AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2023, Vol. 152, No. 6, 1598–1621
https://doi.org/10.1037/xge0001341

Investigating the Influence of Local and Personal Common Ground on Memory for Conversation Using an Online Referential Communication Task

Daniel R. Nault[1], Rohit Voleti[2], Matthew Nicastro[1], and Kevin G. Munhall[1]
[1] Department of Psychology, Queen's University
[2] School of Electrical, Computer, and Energy Engineering, Arizona State University

To maintain efficiency during conversation, interlocutors form and retrieve memory representations for the shared understanding or common ground that they have with their partner. Here, an online referential communication task (RCT) was used in two experiments to examine whether the strength and type of common ground between dyads influence their ability to form and recall referential labels for images. Results from both experiments show a significant association between the strength of common ground formed between dyads for images during the RCT and their verbatim—but not semantic—recall memory for image descriptions about a week later. Participants who generated the image descriptions during the RCT also showed superior verbatim and semantic recall memory performance. In Experiment 2, a group of friends with pre-existing personal common ground were significantly more efficient in their use of words to describe images during the RCT than a group of strangers without personal common ground. However, personal common ground did not lead to enhanced recall memory performance. Together, these findings provide evidence that individuals can remember some verbatim words and phrases from conversations, and partially support the theoretical notion that common ground and memory are intricately linked conversational processes. The null findings with regard to semantic recall memory suggest that the structured nature of the RCT may have constrained the types of memory representations that individuals formed during the interaction. Findings are discussed in relation to the multidimensional nature of common ground and the importance of developing more natural conversational tasks for future work.

Public Significance Statement
Results from the current studies suggest that our ability to recall exact words and phrases from our previous interactions with others is significantly enhanced when we form stronger agreement or common ground with our conversational partner about topics that we discuss. These findings also indicate that humans can, in some instances, remember exact words and phrases from their interactions 7 days later, despite not being told in advance that their memory will be tested.

Keywords: common ground, conversation, recall memory

This article was published Online First February 16, 2023.
Daniel R. Nault https://orcid.org/0000-0003-3066-5145
This project was completed in fulfilment of the first author's master's degree. An earlier version of the present work has been previously uploaded to QSpace, an open access repository for graduate student theses at Queen's University. Portions of this work were previously presented at the annual Psychonomic Society Conference in November 2021. The raw data, stimuli, and analysis code for the present experiments are publicly available on OSF: https://osf.io/hg7d3/

Daniel R. Nault served as lead for conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, writing–original draft, and writing–review & editing. Rohit Voleti contributed equally to software, and served in a supporting role for data curation, formal analysis, and methodology. Matthew Nicastro served as lead for software and served in a supporting role for data curation and validation. Kevin G. Munhall served as lead for funding acquisition, resources, and supervision and contributed equally to project administration, and served in a supporting role for conceptualization and writing–review and editing.

Correspondence concerning this article should be addressed to Daniel R. Nault, Department of Psychology, Queen's University, Humphrey Hall, 62 Arch Street, Kingston, Ontario, K7L 3N6, Canada. Email: daniel.nault@queensu.ca

The 2021 United States Capitol attack in Washington DC provides a striking example of the importance of recall memory for conversation. During the Capitol siege, the House Minority Leader Kevin McCarthy had crucial conversations with President Trump about the need for the president to call off his rioting supporters. Following this call, Representative McCarthy recounted the conversation he had with President Trump to several others, including Representative Jaime Herrera Beutler. On several occasions, Herrera Beutler recounted her recollection of Mr. McCarthy's description of the conversation, which included the much-quoted phrase supposedly uttered by President Trump, "Well, Kevin, I guess these people are more upset about the election than you are" (Gangel et al., 2021). This phrase and its purported verbatim recall could offer crucial insight into President Trump's intentions and mental state before and during the attack. Yet, how accurate was McCarthy's and Herrera Beutler's verbatim recall memory of the conversation with President Trump? What contextual factors and conversational mechanisms may have shaped their memories for the highly charged interaction?

