trials. The use of different flowers of the same color created situations where players did not have established names for the flowers in context and needed to develop shared referring expressions to clearly communicate with their partners.

After the game, players were asked how they would describe each of the images they had played with to their teammates. They were also asked how they would describe each of 4 images from a different color set than the one they had used.

Participants

Mankewitz et al. (2021) recruited 150 participants for 25 3-person games in each of the individual and shared utilities conditions where they had access to a chat box (there were also no-chat-box baseline conditions, which we did not analyse). Some games did not complete the full 24 rounds because participants dropped out early. We excluded incomplete games from analyses, which left us with 18 games in

the individual utilities condition and 21 in the shared utilities condition (117 participants included in the analyses).

Textual annotations

We annotated the chat transcripts to extract all referring expressions and identify which flower image each expression referred to. We also spellchecked and corrected the referring expressions. Annotations were done primarily by the first author, with some done by a research assistant.

Two games consisting of 121 utterances were annotated by both annotators. 117 utterances were identified as containing reference expressions by at least one annotator; of these annotators agreed on the exact reference expression for 105 (90%) of the cases. The second annotator coded the target of the referring expression for 60 of the utterances. Of these, the two annotators agreed on the target in 59 (98%) of cases. We take this high level of interannotator agreement as an indication that the reference spans and targets were identified in a reliable way.

We extracted a total of 3395 referring expressions.

SBERT embeddings

Whereas in the past studies of convention formation have had to rely on manual assessments of similarity or lexical overlap, natural language processing tools now make it possible to quantitatively measure the similarity between phrases. Recent work by (**Hawkins2020?**) and Boyce et al. (2022) have used these tools to assess the dynamics of convergence to conventions.

We embedded each of the extracted referring expressions using SBERT (Reimers & Gurevych, 2019) which maps each utterance to a vector in a high-dimensional semantic vector space. We use cosine distance between pairs of vector embeddings as a measure of similarity between the corresponding referring expressions. When two vectors are near each other, the angle between them is small, and the cosine of this angle is large. Higher cosine values correspond to vectors close together, and thus, reference expressions with more similar meanings.

When doing pairwise comparisons between referring expressions produced during the game, we smoothed out the random sampling and sparseness of which flowers were presented on the same trial by comparing descriptions with those from the same round and the two prior rounds (thus the similarity at round 10 comes from 8-10, 9-10 and 10-10 comparisons). This smoothing was not necessary when comparing with the names produced after the game.

Analytic methods

TODO this is the section where we talk about modelling details

Results

We address each of the questions of interest listed at the end of the Introduction in order.

Reduction of referring expressions

Each round of the game, there were 6 flowers visible on the screen, and participants wanted to each pick a different one. In general, each trial contained references to 2 or 3 distinct flowers (Individual Utilities: mean of 2.3 (sd: 1.34), Shared Utilities: 2.33 (sd: 1.12)). Most players referred to one flower each round (Individual Utilities: 1.04 (sd: 0.92), Shared Utilities: 0.97 (sd: 0.78)).

The amount of referring language decreases over the course of the game, consistent with the dyadic reference game pattern. As shown in Figure 2A, the number of words of referring language decreases across the game (trial: -0.154 CrI=[-0.165, -0.144]. The reduction in referring language was driven by shorter referring expressions later in the game (Figure 2B), while the total number of referring expressions per round remained constant (Figure 2C).

Convergence of referring expressions

The classic reference game phenomenon of convergence to partner-specific referring expressions can be quantified using similarity metrics as the combination of three patterns occurring over time: 1) within a group, descriptions of the same flower converge, 2) within a group, descriptions of different flowers diverge, and 3) between groups, descriptions of the same flower diverge. We assess all of these in two ways: 1) comparing referring expressions within the game to other referring expressions from the same timepoint to get a sense of the dynamic changes, and 2) comparing referring expressions within the game to the names players gave the flowers after the game, to track how referring expressions evolve towards these final conventions.

