Signaling social identity in referential communication

Alicia M. Chen[1], Robert D. Hawkins[2], and Rebecca Saxe[1]

[1]Department of Brain and Cognitive Sciences, MIT

[2]Department of Linguistics, Stanford University

Author Note

Alicia M. Chen https://orcid.org/0000-0001-5521-8566

Robert D. Hawkins https://orcid.org/0000-0001-9089-8544

Rebecca Saxe https://orcid.org/0000-0003-2377-1791

All materials, code, and de-identified data are available at this repository. All experiments were preregistered; the preregistrations are available at `https://osf.io/5j6uk/`. This work was supported by the Simons Center for the Social Brain. We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Alicia M. Chen, Department of Brain and Cognitive Sciences, MIT, 43 Vassar St., Cambridge MA 02139, United States. Email: `aliciach@mit.edu`.

Abstract

Any choice of words simultaneously conveys information about the world and, at the same time, conveys information about the speaker, revealing aspects of their social identity. In this paper, we investigate how speakers strategically modify referential language to signal group membership. Across four experiments using a minimal referential communication paradigm, we find that speakers with the explicit goal of signaling social affiliation (1) choose more concise utterances, (2) preferentially select group-specific referents and descriptions, and (3) resist the otherwise strong tendency to be understood by everyone in the audience. Standard models of referential communication that focus on the trade-off between informativity and efficiency cannot explain these patterns; we argue instead for a model where speakers trade off the referential utility of being understood against the social utility of being identified as an in-group member.

Keywords: social cognition, communication, audience design, computational modeling

Significance Statement

When a group of people interacts repeatedly – from friend groups, to classrooms or schools, to larger online communities – they develop shared ways to refer to concepts and objects. Consequently, speakers can use these expressions as social identity markers, signaling group membership to other group members. In controlled experiments and a formal model, we study how speakers strategically use group-specific terms, similar to slang or jargon, to identify themselves as insiders. This research helps to explain how people use linguistic choices to create and maintain social boundaries between different social groups.

Signaling social identity in referential communication

Human communication is shaped by multiple considerations. Perhaps the most basic consideration is referential efficiency — the balance between referential informativity (ensuring that the audience understands what is being referred to) and parsimony (minimizing communicative effort). A balance between these two goals can explain many aspects of language evolution, processing, and use (e.g., Zipf, 1949; Frank & Goodman, 2012; Kemp & Regier, 2012; Kanwal, Smith, Culbertson, & Kirby, 2017; Zaslavsky, Kemp, Regier, & Tishby, 2018; Gibson et al., 2019; Jara-Ettinger & Rubio-Fernandez, 2022). The influence of these goals is especially apparent in the idiosyncratic group-specific conventions that emerge when people interact with each other repeatedly (Hawkins et al., 2023). Over time, groups build shared systems of meaning that allow them to use referential expressions that are both informative and parsimonious (Krauss & Weinheimer, 1964; Clark & Wilkes-Gibbs, 1986; Clark & Marshall, 1981; Brennan & Clark, 1996).

Notably, when people *interpret* language, they make inferences about the speaker that go beyond identifying the intended referent. Linguistic variation is an especially reliable and observable signal of social groups and communities (Labov, 1973; Eckert, 1989; Lupyan & Dale, 2016; Kinzler, 2021), and so one important type of inference concerns the speaker's group membership or social identity (e.g., Gumperz, 1982; Bucholtz & Hall, 2005; Kinzler, Dupoux, & Spelke, 2007; Rhodes, Leslie, Bianchi, & Chalik, 2018; Smaldino, 2022). While social identity influences many aspects of language (including syntax, grammatical constructions, and accent; Goffman, 1956; Labov, 1966; Scherer & Giles, 1979; Giles, Coupland, & Coupland, 1991; Bucholtz & Hall, 2005; Fedzechkina, Hall Hartley, & Roberts, 2023; Camp & Nowak, 2025), one way listeners make inferences about speaker group membership is from the speaker's word choice in reference. After hearing only brief verbal descriptions from a speaker, and even without explicitly referring to group membership, listeners infer the kinds of communities speakers belong to (Isaacs & Clark, 1987; Walker, Fugelsang, & Koehler, 2025). For example, in one experiment, people described New York City landmarks to each other. New Yorkers could reliably identify a compatriot New Yorker from a single description (e.g., "Washington Square Park" vs. "the fountain with the arch in the background"; Isaacs & Clark, 1987).

The way social identity affects language can be unconscious, or it can be deliberate and strategic (Gumperz, 1982; Bell, 1984; Myers-Scotton, 1998; Woolard, 2004). Just as listeners can infer social identity from language, speakers can intentionally and strategically shape their word choice to signal and emphasize these identities (e.g., Le Page, 1968; Burnett, 2023; Gibson et al., 2019). For instance, a speaker might use a concise term like “rizz” (short for “charisma”) for efficiency, but also to reveal their membership in the communities that use it (Labov, 1973). Field-specific jargon or group-specific slang can signal social affiliation at the expense of being broadly understood (Brown, Anicich, & Galinsky, 2020; Damirjian, 2025). Speakers may choose topics or referents that are uniquely discussed with their in-group, or concepts that are especially salient within their community (Eckert & Brown, 2006; Holmes & Meyerhoff, 1999). In this way, signaling social identity may compete against the tendency to ensure that everyone in a mixed audience can understand what the speaker is saying (S. O. Yoon & Brown-Schmidt, 2018, 2019). Linguistic choices that emphasize social identity can thus lead to negative consequences such as misunderstandings or the exclusion of non-group members (Eckert, 1989; Bullock, Colón Amill, Shulman, & Dixon, 2019; Martínez, Mollica, & Gibson, 2022; Cruz & Lombrozo, 2025), as well as broader polarization in society (Walker et al., 2025).

In this paper, we characterize the influence of strategic social goals in referential language, by integrating two previously separate research traditions. A standard experimental and computational framework in which groups develop arbitrary conventions for talking about abstract shapes (Clark & Wilkes-Gibbs, 1986; Hawkins et al., 2023), has been used to quantify the balance between informativity and parsimony in referential communication, but has largely ignored social signaling motivations. Meanwhile, sociolinguistic research has richly documented social identity signaling in everyday speech, but has not used controlled experiments to quantify its influence compared to other factors in referential word choice. We bridge these traditions by extending the reference game paradigm used by the former tradition, to additionally manipulate speakers’ explicit social signaling goals. We use this paradigm precisely because it provides the minimal conditions for a tension to arise between goals in word choice. If social affiliation systematically competes with efficiency even in minimal experimental contexts, then the existing models, which formalize the strategic decision-making process underlying how people choose referential language, are missing a core driver of language production. By quantifying this trade-off experimentally and formalizing it in a computational model, we can begin to measure how referential and social signaling goals jointly shape even the most basic settings of language use, providing a foundation for understanding more complex real-world cases where these pressures are amplified by more deeply-held social identities and consequences.

To test how speakers navigate these pressures, we developed a computational model that formalizes utterance choice as a graded trade-off between referential clarity and group identification. The experiments are designed to measure the model's predictions. Participants first observe different groups developing their own linguistic conventions in a reference game (Boyce, Hawkins, Goodman, & Frank, 2024); next, the participants themselves must decide what to refer to and how to refer to it, based on their audience (people in the observed groups and/or naive players) and their goal (to refer effectively and/or signal social affiliation). This design preserves experimental control – we experimentally assign which groups talk about which referents, what words they use, and what the speakers' goals are – while maintaining ecological validity, as participants observe real conversations produced naturally by human speakers interacting in groups over time.

In Experiment 1, we find that people selectively choose what to talk about and what to say based on their audience and goal. Then, in Experiments 2-4, we isolate specific comparisons to measure the precise contributions of social signaling on people's choices. The results show that people with a social signaling goal choose referents that afford group-specific labels (Experiment 2), select more concise group-specific labels (Experiment 3), and resist the otherwise strong tendency to prioritize the least informed audience member (Experiment 4). We compare a baseline referential model — capturing the standard trade-off between informativity and effort — to an augmented social signaling model that includes an additional utility for in-group identification. While the baseline model accounts for referential efficiency-driven choices, the social model is needed to explain behavior when participants choose utterances to achieve a social goal.

Transparency and openness

All de-identified data, materials, and code are available at https://github.com/aliciamchen/code-switching and also deposited on OSF at https://osf.io/5j6uk. The sample sizes, design, and analyses for all experiments were all preregistered before data collection. The preregistration for Experiment 1 is available at https://osf.io/uzsc2, the preregistration for Experiment 2 is available at https://osf.io/jyktz, the preregistration for Experiment 3 is available at https://osf.io/z27eb, and the preregistration for Experiment 4 is available at https://osf.io/h4trx. There were minor deviations from the