From the outside, conversations like that between Representative McCarthy and President Trump may seem highly memorable,

especially given the fact that they took place in a time of political crisis. However, decades of research have revealed that verbatim memory for connected discourse is often limited and imprecise. In a seminal study by Sachs (1967), individuals heard a series of passages and were tested for their memory for sentences embedded within the passages at different time intervals. Results show that participants were accurate in their recognition for subtle semantic and syntactic changes made to the sentences immediately after they heard the passages. Importantly, however, recognition accuracy for syntactic—but not semantic—changes dropped to chance levels by the time participants had heard 80 syllables beyond the sentence being tested. While the meaning of sentences was thus retained quite well in memory over time, memory for the original surface form of the sentences was forgotten very quickly (Sachs, 1967). In a related study, Bransford and Franks (1971) presented participants with simple sentences that contained either one, two, or three semantically related ideas. In a recognition memory task, participants were then presented with the same sentences, along with several new complex ones that contained different combinations of all the semantic ideas from the simpler sentences. The results indicate that individuals falsely recognized many of the new complex sentences. Bransford and Franks (1971) concluded by suggesting that subjects had likely integrated the semantic ideas from each of the simpler sentences into a wholistic memory representation that was reflective of the overall meaning of each sentence combined rather than the surface form of each individual sentence. These findings and more recent others (e.g., Gernsbacher, 1985; Potter & Lombardi, 1990) have led to the general notion that linguistic material is primarily encoded and represented in memory in terms of its overall meaning or gist rather than its original surface structure.

While it is acknowledged that semantic memory generally outweighs verbatim memory, individuals can recall verbatim content from conversations under some limited circumstances. For example, Neisser (1981) examined former White House Counsel member John Dean's recall memory for conversations he had with former president Richard Nixon during the Watergate scandal. This study was carried out by comparing the content of the conversations secretly recorded by President Nixon to what John Dean recalled about them during his testimony in front of the Senate Watergate Committee. The analysis shows that John Dean's memory was often reflective of the overall impressions he had of the conversations. He did, however, recall several verbatim words and phrases from those conversations, particularly those he repeated multiple times or spent additional time practicing (Neisser, 1981). Several factors have also been shown to influence the amount of verbatim information that individuals can recall from discourse. For example, Keenan et al. (1977) examined individuals' ability to recognize verbatim statements from a lunchroom discussion 30 hr after the discussion had ended. Individuals were shown to be three times more likely to accurately recognize verbatim statements when they contained information about a speakers' beliefs, intentions, and attitudes toward the listener as opposed to when the information contained emotionally neutral information (Keenan et al., 1977). Individuals have also been shown to exhibit superior recall for the surface forms of utterances when told in advance that their memory will be tested (Johnson-Laird & Stevenson, 1970; Stafford & Daly, 1984), when interlocutors have superior interpersonal competence (Miller & de Winstanley, 2002), and when they are more familiar with one another (Samp & Humphreys, 2007).

One structural aspect of conversation that has recently been suggested to influence memory for conversation is the formation of common ground between interlocutors. A term first proposed by H. H. Clark and Brennan (1991), common ground refers to the collection of mutual knowledge, beliefs, and assumptions of two or more people engaging in conversation together. Achieving common ground is an ongoing process that develops over the course of time, as new ideas and topics are added to an ongoing conversation, and as conversational partners engage in repeated conversational exchanges and learn more about each other. Common ground is also shaped by the mutual knowledge that individuals often bring to a conversation about broader topics and issues, like religion or politics (E. V. Clark, 2015). During conversation, speakers rely on their memories to plan and tailor their utterances to the common ground they have accumulated with their partner over repeated exchanges (H. H. Clark, 1996; H. H. Clark & Marshall, 1981). Moreover, conversational topics are often built on a foundation of memories from previous interactions involving the same topic(s) and individual(s). Without a running record of what was previously shared or agreed upon in a current or past exchange, conversations would suffer from redundancy and inefficiency. Common ground is thus an essential aspect of conversation that allows interlocutors to build their communication together indefinitely without restating redundant information.