TODO table with examples of flower descriptions over time and cos similarities !!!

Within games Within games, the patterns from dyadic reference games predict that similarity for descriptions of the same image will increase over time while similarity for descriptions of different images will decrease over time.

We found that for descriptions of the same flower, cosine similarity increased over time (trial: 0.009 CrI=[0.007, 0.011], see Figure 3 upper left and middle panels). Utterances were more similar if they were produced by the same participant (0.062 CrI=[0.02, 0.104]), although this did not interact with trial number (-0.002 CrI=[-0.006, 0.001]).

For descriptions of different flowers, cosine similarity decreased over time (trial: -0.006 CrI=[-0.007, -0.006], see Figure 3 lower left and middle panels). There was not an effect of utterances being produced by the same person (0.003 CrI=[-0.008, 0.014]), or interactions between who said it and trial number (0 CrI=[-0.001, 0.001]) .

These results mirror are consistent with the reference game phenomenon.

Between games The theory of partner-specificity predictions that descriptions to the same flower in different groups will diverge over time as each group forms their own group-

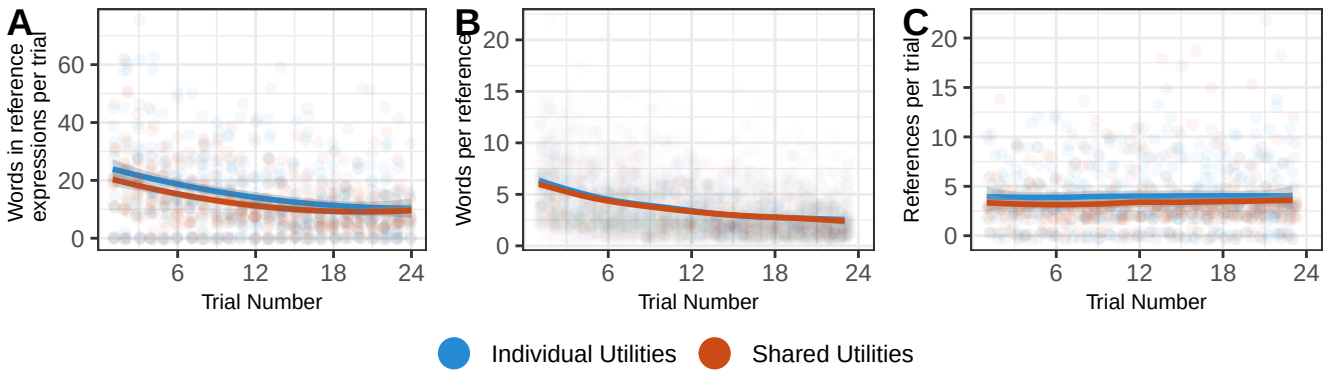

Figure 2: Amount of words produced in referring expressions across trial in games in both conditions. A: Total words of referring language per game per trial. B: Words in each reference expression by trial. C: Total number of reference expressions per game per trial.