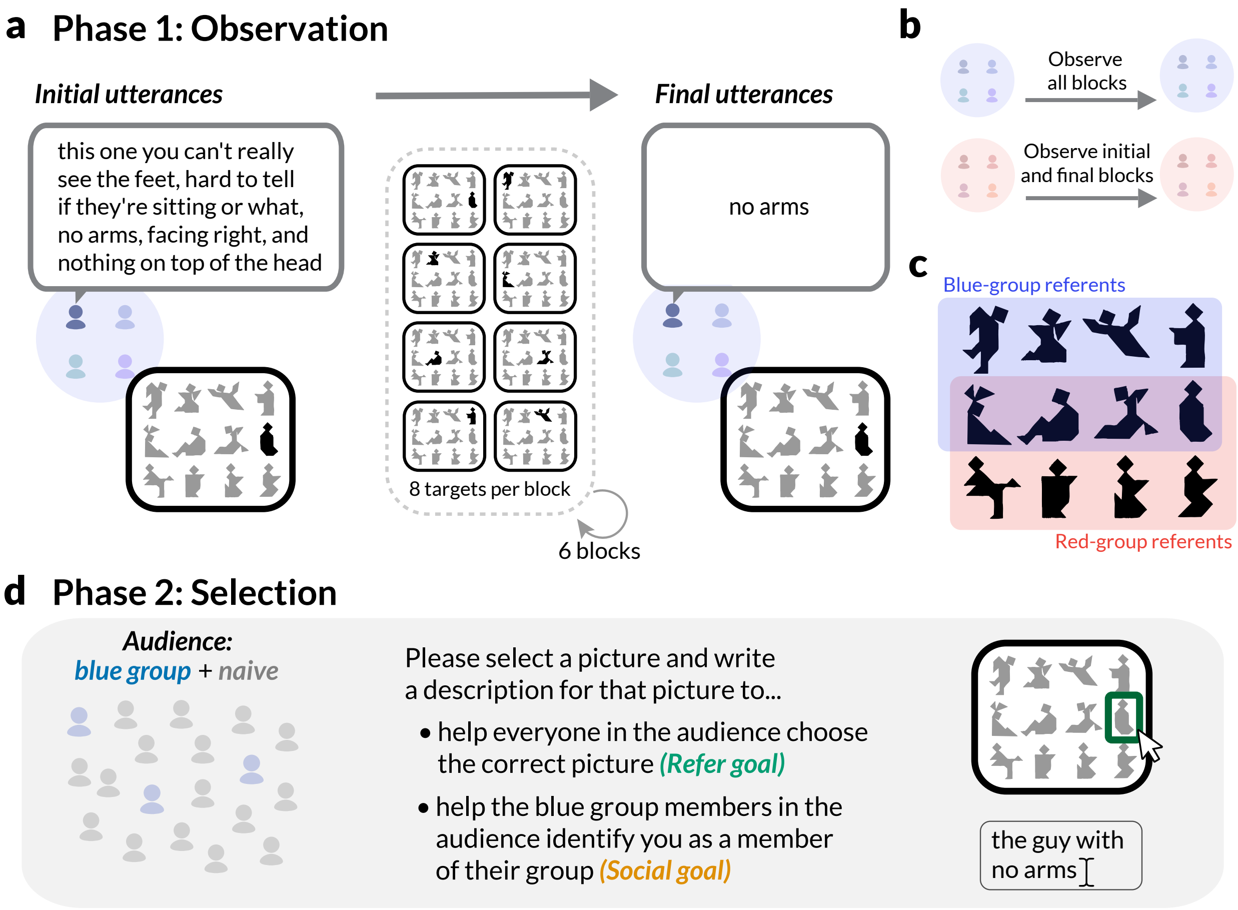
a Phase 1: Observation
Initial utterances
this one you can't really see the feet, hard to tell if they're sitting or what, no arms, facing right, and nothing on top of the head
8 targets per block
6 blocks
Final utterances
no arms
b
Observe all blocks
Observe initial and final blocks
c
Blue-group referents
Red-group referents
d Phase 2: Selection
Audience: blue group + naive
Please select a picture and write a description for that picture to...
• help everyone in the audience choose the correct picture (Refer goal)
• help the blue group members in the audience identify you as a member of their group (Social goal)
the guy with no arms

Fig. 1

*Experiment 1. **(a)** In the first phase of the task, participants were third-party observers viewing a group of four other people playing a multiparty reference game. **(b)** Participants watched the full convergence process of the blue group, and only watched the initial and final blocks for the red group. **(c)** Each group referred to eight of the twelve tangrams. Four of the tangrams were only referred to by the blue group, four only by the red group, and four by both groups. **(d)** In the second phase of the task, participants were shown an audience and a goal, they selected a tangram to refer to, and wrote a description for that tangram.*

preregistrations, described in Appendix B. AI tools were used to aid the generation of analysis and plotting scripts and to improve the readability of the manuscript.

Experiment 1

In Experiment 1, we conducted an initial test of whether people track and deploy group-specific referential conventions for social signaling, in the content of experimental reference games. In the first phase of the task (the ‘observation phase’), participants viewed two groups taking turns writing descriptions to help the other members of their group identify abstract tangram shapes (Figure 1; Clark & Wilkes-Gibbs, 1986). Group-specific conventions naturally emerged over the sequence of interactions (Boyce et al., 2024), such that the two observed groups differed in which tangrams they referred to and the expressions they developed for those tangrams. In the second phase of the task (the ‘selection phase’), we introduced the critical manipulation of communicative goals. Participants were shown an audience composed of members of one of the groups along with naive players who had not played the game before, and were asked either to pursue a purely referential goal (help their audience identify which tangram was being referred to) or a social-signaling goal (help the group members in the audience identify the speaker as a member of their group).

This design allows us to test three key questions about how social signaling operates in referential communication. First, to what extent do participants spontaneously encode and track group-specific linguistic patterns during passive observation? Second, when given a social goal, how do participants identify which specific linguistic choices – among an open-ended set of possibilities – would effectively signal group membership? Third, how do speakers weigh social signaling against referential clarity when these goals conflict? Critically, our focus is not simply on whether participants can follow instructions to signal group identity, but rather *how* they do so: whether they independently recognize which referents and expressions carry social meaning, assess their relative signaling value, and calibrate their use based on audience composition. If participants make systematic trade-offs even without explicit guidance about the social meaning of different choices, this would provide strong evidence that people spontaneously track the language of minimal groups and strategically reason about audience composition to decide what kind of language to deploy.

Methods

Participants

For all experiments, participants were recruited on Prolific and pre-screened to be adult fluent English speakers from the United States, and were excluded if they indicated at the end of the experiment that they did not understand the instructions. For Experiment 1, we recruited 185 (93 female; ages 20-72, M(SD) age = 40.2(11.7)) participants, and no participants were excluded. Participants were paid $12 for completing the experiment, which took an estimated 40 minutes to complete. Participants gave informed consent, and all procedures were approved by the MIT Committee for the Use of Humans as Experimental Subjects.

Design and stimuli

Phase 1: Observation phase. In the *observation phase* (Figure 1a), participants first learned each group's referential conventions through passive observation, without being told what they would do in the second phase of the task. They watched videos showing interleaved rounds of communication by two groups (arbitrarily named the 'red' and 'blue' groups) composed of four people each (see `https://osf.io/f9dxb` for an example round). The groups played the reference game within their group. In each round of the reference game, one player (the 'speaker') described the target tangram in a communal chat; the three other players (the 'listeners') each made a guess by selecting one tangram. Players received feedback on each of the listeners' guesses at the end of the round. Each video corresponded to one round of gameplay and ranged in duration from approximately 5 to 30 seconds.

The gameplay in each group consisted of 6 blocks (matching the four-person 'full feedback' condition described in Boyce et al., 2024). In each block, one speaker described each of the eight tangrams once, in randomized order, before the speaker role rotated to the next player and the next block began. Over the rounds, participants observed expressions converge to shorter, more abstract, and group-specific labels (e.g. "this one you can't really see the feet, hard to tell if they're sitting or what, no arms, facing right, and nothing on the top of the head" becomes "no arms"; see Figure 1a). The task focused on one of the groups (always set to the blue group for simplicity): Participants viewed the full convergence history (6 blocks x 8 targets per block = 48 rounds) for the blue group, but only saw the red group's initial and final blocks (2 blocks x 8 targets per block = 16 rounds; see Figure 1b). Of the total 12 tangrams, four of the tangrams were talked about only in the blue group ("blue-specific referents"), four of the tangrams were talked about only in the red group ("red-specific referents"), and four of the tangrams were talked about in both groups ("common referents"; see Figure 1c). Within each set, the two groups converged on distinct, group-specific labels for all tangrams, ensuring that no overlap in referential labels occurred between groups. Across participants we counterbalanced the assignment of individual tangrams to the "red" versus "blue" groups.

Videos were created by extracting individual conversations about specific tangrams from a large corpus of online multiplayer reference games (Boyce et al., 2024) and splicing them together to satisfy the task's design criteria. Because each round of the repeated reference game is a self-contained conversation centered on a single referent, these spliced sequences appeared natural and coherent to participants.

Phase 2: Selection phase. Participants then moved to the *selection phase*, in which participants selected a tangram and freely produced a referring expression (Figure 1d). On each trial, participants were given a goal and audience, chose one tangram out of the 12 possible tangrams, and then wrote a description of that tangram in a text box. The audience and goal were manipulated fully within-subject. The audience varied continuously in size and composition of in-group members, and was shown to participants as an iconic depiction from 0 to 4 blue-group avatars and 0, 1, 2, 4, 8, or 16 naive avatars. For the goal, participants were told what kind of question the audience would be answering (the 'referential goal' of choosing the best tangram given the label, or the 'social signaling goal' of being identified as a member of their group; Figure 1d)[1].

There were 53 trials in total. To incentivize performance, participants were nominally offered a bonus of $0.10 for each 'correct' response; in practice, all participants were given a $6 bonus and debriefed after the experiment that there were no objectively 'correct' answers. To avoid the trivial strategy of repeatedly selecting the same tangram — which would reduce the variety of communicative choices we could observe — participants were told that they could not select the same tangram they had chosen on the immediately preceding trial.

Implementation details

The experiments were created using jsPsych (de Leeuw, Gilbert, & Luchterhandt, 2023) using the DataPipe tool to automatically send the data to the Open Science Framework (de Leeuw, 2024). Data analysis was performed in R using the lmer, lmerTest, and emmeans packages (Bates, Mächler, Bolker, & Walker, 2014; Kuznetsova, Brockhoff, & Christensen, 2017; Lenth, Singmann, Love, Buerkner, & Herve, 2018).