In conversation research, the formation of common ground has almost exclusively been studied in isolated interchanges using structured communication tasks. This is largely due to the difficulty of establishing reliable and valid dependent measures of common ground formation in naturalistic conversation. In natural conversation, topics can vary widely and thus every conversation will differ substantially from another. Communication tasks constrain the form and content of conversations. With this in mind, H. H. Clark and Wilkes-Gibbs (1986) employed a referential communication task (RCT) initially developed by Krauss and Weinheimer (1964, 1966) to study common ground formation in an experimental setting. In this task, pairs of participants are presented with a series of geometric figures (tangrams) for which there are no obvious or correct descriptive labels that can be used to describe them. One participant is assigned the role of the Director and the other is assigned the role of the Matcher. The goal of the task is for the Director to describe each image in their static matrix with sufficient detail to the Matcher so that the Matcher can rearrange their images into the same configuration in their matrix. Over the course of six trials of matching with the same images, H. H. Clark and Wilkes-Gibbs (1986) showed that partners increasingly used shorter verbal descriptions to refer to the tangrams. This tendency for conversational partners to develop a shared understanding for the labels of images and use more concise descriptions over time has been frequently replicated (e.g., Brennan & Clark, 1996; McKinley et al., 2017; Van der Wege, 2009; Wilkes-Gibbs & Clark, 1992) and is taken as evidence for the formation of common ground in conversation.

Recently, McKinley et al. (2017) sought to investigate the influence of common ground formation on memory for structured conversation. Pairs of participants engaged in a RCT and their recognition memory for images presented to them during the task was subsequently tested. Each individual in a pair played the role of both the Director and the Matcher and engaged in matching with two different partners. This design enabled McKinley et al.

(2017) to test whether image recognition memory differed as a function of the strength of common ground formation, conversational role (i.e., Director vs. Matcher), and context. As expected, the results show that partners established common ground, such that Directors reduced the number of words they used to describe each image to Matchers by an average of 4.77 words over the course of three trials of matching. More importantly, however, a mixed-effects model revealed that image recognition memory was significantly influenced by both the strength of common ground formation and conversational role. For every reduction in one word used by Directors to describe an image in the RCT, participants were 1.03 times more likely to correctly recognize an image. Further, participants were 2.64 times more likely to accurately recognize images when they interacted with them as a Director rather than as a Matcher (McKinley et al., 2017). Individuals were also 1.02 times more likely to be accurate in their identification of which partner they saw an image with when they were playing the role of the Director. These results suggest that the development of common ground in conversation may promote improved memory for conversation among both speakers and listeners (McKinley et al., 2017). It is worth noting, however, that McKinley et al. (2017) did not directly investigate the impact of common ground formation on memory for conversational content. Rather, the association between common ground formation and memory was indirectly inferred via a measure of image recognition memory performance. Thus, it is possible that the strength of common ground formed between dyads for images during the interaction may not have influenced the strength of their memory representations for true conversational content.

As noted above, the common ground that exists between conversational partners and that enhances their ability to efficiently communicate often includes information beyond what has been grounded in one isolated interaction (E. V. Clark, 2015). Romantic partners or long-term friends, for example, bring to each conversation a collection of previous experiences and shared knowledge about each other that can be retrieved from their memories to ignite new discussions or build on a previous one. Conversational partners who belong to the same religious group or reside in the same university residence may also have shared knowledge that they can draw upon to facilitate discussion about topics that are of common interest to them. This suggests that there may be differences in the way that conversational partners with various types of social relationships and thus, varying degrees of previously existing common ground, are able to form common ground for new information.

Studies that have investigated whether conversational dyads with pre-existing personal common ground have greater communicative efficiency than those who do not have surprisingly reported mixed findings. For example, Boyle et al. (1994) had pairs of friends and strangers complete a map task, wherein one partner was tasked with providing instructions to the other about how to draw a route on a map. The results showed that pairs of strangers used significantly less words to complete the task than friends, although friends interrupted and spoke over each other significantly less than strangers (Boyle et al., 1994). In another study, conversational dyads from New York City were shown to reach common ground for pictures of New York City landmarks more efficiently than conversational dyads who had never been to New York City (Isaacs & Clark, 1987). In contrast, Schober and Carstensen (2009) found no significant difference in the time it took for married couples and strangers to establish common ground for unfamiliar objects or people. Pollmann and Krahmer (2018) also reported no significant difference in the communicative efficiency of married partners as compared to strangers during a game of Taboo.