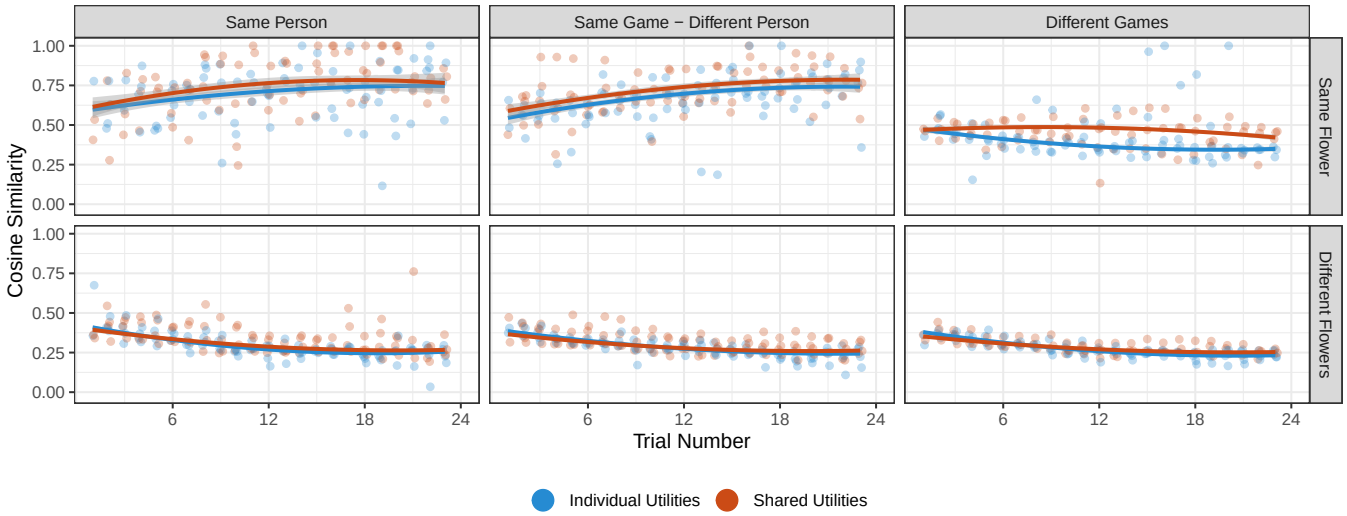

Figure 3: Cosine similarities between SBERT embeddings of utterances produced within 2 trials of each other. Utterances were paired with other utterances from different games; from the same game, but a different utterer speaker; or from the same speaker (vertical panels). Utterances were either in reference to different flowers or both in reference to the same flower (horizontal panels).

specific convention.

For descriptions of the same flower across games, we found that cosine similarities decreased slightly over time (trial: -0.005 CrI=[-0.006, -0.004], see Figure 3 upper right panel). This is weakly consistent with the predictions.

Comparison with post-game names After the game, players were asked to provide the description they would use to identify each flower to their teammates. We treated these descriptions as the conventionalized names and looked at how the in-game descriptions developed into these conventions.

Referring descriptions later in a game were more similar to the convention than descriptions earlier in the game (trial: 0.006 CrI=[0.005, 0.007], see Figure 4 upper panels). Utterances were more similar to the convention given by the same person (versus the convention given by a groupmate) (0.053 CrI=[0.023, 0.083]), although this did not interact with trial number (0.003 CrI=[0.001, 0.005]). This is consistent with the expected pattern of convention formation.

We also examined how referring expressions related the conventionalized names for other flowers. Utterances did not significantly diverge from the conventionalized names of other flower over time (trial: -0.003 CrI=[-0.003, -0.003], see Figure 4 lower panels). There was not an effect of being said by the same person who gave the convention (0.001 CrI=[-0.005, 0.007]), or interactions between who said it and trial number (0 CrI=[0, 0.001]).

Individual and Shared Utility conditions

The last question of interest is whether there were differences between the individual and shared utility conditions. Here we report the extent to which the shared utility condition differed from the individual utility condition.

There was not a significant difference in the amount of re-

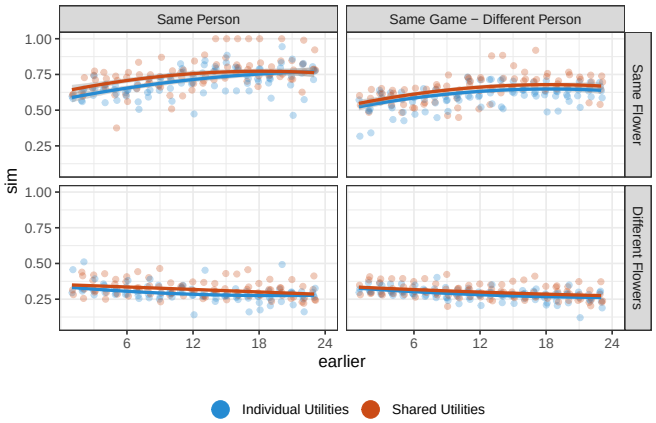

Figure 4: Cosine similarities between SBERT embeddings of utterances produced during a game and the post-game descriptions of the flowers.