Results

The purpose of Experiment 1 was to test how people strategically adjusted their referential language based on their audience and goal, in an unconstrained setting where participants could pick anything to talk about and also write anything to say. We analyzed three dependent measures: (1)

[1] In practice, in all experiments we omitted the conditions where the 'social signaling goal' was directed to only naive or out-group members because social signaling isn't meaningful when there aren't any in-group members in the audience.

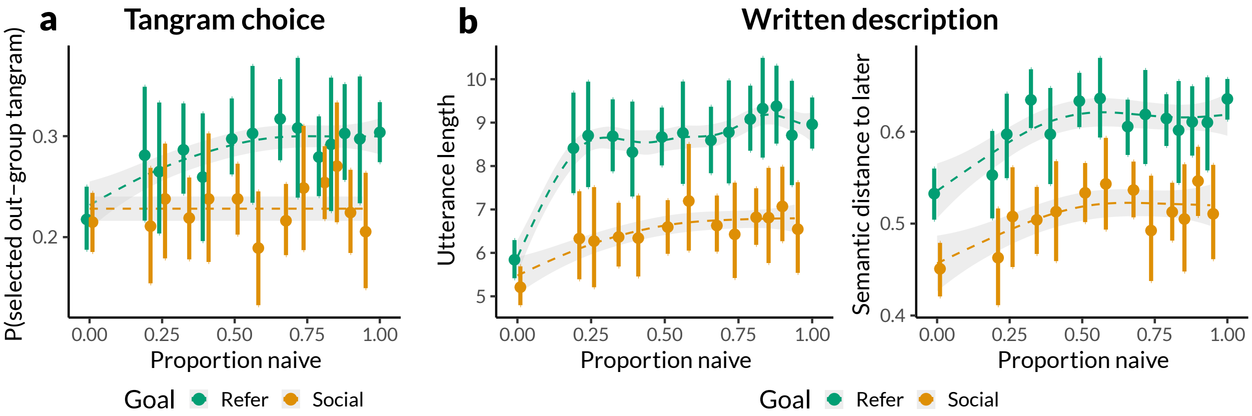

Fig. 2

Experiment 1 results. Given a goal of social identification, participants were more likely to ***(a)*** *choose tangrams that the in-group group talked about, and* ***(b)*** *write shorter descriptions that were also more similar to the descriptions that the in-group used. Error bars are bootstrapped 95% confidence intervals.*

participants' choice of tangrams ('in-group' tangrams that the group referred to, versus 'out-group' tangrams that only the other group referred to), (2) the number of words that participants wrote for their descriptions, and (3) the semantic distance of their descriptions to the converged labels for those tangrams that participants observed in the first phase of the task (by comparing SBERT embeddings using the `paraphrase-MiniLM-L6-v2` model; Reimers & Gurevych, 2019). We used mixed-effects regression models to test our predictions, predicting each of the above measures from audience composition (proportion of naive people in the audience) and goal ('refer' vs. 'social'), along with their interaction, with random effects for participant, tangram set, and selected tangram. We also included a quadratic term for proportion naive, to account for nonlinear effects of audience composition.

We first examined whether behavior in the 'refer' condition matches the standard referential efficiency findings. Standard accounts of referential efficiency predict that people want to be both informative and concise, and consider the knowledge level of their audience when computing the value of producing an utterance (Brown-Schmidt, Yoon, & Ryskin, 2015; Goodman & Frank, 2016; Hawkins et al., 2023). This account predicts that participants are more likely to tailor their expressions to naive people if their audience has a larger proportion of them (S. O. Yoon & Brown-Schmidt, 2019). Indeed, when addressing an audience of only in-group members, participants were more likely to choose tangrams that the in-group had talked about, and write shorter descriptions that are similar to the words that the in-group used; and as the number of naive people in the audience increased, so did participants' likelihood of choosing out-group referents and writing longer descriptions that were less similar to the in-group expressions (tangram choice $b_x = 0.32$, $z = 2.43$, $p = 0.015$; utterance length $b_x = 2.14$, $z = 5.52$, $p < 0.001$; semantic distance $b_x = 0.06$, $z = 3.09$, $p = 0.002$) (Figure 2). As predicted, we observed nonlinear effects of audience composition (tangram choice $b_{x^2} = -0.86$, $z = -2.66$, $p = 0.008$; utterance length $b_{x^2} = -5.12$, $z = -8.28$, $p < 0.001$; semantic distance $b_{x^2} = -0.14$, $z = -3.63$, $p < 0.001$): our results show that, compared to when there are no naive people in the audience, the presence of even a few naive people drastically changes people's behavior to produce longer, more descriptive utterances that are understood by the whole audience, consistent with previous studies showing that people 'aim low' to the least knowledgeable members of their audience (Bell, 1984; S. O. Yoon & Brown-Schmidt, 2018, 2019).

How does a goal of social signaling change participants' responses? For all three measures (tangram selection, utterance length, semantic distance), we found main effects of the social signaling goal. When they were told to explicitly signal social identity to the in-group members of their audience, people were more likely to choose tangrams that the in-group talked about ($b = -0.17$, $z = -3.96$, $p < 0.001$) (Figure 2a). They produced shorter utterances overall, counteracting the otherwise strong tendency to make sure everyone in the audience understood ($b = -0.98$, $t(7.14) = -5.95$, $p < 0.001$) (Figure 2b, left). The utterances were more semantically similar to the observed in-group utterances ($b = -0.04$, $t(15.1) = 4.48$, $p < 0.001$), for all audiences (Figure 2b, right). Furthermore, there were significant interaction effects for the utterance measures (utterance length $p < 0.001$; semantic distance $p = 0.014$): The effect of the 'social' goal was most pronounced when the in-group members were small minorities in the audience.

The influence of a social goal in this experiment depends on the extent to which participants learned and remembered the in-group's language in the observation phase. Because the observation phase was entirely passive and also relatively long (participants clicked through 72 videos), participants likely varied in their attention, learning, and memory for the group's labels. We therefore tested whether the effect of social goals would be more detectable in participants who had learned the group-specific expressions. We computed a 'learning score' for each participant, by comparing the average semantic similarity of participants' referential-goal descriptions for homogeneous in-group audiences, to the converged in-group labels they had observed in the first phase of the task. In other words, this metric captures whether people produce the same expressions that the in-group produced, for the same task that the in-group was observed doing. Indeed, participants with higher learning scores showed stronger effects of the social goal across all three dependent measures (tangram choice $r = 0.27$, $t(183) = 3.84$, $p < 0.001$; utterance length $r = 0.44$, $t(183) = 6.70$, $p < 0.001$; semantic similarity $r = 0.29$, $t(183) = 6.26$, $p < 0.001$). In other words, the more participants remembered the group's labels, the more they leveraged them to signal social identity. Considering only the top 50% of participants, in terms of learning score, the effects of the social goal were even more pronounced (Figure A1).

In sum, participants spontaneously tracked the observed linguistic context between minimal groups, and used this knowledge to tailor their utterances to the knowledge of the group.

Experiment 2

Experiment 1 replicated extensive prior evidence that people select language (here, what to talk about and what to say) by considering the referential informativity for the audience (will the intended listeners understand), and also showed that people strategically adjust their language based on whether their goal is referential or social signaling. However, because people freely chose any referent they wanted to talk about, and wrote any expression in a free-response format, participants' responses combined a number of decisions: which referent best affords social signaling, how conciseness serves social goals, and how audience composition modulates these trade-offs.

In the subsequent experiments, we decompose these bundled decisions using a forced-choice paradigm. By systematically controlling what varies across the choices (referent vs. expression, group-specific vs. shared, early vs. late), we can formalize these trade-offs in a computational framework that specifies how social utility competes with referential efficiency. Experiment 2 specifically measured speakers' choice of referent – what to talk about – to reveal their social identity.

Methods

Participants

We recruited 65 (40 female; ages 20-70, M(SD) age = 35.8(12.4)) adult fluent English speakers from the United States on Prolific. No participants were excluded. Participants gave informed consent. They were paid $15 for completing the experiment, which took an estimated 60 minutes to complete.