It is important to consider that these mixed findings may be explained by differences in the specific tasks that were used in each of these studies to evoke conversations. In some cases, the task itself may have limited the opportunity for dyads to draw upon and thus benefit from their pre-existing personal common ground. In Boyle et al. (1994), for instance, the task goal of providing instructions for drawing a route on an unfamiliar map would presumably give little reason for dyads to use their shared experiences and common knowledge of each other to their advantage. Recalling and using information from personal common ground would be more likely to hinder task success by derailing the exchange and making it less efficient. In the study by Isaacs and Clark (1987), on the other hand, having personal common ground about New York City offers a direct advantage to interlocutors who are being asked to form local common ground for pictures of New York City landmarks. Hence, more research is needed to better understand how pre-existing personal common ground influences the grounding of new information, particularly in conversational contexts that provide sufficient opportunity for the natural usage of such personal information. Moreover, little is known about the impact that this shared knowledge may have on the memory representations that are built and can later be accessed by interlocutors (e.g., Samp & Humphreys, 2007).

The Present Research

The aim of the current set of studies was to provide a comprehensive investigation of the association between different types of common ground and memory for conversation. We conducted two separate experiments using an online version of the RCT (Krauss & Weinheimer, 1964, 1966) that we adapted for use during the COVID-19 pandemic. Based on consistent findings from several previous RCT experiments (e.g., Krauss & Weinheimer, 1964, 1966; McKinley et al., 2017; Wilkes-Gibbs & Clark, 1992), we expected that conversational dyads in both experiments, albeit in a virtual format, would develop shared referential labels to describe images. We further predicted that dyads would use shorter referential labels over time, thus providing evidence of local common ground formation (McKinley et al., 2017; Wilkes-Gibbs & Clark, 1992).

In both experiments, we extended the memory work of McKinley et al. (2017) by testing whether the relative strength of local common ground formed between dyads in a RCT could predict their verbatim and semantic recall memory for image descriptions in a follow-up cued recall memory task 7 days later. In Experiment 1, all dyads knew each other prior to participating. In Experiment 2, half of the dyads were friends, and half were strangers. Our general hypothesis for both experiments was that relatively stronger local common ground formation between dyads for images during the RCT would be related to stronger recall memory for image descriptions at follow-up. We chose to test participants' recall memory following a 7-day period, as recall memory for linguistic information like words and narrative text has been shown to drop considerably about 7 days after initial encoding (see Fisher & Radvansky, 2018; Radvansky et al., 2022). To test verbatim recall memory, we used word accuracy methods akin to approaches used in speech intelligibility research (e.g., Bamford & Wilson, 1979; Rosen &

Corcoran, 1982; see Methods below). Given the evidence showing that humans have a limited capacity to recall verbatim words and phrases that they encounter during connected discourse (Bransford & Franks, 1971; Gernsbacher, 1985; Potter & Lombardi, 1990; Sachs, 1967; Stafford & Daly, 1984), we also opted to include a measure of semantic recall memory. Semantic memory is less susceptible to decay and more reflective of the overall meaning or gist that individuals encode and remember about events and objects like conversations and images (Tulving, 1972). Semantic recall memory performance was determined using a natural language processing (NLP) approach to estimate the degree of semantic similarity between conversational content shared during the RCT and recall memory for that same information (see Methods). Our aim in using NLP was to isolate any semantic similarity between image descriptions used during the RCT and image descriptions recalled from memory that may exist independent of verbatim similarity.

Our design for both experiments also included a recognition memory task that was presented immediately following the RCT. This task was mainly used as a manipulation check to ensure that participants were paying attention to and forming memories for images in an online environment. Our reasoning was that, if participants performed near ceiling levels on a recognition memory task, we could be relatively confident that our memory data were valid and reliable. While humans are known to perform exceptionally well on immediate picture recognition tasks (Standing, 1973), divided attention can lead to lower performance (Hicks & Marsh, 2000).

Experiment 1

In Experiment 1, we directly examined whether the relative strength of local common ground formed between dyads for basic category object images during an online RCT could predict their ability to individually recall verbatim and semantic conversational content (i.e., image descriptions) from the RCT about a week later. As noted above, our prediction was that relatively stronger local common ground formation during the RCT would lead to superior verbatim and semantic recall memory performance among participants about a week later.

Method

Transparency and Openness

The raw data, stimuli, and analysis code are publicly available on OSF here: https://osf.io/hg7d3/. The present experiments were not preregistered.