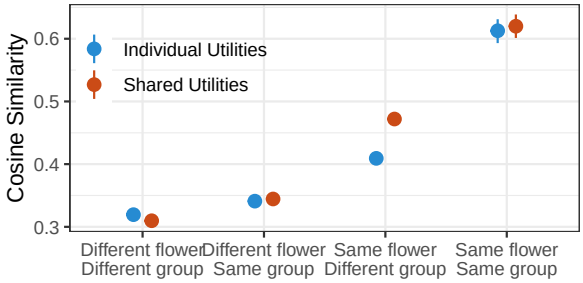

Figure 5: Cosine similarities between flower descriptions provided by participants after the game. Descriptions were more similar to each other if they came from the same group and were of the same flower.

ferring language produced in the two conditions (0.076 CrI=[-0.661, 0.808]).

Within games Games in the shared utilities condition had slightly higher cosine similarities between descriptions of the same flower within games (0.056 CrI=[-0.023, 0.131]), but this did not interact with trial number (0 CrI=[-0.003, 0.003]) or whether the utterances were produced by the same participant (-0.014 CrI=[-0.077, 0.051]). As a complement to this, games in the shared utilities condition had slightly lower similarities between descriptions of different flowers within games, but this did not interact with trial (0.002 CrI=[0.001, 0.002]) or whether the utterances were produced by the same player (0.016 CrI=[0.001, 0.032]). Taken together, groups in the shared utility had more internal alignment on the mapping between flowers and names.

Comparing to the end conventions, games in the shared utility condition produced flower descriptions that were more similar to the convention for that flower (0.032 CrI=[0.007, 0.056]), but this did not interact with trial number (0 CrI=[-0.002, 0.002]) and did not significantly interact with same participant versus different participant (0.032 CrI=[-0.011, 0.075]). In the shared utility condition, descriptions were slightly less similar to the conventions for other flowers (0.016 CrI=[0.011, 0.021]), but this did not interact with trial (0 CrI=[0, 0]). There was an interaction between shared utility condition and the referring expression coming from the same person who gave the convention for the different flower (0.019 CrI=[0.009, 0.027]).

Taken together, groups in the shared utility had more internal alignment on the mapping between flowers and names.

Between games Across different games, games in the shared utility condition tended to have more similar descriptions for the same flower (0.052 CrI=[0.034, 0.071]); this did not interact with trial (0.003 CrI=[0.002, 0.005]).

Lastly, we can look at how the conventionalized names compared, both between members of the same group and across players in different groups that saw the same color palette of flowers, and whether this differs by condition.

We treat different flowers described by different games as the baseline condition. For individual utility games, descriptions from groupmates are more similar (even for different flowers, 0.022 CrI=[0.016, 0.027]), descriptions of the same flower are more similar (even across games, 0.09 CrI=[0.083, 0.097]), and there is a large interaction effect, where descriptions of the same flower are very similar among groupmates (0.182 CrI=[0.164, 0.199]).

For shared utility games compared to individual utility games, descriptions are roughly equivalent at baseline (-0.01 CrI=[-0.012, -0.007]) and for descriptions from groupmates (0.013 CrI=[0.006, 0.02]). Descriptions are much more similar for the same flower (even across games, 0.072 CrI=[0.064, 0.081]), but this is balanced out by a negative interaction effect between same flower and same game (-0.068 CrI=[-0.092, -0.045]).

These findings are suggestive that in the shared utility condition, games differentiated less from each other, and their conventions may have been influenced more by shared priors for how to describe the flowers.