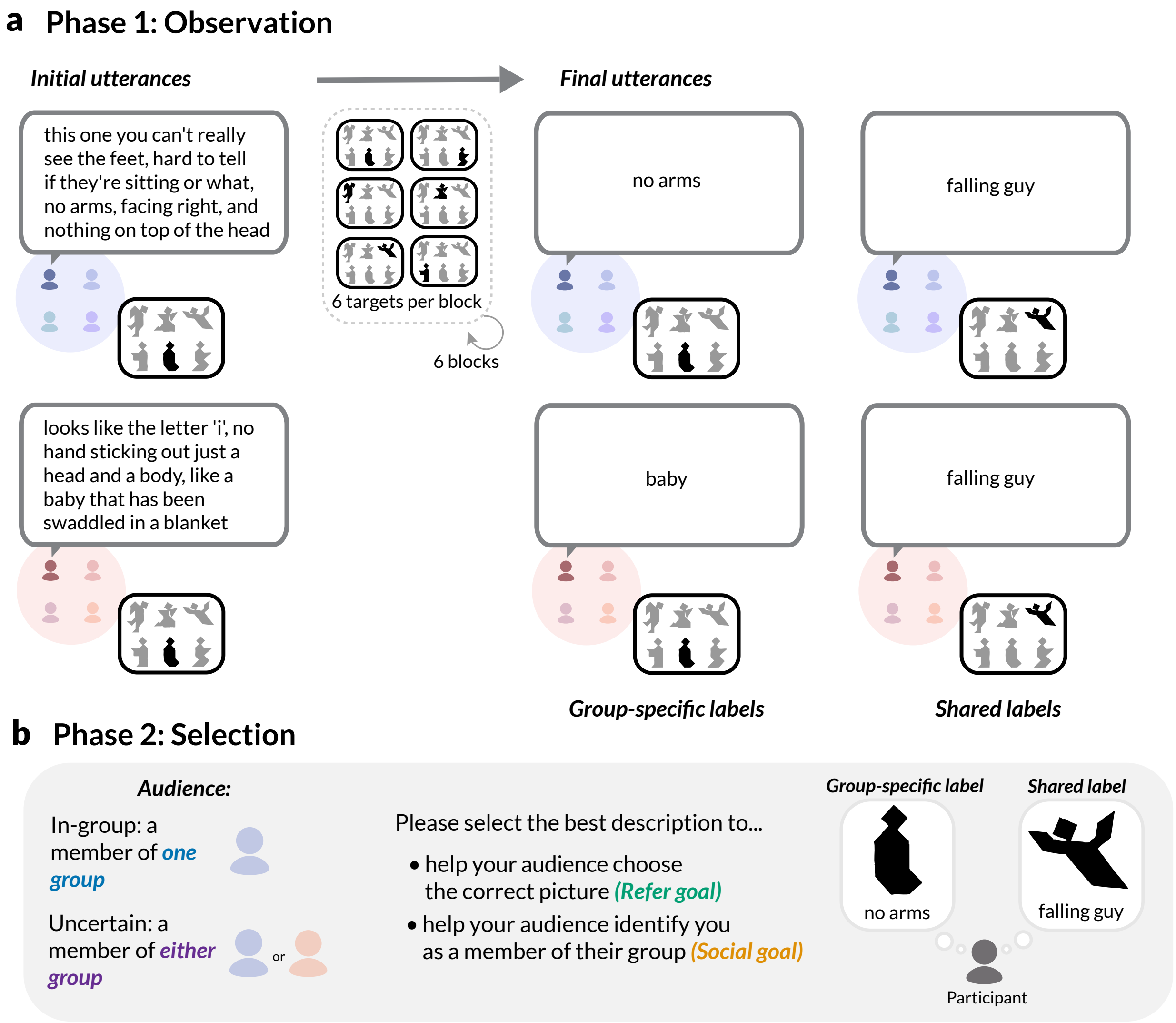
a
Phase 1: Observation
Initial utterances
Final utterances
this one you can't really see the feet, hard to tell if they're sitting or what, no arms, facing right, and nothing on top of the head
6 targets per block
6 blocks
no arms
falling guy
looks like the letter 'i', no hand sticking out just a head and a body, like a baby that has been swaddled in a blanket
baby
falling guy
Group-specific labels
Shared labels
b
Phase 2: Selection
Audience:
In-group: a member of one group
Uncertain: a member of either group
or
Please select the best description to...
• help your audience choose the correct picture (Refer goal)
• help your audience identify you as a member of their group (Social goal)
Group-specific label
Shared label
no arms
falling guy
Participant

Fig. 3

Experiment 2. ***(a)*** *In the first phase of the task, participants were third-party observers viewing two groups of people playing a multiparty reference game. Half of the tangrams converged to group-specific labels, and half of the tangrams converged to shared labels.* ***(b)*** *In the second phase of the task, participants selected between a group-specific label and a shared label.*

Procedure

Participants observed two distinct groups playing the same referential communication game. Unlike in Experiment 1, both groups referred to all tangrams in the context. To keep the experiment to a manageable length under the fully within-subject design, the context contained 6 tangrams instead of 12. This reduction was necessary to ensure participants could experience all possible combinations during the critical selection-phase trials without the task becoming prohibitively long.

A key feature of Experiment 2 was that each tangram was assigned to have either a shared label or a group-specific label. For the ‘shared-label’ tangrams, both groups independently converged on the same label (e.g., both the red and blue groups calling one tangram “falling guy”). For the ‘group-specific label’ tangrams, each group developed its own unique label for the same tangram (e.g., one group calling a tangram “baby,” while the other called it “no arms”) (Figure 3a, right). Thus for the ‘shared label’ tangrams, both groups were observed as coincidentally having converged to the same labels for those tangrams.

Then, participants proceeded to the selection phase, which manipulated the audience and goal. The audience was described as as a member of a specified group (a member of the red group, or a member of the blue group), or an ambiguous member whose group identity was uncertain (a member of ‘either the red group or the blue group’). The goal manipulation was the same as in Experiment 1: participants were either trying to refer (help the audience identify the target tangram) or signal social identity (indicate their own group membership).

On each critical trial, participants chose between two converged tangram–label pairs: one associated with a group-specific label and one with a shared label (Figure 3b). This design allows us to simultaneously isolate the effect of a social signaling goal on tangram choices (referential efficiency directed to the in-group is matched, so any difference in choices to the in-group is due to a social goal), while testing the standard referential efficiency predictions that should arise in the effect of audience identity (the ‘shared-label’ tangrams are referentially helpful for either group, while the ‘group-specific label’ tangrams are only referentially helpful for their specific group).

The assignment of tangram and label to condition (group-specific vs. shared) was counterbalanced across participants, and the design was fully within-subject (participants chose between all possible tangram-label pairs, for all audience conditions). Manipulation check trials were included to measure how well participants learned the observed tangram-label pairs and their group-specific sources. In the manipulation check trials, participants chose between two converged labels, each corresponding to the same tangram. For each of the ‘shared label’ tangrams, participants were given the ‘refer’ goal and asked to choose between a label that they saw in the gameplay, and a plausible label that they did not see. For each of the ‘group-specific label’ tangrams, participants were given the ‘social’ goal, and asked to choose

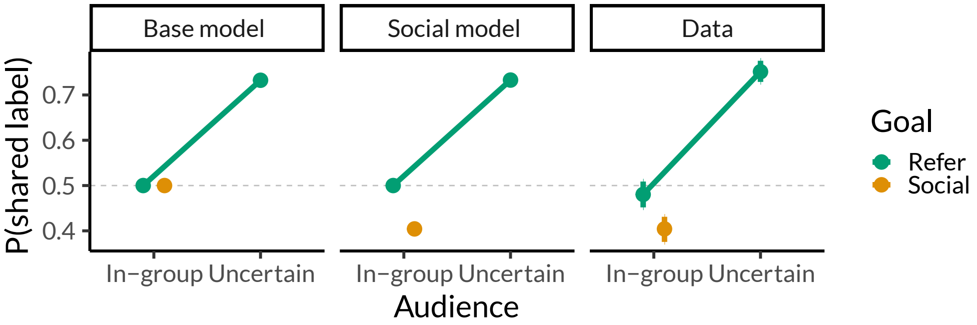

Fig. 4

Experiment 2 results. The base model predicts adaptation based on audience, but only the social model also captures the additional pressure in social signaling to choose group-specific label (vs. shared-label) referents. Error bars are 95% bootstrapped confidence intervals.

between each of the group's labels based on the audience group (the 'red' group or the 'blue' group). There were 15 manipulation check trials and 54 critical trials, for a total of 69 trials. Performance was incentivized as in Experiment 1.

Results

We tested our predictions using logistic mixed-effects regression models predicting participants' binary responses, from the goal-audience condition ('refer' to a specified group, 'refer' to 'either' group, and 'social' goal to a specified group), with random effects for participant, tangram set, and tangram.

We first examine the critical trials, in which participants chose between a 'shared label' tangram versus a 'group-specific label' tangram. When participants were uncertain about the audience, the label for the shared-label tangram provided higher referential efficiency, since it would be interpretable by members of either group. Consistent with this, when participants' goal was to refer, they more often chose the shared-label tangram when addressing an uncertain audience ($M = 1.60\,[1.14, 2.06]$, $z = 6.81$, $p < 0.001$; Figure 4 right, green).

Notably, the shared labels had also been experienced twice as often during the observation phase, potentially making them more accessible. However, participants only preferentially chose the shared-label tangrams when they would be referentially more informative. When addressing a specific group – where referential informativity was matched for both tangram-label pairs – participants selected shared and group-specific tangrams at similar rates ($M = -0.04\,[-0.46, 0.37]$, $z = -0.20$, $p = 0.840$). The observation that participants chose the shared-label tangram for the uncertain audience is therefore consistent with prior evidence that speakers are sensitive to their audience when considering referential informativity.

Next, we tested whether participants further adjust their selection based on an explicit social goal. When speaking to the in-group, both tangram-label pairs are equally informative from a referential perspective, but only the group-specific label provided an opportunity to signal group membership. Consistent with this prediction, participants were more likely to choose the group-specific tangram to achieve this goal ($M = -0.50\,[-0.90, -0.10]$, $z = -2.47$, $p = 0.014$) ($\text{diff} = 0.46\,[0.19, 0.73]$, $z = 3.98$, $p < 0.001$) (Figure 4 right, orange).

We examined whether individual variability in the strength of the social-goal effect was related to their performance on the manipulation check trials. Participants were near-ceiling in distinguishing the shared labels from plausible unobserved labels (98.4% correct; $\text{intercept} = 4.12$, $z = 5.85$, $p < 0.001$). However, overall accuracy in identifying which group-specific labels originated in each group was moderate (72.6% correct; $\text{intercept} = 0.98$, $z = 4.63$, $p < 0.001$). Individual participants who more accurately remembered the origins of group-specific labels were better able to use tangram selection to achieve a social goal ($r = 0.45$, $t(63) = 4.04$, $p < 0.001$).

Computational framework

To better understand the mechanisms driving participants' choices, we tested whether a computational model incorporating social-signaling utility explained behavior better than a baseline referential model. Our approach provides a principled way to formalize how participants might integrate different goals (referential clarity vs. social identification), along with their audience, when selecting an utterance.