Participants

Thirty participants (i.e., 15 pairs) who knew each other prior to the study were recruited from a university Facebook group. Six participants were excluded from the study, two due to an internet connection issue and the other four due to data saving issues. The remaining 24 participants (16 females; 8 males) ranged in age from 20 to 30 years of age ($M_{age} = 24.25$, $SD_{age} = 3.07$). All participants reported speaking Canadian English as their first language, although nine subjects reported being able to speak one language other than English fluently. All participants had normal or corrected to normal vision, had no concerns about their hearing, and had no history of speech or language impairments. They provided informed consent online prior to participating and were financially compensated for their time. All experimental procedures were approved by the Institutional Review Board at Queen's University.

Materials

Sixty-four images belonging to four different basic object categories were selected from Version 6 of the Open Images Dataset (Kuznetsova et al., 2018). Images were cropped to be equal in size and resolution (200×200 pixels). The object categories were birds, horses, bowls, and flowers.

For the RCT, only half of the stimulus set (i.e., 32 images) was shown to participants. The same 32 images (eight images in each of the four different basic object categories) were presented to each dyad during the RCT. In trials 1–3, eight birds and eight horses were presented, while in trials 4–6, eight bowls and eight flowers were presented. The remaining 32 visual stimuli (eight additional images in each of the four different basic object categories) were only shown to participants during the recognition memory task. In this task, participants were shown all 64 images in the dataset, 32 of which they had seen during the RCT and 32 of which they had not seen during the RCT. Out of the 32 images participants had previously seen during the RCT, 16 had been presented to them while they were playing the role of the Director in the RCT and 16 had been presented to them while they were playing the role of the Matcher in this task. In the recall memory task, participants were shown and asked to describe from memory the same 32 images that they had previously seen during the RCT.

Software

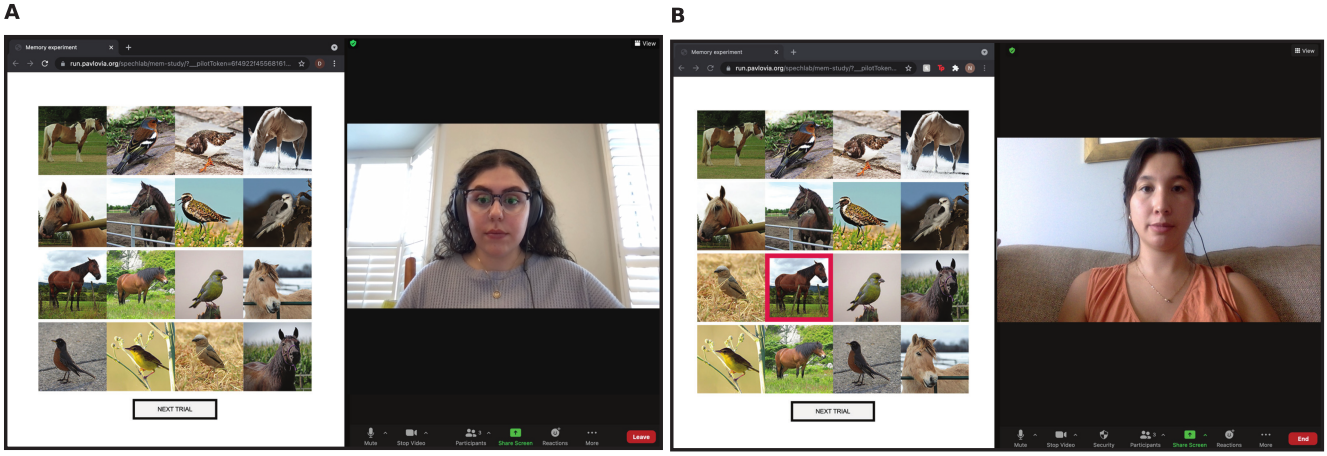
The RCT and the recognition memory task were both hosted by Pavlovia (Peirce et al., 2019), a website that can be used to host and run experiments online. The RCT was programmed using jsPsych (De Leeuw, 2015), a JavaScript-based web browser experiment builder. The recognition memory task was built using PsychoPy (Peirce et al., 2019), a Python-based application that has a builder interface. File management for both of the tasks was handled by the Gitlab repository. The recall memory task was administered through the online survey platform Qualtrics (Qualtrics^{XM}, Provo, Utah). The videoconference software platform Zoom was used as a virtual replacement for in-person interaction due to COVID-19 pandemic restrictions.