Discussion

This analysis of the referring expressions in Mankewitz et al. (2021) examined three key questions: the reduction of referring expressions, the convergence of referring expressions to conventions, and the differences between the shared and individual utility conditions.

First, we saw reduction, as referring expressions tended to decrease in length over the course of the game. Second, we found the expected trifecta of convergence and divergence results. Over the course of games, references to the same flower within a game converged both to other descriptions used at that timepoint and to the post-game descriptions. This convergence was specific to flower and games, as references to different flowers within a game and references to the same flower across games both diverged.

These results align with the expected pattern of partner-specific convention formation established in dyadic reference games. These patterns of reduction and convergence con-

firm some generalizability of the reference game patterns to the freer form and more naturalistic domain of a negotiation game. While still an artificial situation, this game differs from classic reference games in that the set-up was symmetric, with all players having equivalent amounts of knowledge and authority, and that the need for reference was embedding within a more complex goal structure. This game also was aytpical for reference games in that it had more than 2 players (Boyce et al., 2022; although see Yoon & Brown-Schmidt, 2019) and used photos as target images (although see Weber & Camerer, 2003).

Our last consideration was how these reduction and convergence effects might be moderated by the shared and individual utility conditions. We found slight condition differences where being in the shared utility condition led to stronger alignment between players throughout. We hypothesize that they might be due to a perceived greater importance of communicating with each other or paying greater attention to what groupmates are seeing. In addition, groups in the shared utility condition diverged less from each other. This reduced divergence suggests that their descriptions and conventions were less ideosyncratic and perhaps closer to shared underlying priors about how to describe the images. We hypothesize that the explicitly more cooperative goal might have led to participants caring more about being understood by their partners. This is interesting in light of Guilbeault, Baronchelli, & Centola (2021) finding that larger groups tend to all converge to similar conventions whereas smaller groups can support ideosyncratic conventions.

However, we are cautious not to over-interpret the condition differences. There were a limited number of groups in each condition, so these effects may have been driven or exacerbated by heterogeneity in which participants were in each group. The differences are suggestive, but these ideas will need to be tested further with larger samples and more different goal structures.

Our interpretation of results is limited by the dataset, which is small, especially compared to recent datasets of convention formation in reference games Boyce et al. (2022). The dataset also has large item-level variation, as the images differ in how nameable the flowers are or whether their are salient features.

In conclusion, this analysis showed that the phenomenon of reduction and conventionalization occurs even in complex games with different incentives, asymmetric knowledge, and more open-ended negotiation and dialogue than that found in reference games. Additionally, we saw hints that different incentives may lightly alter moderate the formation of conventions, but that reduction and convention formation took place regardless of incentive condition.

Acknowledgements

Place acknowledgments (including funding information) in a section at the end of the paper.

References

10 Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2020, December 30). *Empirica: A virtual lab for high-throughput macro-level experiments*. Retrieved from http://arxiv.org/abs/2006.11398

Boyce, V., Hawkins, R., Goodman, N., & Frank, M. (2022). Two's company but six is a crowd: Emergence of conventions in multiparty communication games. Retrieved January 16, 2023, from https://escholarship.org/uc/item/3gd9j28x

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Retrieved from http://www.speech.kth.se/~edlund/bielefeld/references/clark-and-wilkes-gibbs-1986.pdf

Guilbeault, D., Baronchelli, A., & Centola, D. (2021). Experimental evidence for scale-induced category convergence across populations. *Nature Communications*, *12*(1, 1), 327. http://doi.org/10.1038/s41467-020-20037-y

Mankewitz, J., Boyce, V., Waldon, B., Loukatou, G., Yu, D., Mu, J., ... Frank, M. C. (2021). *Multi-party referential communication in complex strategic games* (preprint). PsyArXiv. Retrieved from https://osf.io/tfb3d

Reimers, N., & Gurevych, I. (2019, August 27). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv. http://doi.org/10.48550/arXiv.1908.10084

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

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