Consistent with previous models of referential efficiency (Frank & Goodman, 2012; Goodman & Frank, 2016), our computational model formalizes people's choice of utterances as a utility-driven decision process, where an agent selects an action proportional to its total utility, modulated by the softmax temperature parameter α. In our paradigm, agents select an utterance u, given an audience a and goal g:

$$P(u|a,g) \propto \exp(\alpha \cdot U_{\text{tot}}(u|a,g)). \tag{1}$$

In the *social signaling* model, the total utility of the action is the sum of referential informativity

$U_{\text{inf}}(u|a,g)$ (how well the utterance allows the listener to identify the correct referent), the social utility $U_{\mathbf{soc}}(u|a,g)$ (the extent to which the utterance signals in-group identity), and the cognitive cost of producing the utterance $c(u|a,g)$ (here, captured by utterance length), the balance of which is modulated by the weights w_r, w_s, and w_c[2]:

$$U_{\text{tot}}(u|a,g) = w_r \cdot U_{\text{inf}}(u|a,g) + w_s \cdot U_{\text{soc}}(u|a,g) - w_c \cdot c(u) \tag{2}$$

The difference between the referential baseline and social signaling models is that in the baseline referential model, there is no social utility (i.e. $w_s = 0$). In the 'social signaling' condition, the social model, but not the baseline referential model, considers the value of social identification rather than referential informativity.

Model comparison

The experiments were specifically designed to manipulate conditions that change the balance between referential and social-signaling utilities. Thus, to fit the two models, we stipulated the utility terms as binary values reflecting the structure of the experimental design. Encoding utilities as binary values allowed us to capture these intended contrasts cleanly and test whether differences in behavior aligned with the experimental manipulations.

When the goal was reference, we set the referential informativity of an utterance to 1 if the audience's group had used the utterance to successfully refer, and 0 otherwise. When the audience was 'either' group, we marginalized over the speaker's utilities for both groups. For the social-signaling term, we grounded its value in the assumption that the social utility is proportional to the distinctiveness of the label. Group-specific labels are more diagnostic of group membership than shared labels, and thus we stipulated that they would carry greater U_{soc} in the social goal condition.

We implemented the computational models using memo, a probabilistic programming language for performing fast enumerative inference (Chandra, Chen, Tenenbaum, & Ragan-Kelley, 2025). We fixed $\alpha = 1$ and fit the other parameters by maximum likelihood estimation, using stochastic gradient descent;

[2] We include the cost term here for theoretical completeness, though it does not influence predictions in this experiment because participants always select between converged descriptions. It becomes relevant in later experiments where participants choose between utterances of varying length.

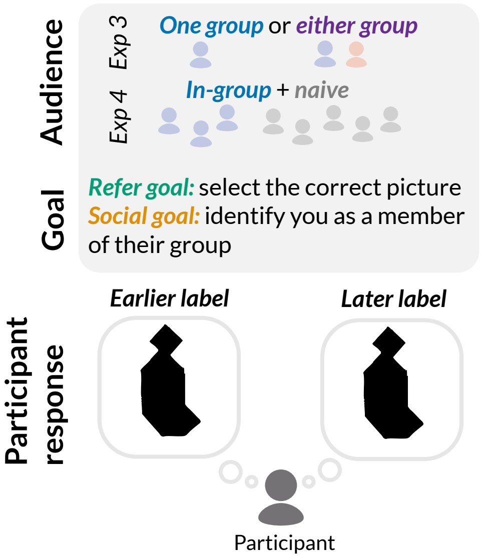
Audience
Exp 3
One group or either group
Exp 4
In-group + naive
Goal
Refer goal: select the correct picture
Social goal: identify you as a member of their group
Participant response
Earlier label
Later label
Participant

Fig. 5

Experiments 3 and 4. In both experiments, participants chose between a longer label that was produced earlier in the group's conversations, versus a shorter label that was produced later in the group's conversations. The experiments differed in audience manipulation. Experiment 3 manipulated whether the audience was from a specified group, or from either group. Experiment 4 manipulated the proportion of naive people in the audience.

all plots shown are predictions under the best-fitting parameters. The social signaling model (best-fitting params $w_r = 3.34$, $w_s = 0.39$, $\mathrm{NLL} = 2257.7$), but not the baseline referential model (best-fitting params $w_r = 3.31$, $\mathrm{NLL} = 2279.3$) captured the patterns in participants' responses, specifically the difference between the 'social' and the 'refer to one group' conditions (Figure 4), and explained the data better (likelihood-ratio test $\chi^2(1) = 43.24$, $p < 0.001$).

In summary, in Experiment 2, we found that when referential informativity is matched, people can further choose the option that reveals social group membership to an in-group member. A model that considers both referential and social-signaling utilities is needed to capture these patterns.

Experiment 3

In Experiment 2 participants simultaneously selected what to talk about and what to say, choosing between two different tangrams with known labels. In Experiment 3, we shifted focus to a different but complementary trade-off: the decision between alternative labels for the same tangram. Specifically, participants chose between an earlier, longer label drawn from the group's initial conversations, which is more descriptive but less efficient — and a later, shorter label that emerged through repeated interactions within the group (Figure 5). This setup allows us to examine the standard contrast of how participants balance informativity and efficiency (e.g., Isaacs & Clark, 1987), and how this balance shifts when a social-signaling goal is introduced.

Because participants are choosing between earlier and later labels, they now have to consider the balance of referential informativity and efficiency, depending on the audience. Earlier labels are more descriptive and universally comprehensible but more effortful for both the speaker and the listener, whereas later labels are concise and efficient but only informative to a group that generated them (Schober & Clark, 1989). The observation phase included (as above) both shared-label tangrams (generated independently by both groups) and group-specific label tangrams.

Methods

Participants

We recruited 61 (27 female; ages 19-67, M(SD) age = 36.6(12.5) adult fluent English speakers from the United States on Prolific, and excluded 1 participant who indicated that they did not understand the instructions. Participants gave informed consent, and were paid $15 for completing the experiment, who took an estimated 60 minutes to complete.

Procedure

The observation phase procedure, the bonus structure, and the structure of the manipulation check trials were the same as in Experiment 2. However, in the critical selection phase trials (fully within-subject design), participants choose between two labels for the same tangram: a label that appeared earlier in the given group's conversation (e.g., "this one you can't really see the feet, hard to tell if they're sitting or what, no arms, facing right, and nothing on top of the head"), versus a label that appeared later in the same conversation (e.g., "no arms"). There were 15 manipulation check trials and 36 critical trials, for a total of 51 trials in the experiment.

Results

The baseline referential account predicts that people use later labels when addressing the in-group, but switch to earlier labels when addressing an uncertain audience (Hawkins et al., 2023). This is indeed

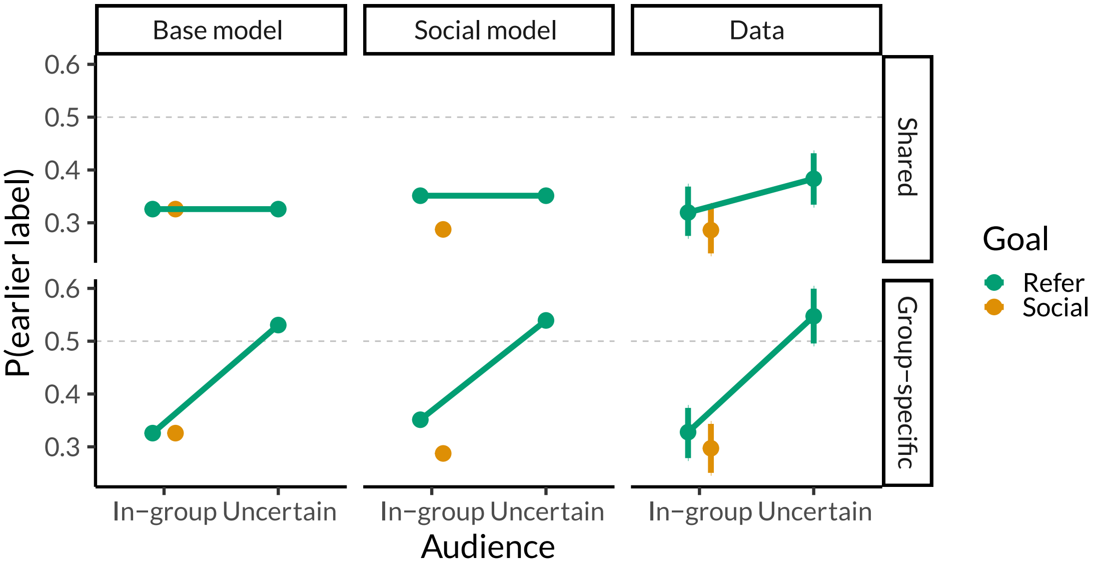

Fig. 6

Experiment 3 results. Top row: results for the 'shared-label' tangrams. Bottom row: Results for the 'group-specific label' tangrams. Error bars are bootstrapped 95% confidence intervals.

what we saw with the 'group-specific label' tangrams ('refer to one group' $M = -0.87\,[-1.26, -0.48]$, $z = -4.36$, $p < 0.001$; 'refer to either group' $M = 0.29\,[-0.27, 0.84]$, $z = 1.02$, $p = 0.310$) (Figure 6, bottom row). For the 'shared label' tangrams, the later labels are concise *and* informative to both groups, so participants chose them regardless of the audience ('refer to one group' $M = -0.93\,[-1.34, -0.53]$, $z = -4.54$, $p < 0.001$; 'refer to either group' $M = -0.76\,[-1.37, -0.16]$, $z = -2.47$, $p = 0.014$) (Figure 6, top row). Thus our data confirm that people are sensitive to their audiences and adjust what they say based on anticipated referential efficiency.

Can a social signaling goal itself lead to the same pressure of producing shorter and concise utterances to the in-group? While both the earlier and later expressions *occurred* within specific groups, the later utterances were *generated* through within-group interactions. Knowing how to use these later expressions for effective reference could thus reveal a shared in-group history. Indeed, participants selected the concise converged expressions when social signaling to the in-group, compared to referring to an uncertain audience ($\text{diff} = 0.87\,[0.37, 1.36]$, $z = 4.10$, $p < 0.001$).

Thus in our choice framework, the same later, shorter utterances could be used effectively to achieve two goals, when the audience is composed only of in-group members: reference and social signaling. When addressing the in-group, both kinds of goals created a pressure toward selecting the short grou-specific labels. We tested whether social signaling goals lead to an even greater pressure toward shorter utterances than referential goals, as potentially suggested by the Experiment 1 results (Figure 2b, left, 0% naive). This contrast was not significant in the preregistered dataset of Experiment 3 ($\text{diff} = 0.20\,[-0.12, 0.53]$, $z = 1.47$, $p = 0.306$) (but see the model comparison below, and another test of this comparison in Experiment 4).

Model comparison

As in the previous experiment, we stipulated the utility values based on the experimental design to directly test the hypothesized trade-offs. Earlier labels were assigned referential value for all audiences, reflecting the assumption that they were produced for an audience that the speaker had not interacted with before. Later labels, by contrast, were assigned referential value only when directed to the in-group, reflecting the assumption those expressions would not be as interpretable by out-group or naive listeners[3]. For social utility, the later labels, but not the earlier labels, were assigned to carry social signaling value, reflecting the assumption that conventions formed through repeated social interaction are idiosyncratic to a group and therefore can signal group membership. Finally, we incorporated an efficiency cost for the earlier labels, which were longer and more complex, reflecting the assumption that even in a forced-choice paradigm, the cognitive cost term $c(u)$ captures the perceived or simulated effort associated with selecting longer labels. Participants may implicitly simulate producing or processing these labels, making them less attractive except when their referential or social utility is sufficient to offset that cost.

As in Experiment 2, the social model (best-fitting params $w_r = 1.59$, $w_s = 0.29$, $w_c = 0.61$, $\text{NLL} = 1377.6$), but not the baseline referential model (best-fitting params $w_r = 1.75$, $w_c = 0.72$), captured the distinction between goals (likelihood ratio test $\chi^2(1) = 5.39$, $p = 0.020$; Figure 6). Thus model comparison identified a significant effect of social signaling value in participants' utterance selection, in addition to referential efficiency.

Experiment 4

Experiment 3 showed that social signaling goals, like referential goals, lead to a pressure to produce concise, in-group-generated utterances. However, in the design of Experiment 3 the short

[3] In practice, all labels carry some degree of referential utility. We revisit this in Experiment 4.

utterances could achieve both goals efficiently and simultaneously, so that experiment did not manipulate any *tension* between these goals. In Experiment 4, we systematically tested whether, as in Experiment 1, people make systematic trade-offs between social signaling considerations and referential clarity, based on their audience. Consider, for example, a scientist lecturing to a mixed audience of experts and novices. Prior research shows that when speakers in such situations pursue a referential goal, they tend to 'aim low,' adapting their language to the least-informed audience members: Speakers use concise in-group specific labels when the audience is entirely composed of in-group members, but sensitively shift to less efficient, more descriptive labels when even a small fraction of the audience is naive, prioritizing referential efficiency to the least informed members of the audience (Bell, 1984; S. O. Yoon & Brown-Schmidt, 2018, 2019). By contrast, we hypothesized that when a social signaling goal is introduced, speakers will resist the urge to 'aim low', and instead use the more concise group-specific labels, even when the in-group represents only a small fraction of the audience.

Methods

Participants

We recruited 180 (95 female; ages 18-72, M(SD) age = 37.4(11.9) adult fluent English speakers from the United States on Prolific, and excluded 1 participant who did not understand the instructions. Participants gave informed consent, and were paid $10 for completing the experiment, which took an estimated 30 minutes to complete.

Procedure

In the observation phase, participants viewed interactions from a single group, allowing them to learn that group's conventions. Then, in the selection phase, participants chose between earlier and later labels for a single tangram. The audience composition varied continuously in size and proportion of in-group members, and was depicted iconically as in Experiment 1 (Figure 5). The audience-goal manipulation was within-subject. To keep the task length manageable given the number of sets of audiences, participants saw one set of tangrams and labels within each condition manipulation ('refer' vs. 'social'), with the appearance of each tangram-label pair balanced within and across participants. Manipulation-check trials tested whether participants could reliably distinguish between observed labels and plausible but unobserved alternatives. There were 53 critical trials and 6 manipulation check trials, for

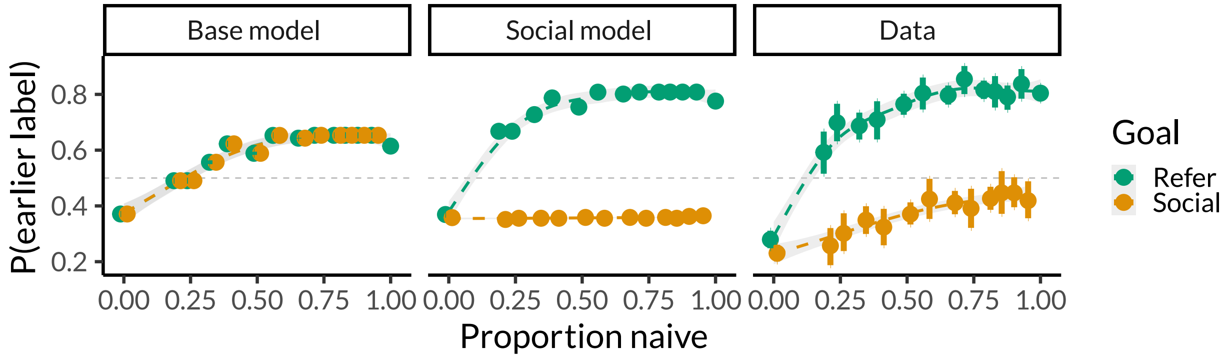

Fig. 7

Experiment 4 results. Error bars are bootstrapped 95% confidence intervals.

a total of 59 trials.

Eliciting transparency

When the audience includes naive people, we expected that participants' choices would partly depend on how relatively transparent the labels are (i.e. how interpretable they are to naive listeners). For example, if a later label is very opaque, participants might think that naive people won't understand what it is referring to, and instead be more likely to pick the earlier, longer label. To capture this item-wise variability, a separate sample of participants ($N = 201$) on Prolific guessed the correct tangram given a label (obtaining a measure for how well naive participants can infer the correct referent from each label). Each participant provided one response per tangram per tangram set (18 trials total), with the assignment of label to participant balanced across participants. These judgments were used to generate an estimate of transparency for each of the labels. Earlier labels were overall more transparent (83.0% correct; 95% CI: [81.3, 84.5]) than the later labels (70.2% correct; 95% CI: [67.7, 72.6]). Note that the transparency of both kinds of labels were well above chance (16.7% correct), consistent with previous estimates of transparency in tangram reference games (Boyce et al., 2025). The item-wise transparency measure was incorporated into the main analysis to account for item-level differences in how informative each label was to a naive audience.

Results

We tested our predictions using logistic mixed-effects regression models to predict participants' binary responses as a function of the proportion of naive people in the audience (coded as a quadratic term), the speaker's goal, and the interaction of these factors. We also included total audience size and the measured transparency differences as predictors. Participants remembered the converged labels from the observation phase, performing well on the manipulation check trials (94.0% correct; intercept $= 2.75$, $z = 10.71$, $p < 0.001$).

When speakers chose labels to 'refer', there were linear ($b_x = 3.83$, $z = 12.40$, $p < 0.001$) and quadratic ($b_{x^2} = -4.49$, $z = -10.15$, $p < 0.001$) effects of audience composition, showing the classic tradeoff of referential informativity and efficiency given the composition of the audience, and replicating the 'aim low' effects showed in Experiment 1 and in previous research (S. O. Yoon & Brown-Schmidt, 2018, 2019; see Figure 7, right). The label transparency for each item influenced participant's choices: people were overall more likely to select the earlier label, when the later label was less interpretable to a naive audience ($b = 0.66$, $z = 3.35$, $p < 0.001$). There was no effect of total audience size ($p = 0.913$).

We also observed a main effect of and interaction with goal ($b = -1.53$, $z = -9.80$, $p < 0.001$; both linear and quadratic interaction terms $p < 0.001$). As the in-group became more diluted in the audience, participants were comparatively less likely to switch to the earlier more transparent expressions for the 'social' goal than the 'refer' goal (Figure 7, right). That is, when given an explicit social goal, participants continued to use opaque in-group labels, despite the presence of many naive listeners who would not understand.

Model comparison

We implemented the computational models by considering each audience member as separately contributing to the total utility of producing an utterance (as in, e.g., Frank & Liu, 2018). Because the ceiling for participants' responses in the 'refer' condition asymptotes below 1, we fit an additional noise parameter ε corresponding to people's probability of choosing randomly on each trial. The baseline referential model (best-fitting params $w_r = 2.32$, $w_c = 2.45$, $\varepsilon = 0.69$, NLL $= 6247.9$) successfully captured the nonlinear effect of proportion naive people (i.e. the 'aim low' results). However, only the social-signaling model (best-fitting params $w_r = 2.13$, $w_s = 0.04$, $w_c = 0.091$, $\varepsilon = 0.38$, NLL $= 5609.5$) captured the distinction between the 'refer' and the 'social' goals, and was a better fit to the data (likelihood ratio test $\chi^2(1) = 1276.86$, $p < 0.001$; Figure 7).

In sum, Experiment 4 showed that participants adapted to their audience under referential goals, prioritizing clarity when naive people were present. But when signaling identity was the goal, participants maintained concise in-group expressions, even when speaking to mixed audiences for whom those labels would reduce overall comprehension. These findings highlight the trade-off between referential clarity and social signaling, demonstrating that speakers flexibly balance these competing pressures based on their audience and communicative goals.

Discussion

Language is not only a way to transmit information, but also a powerful tool for expressing social identity. Across four experiments, we measured how people strategically adjust their referential language when their goal is to signal group membership. When people were given an explicit social signaling goal, they consistently selected referents and labels that maximized in-group identification (Experiments 1 and 2). They also favored concise, group-specific labels that emerged during repeated interactions, even when longer, more descriptive labels would have been more universally understandable (Experiments 3 and 4). Finally, speakers resisted the well-documented tendency to adapt their language to the least knowledgeable audience members (Experiments 1 and 4), prioritizing identity signaling over universal comprehension. Formal cognitive modeling showed that extending referential efficiency models to include a social signaling term was critical for capturing these patterns.

These findings align with decades of sociolinguistic research showing that speakers dynamically adjust their language to signal affiliation and solidarity with social groups (e.g., Bell, 1984; Eckert, 1989; Giles et al., 1991). Such language use is documented in many real-world contexts. Scientists and lawyers, for example, often deploy technical jargon as a deliberate marker of expertise or group membership (Cruz & Lombrozo, 2025; Martínez, Mollica, & Gibson, 2024). Political actors use dogwhistles and coded language to signal affiliation with particular constituencies (Henderson & McCready, 2017; Van Der Does, Galesic, Dunivin, & Smaldino, 2022). Online communities rapidly develop shared slang that functions both as an efficient shorthand and a marker of belonging (Damirjian, 2025). Our controlled experiments demonstrate that the cognitive pressures underlying these patterns can emerge even in minimal, low-stakes settings, suggesting that the drive to balance referential efficiency and social signaling is a fundamental feature of human communication.

To uncover the cognitive mechanisms behind these behaviors, we designed a tightly controlled experimental paradigm that isolates the mechanisms of interest. Specifically, our approach allowed us to model and measure the strategic use of referential conventions — separate from other sources of linguistic variation (such as accent, prosody, or grammatical structure; Poplack & Sankoff, 1984; Myers-Scotton, 1998). Also, in our task, participants were placed in arbitrary novel groups that developed linguistic conventions on the fly, often within minutes; allowing us to examine the weakest conditions under which participants' utterances depend on social group structure and isolate the effects of these minimal groups on participants' responses. Even under these simplified conditions – and without any forewarning during the observation phase that the test phase would require social signaling – participants not only encoded group-specific language, but also deployed it strategically, adapting their language to their audience and communicative goals. This paradigm provided the most minimal scenario in which referential labels are tied to social identity, and so using this paradigm allowed us to cleanly compare the effects of signaling social affiliation to the standard referential efficiency accounts.

One discrepancy between our data and our formal social-signaling model concerns the implicit goals people bring to referential language use. Our experiments were designed to isolate deliberate goals, explicitly prompting participants to focus either on referential efficiency or on social signaling. Consistent with this design, the model predicts that under a purely social-signaling goal, the proportion of out-group or naive people in the audience should have no effect, since the value of signaling in-group identity is independent of out-group presence. However, in both Experiments 1 and 4, participants' behavior was sensitive to the proportion of out-group members even when instructed to signal group membership (Figure 2, Figure 7). This pattern suggests that speakers may balance multiple utilities – such as maintaining a baseline level of communicative clarity for all listeners – alongside their explicit goal of signaling group identity. Understanding how people integrate these competing pressures, often without conscious deliberation, is an important direction for future research.

Overall, for a speaker to choose utterances that convey social information requires knowledge of how listeners will interpret this information. That is, for an utterance to function as a social marker, both speakers and listeners must recognize the potential social meaning of an utterance or label. Our models stipulated that utterances had social meaning if they were similar to the language that the group used (Experiment 1), emerged uniquely within one group (Experiment 2), or were generated through the process of repeated interaction within a group (Experiment 3, 4). In Experiments 2-4, which aimed to isolate the contribution of a social signaling goal, social information was experimentally assigned to one of the two response options on every trial and the social utility in our model was defined based on this experimental design.

Note that this experimental design and model operationalization does not test or express the recursive social reasoning processes that human minds likely use to reason about audience perceptions of speakers' social identity. For a speaker, reasoning about the social information of an utterance requires reasoning about listeners' beliefs (about the speaker's social identity), and how their beliefs would change given a produced utterance. Formalizing such a reasoning process would require representations of how listeners might interpret a variety of referential terms in general, within and across different social groups and contexts. For example, richer approaches could estimate utilities on a continuous scale, informed by measuring what speakers think about the population-level distinctiveness of utterances and which listeners have access to those utterances. When such granular knowledge of utterances and their alternatives is available, it could be incorporated into Bayesian models of recursive reasoning, which already capture different kinds of socially motivated communication (Burnett, 2017; E. J. Yoon, Tessler, Goodman, & Frank, 2020; Papineau & Degen, 2024), and incorporate elements of audience adaptation based on existing and updated knowledge (Hawkins et al., 2023). Such formal models of recursive reasoning could simultaneously capture and test the reasoning process (i.e. about the referent the speaker is referring to, about the speaker, and about the broader population) that *generates* the social utility of utterances in the first place. Our approach lays a foundation for more expressive models that represent these reasoning processes in richer detail.

Language can be used for social signaling because language usage patterns covary with the history of intragroup interaction. Why is this the case? One possibility is that repeated intra-group interaction leads, purely for efficiency, to idiosyncratic expressions. Referential efficiency is sufficient to drive the emergence of idiosyncratic group-specific conventions, as shown in the selection phase (Hawkins et al., 2023). Once such variation arises, group-specific expressions could then be exploited for social signaling. This view is consistent with broader hypotheses that linguistic diversity arises by passive 'drift' when communities are relatively isolated from one another (Livingstone & Fyfe, 1999; Trudgill, 2004).

A second possibility is that social signaling is one key driving force behind the emergence of group-specific referring expressions in the first place (Roberts, 2012, 2013), even in minimal reference game contexts. Sociolinguistic and ethnographic studies suggest that people continuously create and use new terms in order to distinguish their own groups from other groups, and to identify themselves with certain social categories: either the ones they are in or the ones they want to be associated with (Labov, 1973; Milroy & Milroy, 1992; Nettle, 1999). In fact, adolescence, a stage where distinguishing social categories and affiliations is especially important, has been shown to be a primary time when new linguistic terms are created, leading to large-scale language change (Eckert, 1989). While our experiments cannot distinguish between these possibilities, our results – which show that speakers continuously manage the trade-off between being understood and signaling in-group affiliation – highlight the mechanisms that allow local decisions about language use and social identity to shape broader patterns of language evolution and change.

Constraints on generality

In naturalistic settings, language is embedded within a wide variety of social contexts – ranging from school or hobby-based communities to regional, cultural, or ethnic groups – where different linguistic cues, such as accent, dialect, or register, can signal social identity. By focusing on linguistic conventions generated on the fly in a controlled reference game (and only testing native English speakers from the United States), our experiments deliberately stripped away these broader cultural and linguistic factors to isolate the core cognitive mechanisms that guide social signaling in reference. However, in real-world contexts, these broader factors could also influence behavior. Future work should examine whether similar trade-offs between social and referential goals emerge across different languages and cultural settings, particularly in communities with distinct norms around indirectness, hierarchy, or group identity. Moreover, while our controlled paradigm enabled precise measurement and modeling, real-world communication involves more nuanced social dynamics, overlapping goals, and higher stakes. For example, when the costs of outsiders discovering one's social identity are high, people may develop more covert linguistic signals (Smaldino & Turner, 2022), taking into account the knowledge level of outsiders and overhearers. Future work should investigate how social-signaling goals interact with other communicative priorities and constraints that shape language use in the complex dynamics of real-world social identities.

Acknowledgments

We thank Eric Martínez, Camille Osumah, Lio Wong, Kartik Chandra, Tom Griffiths, Adele Goldberg, Ted Gibson, Josh Tenenbaum, and members of the Saxelab and Social Interaction Lab for helpful discussions and feedback. This research was supported by award #6951949 from the Simons Center for the Social Brain.

Appendix A

Supplementary figures

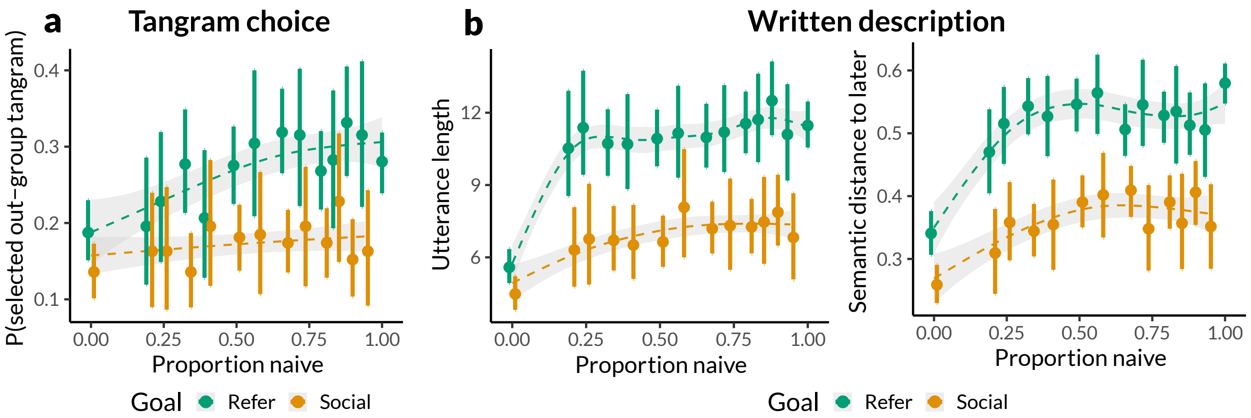

Fig. A1

Experiment 1 results, for just the participants who scored above the median on the learning metric. Error bars are bootstrapped 95% confidence intervals.

Appendix B

Deviations from preregistration

Experiment 1

For Experiments 1 and 4, we preregistered analyses testing for an overall main effect of the proportion of naive audience members. However, we determined that our primary theoretical interest was whether the proportion of naive people influenced behavior specifically within the ‘refer’ condition of the goal predictor. Focusing on the ‘refer’ condition allows for more precise comparison with existing research on referential efficiency. Consequently, we reported the marginal effects (estimated slopes within the ‘refer’ level of the factor) rather than focusing solely on the overall main effect, thereby providing a more targeted and theoretically relevant test.

Experiment 2

The preregistered model treated audience and goal as separate predictors. In the final analyses, these factors were combined into a single ‘condition’ variable, for consistency with Experiment 3. Additionally, given the unbalanced design (with the social goal only present in the ‘uncertain’ audience condition), combining these predictors provided a statistically more appropriate and coherent model.

Experiment 3

No deviations.

Experiment 4

The preregistered model did not include label transparency as a predictor. We incorporated this variable into the model to account for additional variance in participants’ choices attributable to item-level differences in transparency.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using the lme4 package in r. *J Stat Softw*, *67*, 1–48.

Bell, A. (1984). Language style as audience design. *Language in society*, *13*(2), 145–204.

Boyce, V., Hawkins, R. D., Goodman, N. D., & Frank, M. C. (2024). Interaction structure constrains the emergence of conventions in group communication. *Proceedings of the National Academy of Sciences*, *121*(28), e2403888121.

Boyce, V., Prystawski, B., Tan, A., & Frank, M. C. (2025). Idiosyncratic but not opaque: Linguistic conventions formed in reference games are interpretable by naïve humans and vision–language models.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, *22*(6), 1482.

Brown, Z. C., Anicich, E. M., & Galinsky, A. D. (2020). Compensatory conspicuous communication: Low status increases jargon use. *Organizational Behavior and Human Decision Processes*, *161*, 274–290.

Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as contexts in conversation. In *Psychology of learning and motivation* (Vol. 62, pp. 59–99). Elsevier.

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, *7*(4-5), 585–614.

Bullock, O. M., Colón Amill, D., Shulman, H. C., & Dixon, G. N. (2019). Jargon as a barrier to effective science communication: Evidence from metacognition. *Public Understanding of Science*, *28*(7), 845–853.

Burnett, H. (2017). Sociolinguistic interaction and identity construction: The view from game-theoretic pragmatics. *Journal of Sociolinguistics*, *21*(2), 238–271.

Burnett, H. (2023). *Meaning, identity, and interaction: Sociolinguistic variation and change in game-theoretic pragmatics*. Cambridge University Press.

Camp, E., & Nowak, E. (2025). Linguistic variation, agency, and style.

Chandra, K., Chen, T., Tenenbaum, J. B., & Ragan-Kelley, J. (2025). A domain-specific probabilistic programming language for reasoning about reasoning (or: a memo on memo). *psyarxiv preprint*. Retrieved from https://doi.org/10.31234/osf.io/pt863

Clark, H. H., & Marshall, C. R. (1981). Definite knowledge and mutual knowledge.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Cruz, F., & Lombrozo, T. (2025). How laypeople evaluate scientific explanations containing jargon. *Nature Human Behaviour*, 1–16.

Damirjian, A. (2025). The social significance of slang. *Mind & Language*, *40*(2), 138–156.

de Leeuw, J. R. (2024). Datapipe: Born-open data collection for online experiments. *Behavior Research Methods*, *56*(3), 2499–2506.

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), 5351.

Eckert, P. (1989). *Jocks and burnouts: Social categories and identity in the high school*. Teachers college press.

Eckert, P., & Brown, K. (2006). Communities of practice. *Concise encyclopedia of pragmatics*, 109–112.

Fedzechkina, M., Hall Hartley, L., & Roberts, G. (2023). Social biases can lead to less communicatively efficient languages. *Language Acquisition*, *30*(3-4), 230–255.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frank, M. C., & Liu, L. (2018). Modeling classroom teaching as optimal communication.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in cognitive sciences*, *23*(5), 389–407.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, *1*, 1–68.

Goffman, E. (1956). *The presentation of self in everyday life*. Routledge.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Gumperz, J. (1982). Discourse strategies. *Cambridge UP*.

Hawkins, R. D., Franke, M., Frank, M. C., Goldberg, A. E., Smith, K., Griffiths, T. L., & Goodman, N. D. (2023). From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*, *130*(4), 977.

Henderson, R., & McCready, E. (2017). How dogwhistles work. In *Jsai international symposium on* *artificial intelligence* (pp. 231–240).

Holmes, J., & Meyerhoff, M. (1999). The community of practice: Theories and methodologies in language and gender research. *Language in society*, *28*(2), 173–183.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of experimental psychology: general*, *116*(1), 26.

Jara-Ettinger, J., & Rubio-Fernandez, P. (2022). The social basis of referential communication: Speakers construct physical reference based on listeners' expected visual search. *Psychological review*, *129*(6), 1394.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45–52.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kinzler, K. D. (2021). Language as a social cue. *Annual Review of Psychology*, *72*(1), 241–264.

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, *104*(30), 12577–12580.

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*, 113–114.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *Journal of statistical software*, *82*(13).

Labov, W. (1966). *The social stratification of english in new york city*. Cambridge University Press.

Labov, W. (1973). *Sociolinguistic patterns* (No. 4). University of Pennsylvania press.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. *R package version*, *1*(1), 3.

Le Page, R. B. (1968). Problems of description in multilingual communities. *Transactions of the Philological Society*, *67*(1), 189–212.

Livingstone, D., & Fyfe, C. (1999). Modelling the evolution of linguistic diversity. In *European conference on artificial life* (pp. 704–708).

Lupyan, G., & Dale, R. (2016). Why are there different languages? the role of adaptation in linguistic diversity. *Trends in cognitive sciences*, *20*(9), 649–660.

Martínez, E., Mollica, F., & Gibson, E. (2022). Poor writing, not specialized concepts, drives processing difficulty in legal language. *Cognition*, *224*, 105070.

Martínez, E., Mollica, F., & Gibson, E. (2024). Even laypeople use legalese. *Proceedings of the National Academy of Sciences*, *121*(35), e2405564121.

Milroy, L., & Milroy, J. (1992). Social network and social class: Toward an integrated sociolinguistic model1. *Language in society*, *21*(1), 1–26.

Myers-Scotton, C. (1998). *Codes and consequences: Choosing linguistic varieties*. Oxford University Press.

Nettle, D. (1999). Using social impact theory to simulate language change. *Lingua*, *108*(2-3), 95–117.

Papineau, B., & Degen, J. (2024). Biological males' and'trans (gender) women': Social considerations in the production of referring expressions. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Poplack, S., & Sankoff, D. (1984). Borrowing: the synchrony of integration.

Reimers, N., & Gurevych, I. (2019, 11). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing.* Association for Computational Linguistics. Retrieved from http://arxiv.org/abs/1908.10084

Rhodes, M., Leslie, S.-J., Bianchi, L., & Chalik, L. (2018). The role of generic language in the early development of social categorization. *Child Development*, *89*(1), 148–155.

Roberts, G. (2012). An experimental study of social selection and frequency of interaction in linguistic diversity. In *Experimental semiotics: Studies on the emergence and evolution of human communication* (pp. 139–160). John Benjamins Publishing Company.

Roberts, G. (2013). Perspectives on language as a source of social markers. *Language and Linguistics Compass*, *7*(12), 619–632.

Scherer, K. R., & Giles, H. (1979). *Social markers in speech*. Cambridge University Press; Ed. de la Maison des sciences de l'homme.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive psychology*, *21*(2), 211–232.

Smaldino, P. E. (2022). Models of identity signaling. *Current Directions in Psychological Science*, *31*(3), 231–237.

Smaldino, P. E., & Turner, M. A. (2022). Covert signaling is an adaptive communication strategy in diverse populations. *Psychological review*, *129*(4), 812.

Trudgill, P. (2004). *New-dialect formation: The inevitability of colonial englishes*. Oxford University Press, USA.

Van Der Does, T., Galesic, M., Dunivin, Z. O., & Smaldino, P. E. (2022). Strategic identity signaling in heterogeneous networks. *Proceedings of the National Academy of Sciences*, *119*(10), e2117898119.

Walker, A. C., Fugelsang, J. A., & Koehler, D. J. (2025). Partisan language in a polarized world: In-group language provides reputational benefits to speakers while polarizing audiences. *Cognition*, *254*, 106012.

Woolard, K. A. (2004). Codeswitching. *A companion to linguistic anthropology*, 73–94.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, *4*, 71–87.

Yoon, S. O., & Brown-Schmidt, S. (2018). Aim low: Mechanisms of audience design in multiparty conversation. *Discourse Processes*, *55*(7), 566–592.

Yoon, S. O., & Brown-Schmidt, S. (2019). Audience design in multiparty conversation. *Cognitive science*, *43*(8), e12774.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.

Zipf, G. K. (1949). Human behavior and the principle of least effort.