Procedure

Dyads joined the experimenter on a Zoom call from their own computer and in separate physical locations from each other. They were then familiarized with Zoom and instructed to disable their self-view to better mimic an in-person conversation. Before beginning the RCT, participants were asked to arrange their computer screens such that half of their screen showed the 4×4 matrix of images (i.e., the experiment browser window) and the other half of their screen showed their partner's video on Zoom. See Figure 1 for a visual display of the online RCT. The experimenter ensured that all participants wore headphones, were seated in a relatively quiet physical location, and disabled any sound notifications on their cell phones and laptops to reduce distraction during the experiment.

Figure 1

Computer Screen Setup for the RCT via Zoom



Note. The RCT is shown on the left side of the split-screen, while the Zoom window is shown on the right side of the split-screen. Participants cannot see themselves and can only see their partner. The Director (panel A) describes each image from top to bottom and from left to right to the Matcher (panel B) who clicks on the respective image in their matrix and swaps it in the correct position to match their partner. See the online article for the color version of this figure.

Instructions for the RCT task were provided by the experimenter using a demonstration image matrix. Each dyad was instructed that they would each play two different roles during the experiment (i.e., Director and Matcher). Participants were told that on any given trial they would each see 16 images in a 4 × 4 matrix appear on their respective screens, and the only difference between the images in their matrix and the images in their partner's matrix would be the order of the pictures. As Directors, participants were informed that their role was to describe each picture in their matrix one-by-one to their partner (i.e., the Matcher) from top to bottom and from left to right, so that their partner could rearrange their images into the same order. They were told that only the Matcher would have the ability to swap images in their matrix (i.e., the Director's matrix remained static). Participants were instructed to proceed to the next trial once all 16 images were described by the Director and swapped into their proper location in the matrix by the Matcher. The conversations held between dyads were recorded over Zoom and the audio and video of the experimenter was disabled during the experiment to minimize any possible distraction to the participants.

Participants took part in two rounds of the RCT, switching roles in-between rounds. Each round consisted of three trials of matching under the same role assignments. For each trial within the same round (i.e., Round 1: trials 1–3; Round 2: trials 4–6), the same 16 pictures were randomly reorganized in the 4 × 4 matrix again. Upon completion of the RCT, pairs of participants remained on Zoom with the experimenter and were provided with instructions for the recognition memory task. They completed the recognition task at a self-paced rate with 64 images being presented (i.e., 32 "old" from the RCT rounds (16 from each round) and 32 "new" that had not been seen in the experiment) one at a time, and in a randomized order. Each participant received a different order of image presentation. Participants were unaware that their memory would be tested in follow-up tasks. To reduce distraction and to eliminate any possible communication between pairs, participants were asked to mute themselves on Zoom while completing the recognition memory task. Dyads could not see each other while completing the task, as they were instructed to minimize the Zoom window and run the experiment in full screen.

Upon completion of the recognition memory task, participants were informed of a follow-up questionnaire that they had the option of completing a week later. All 12 dyads agreed to participate, although one subject did not complete it. Participants were not informed of what the task would entail and were simply told by the experimenter that they would be sent a Qualtrics survey to complete in 7 days. Previous investigations have shown that participants recall significantly more conversational content in recall memory tasks if they are provided with instructions informing them that their memory will be later tested (e.g., Stafford & Daly, 1984; Stafford et al., 1987). The survey asked each participant to recall word-for-word the most efficient description that they and their partner used to describe each of the 32 images presented to them during the RCT. The images were randomly presented, and participants were provided with the image itself along with an accompanying text box to enable entry of each of their responses. They were instructed to complete the survey individually without the aid of their partner.

Dependent Measures

Common Ground Formation. Following previous work (e.g., McKinley et al., 2017; Yoon & Brown-Schmidt, 2014), image description lengths in number of words served as the primary analysis tool for measuring common ground formation during the RCT. To determine description lengths, stereo audio files from each recorded conversation were transcribed verbatim by a professional transcription company (Scribie or iScribed) and the transcriptions were checked for accuracy by two research assistants. The total word count for each image was determined for Directors and Matchers and included all descriptive words and phrases, any lexical dysfluencies (e.g., like, um), and any backchanneling from the Matcher (e.g., "OK, got it"). Lexical dysfluencies were included in

the present analysis as they have been shown to reflect description difficulty or the degree of planning or cognitive effort that a talker exerts when formulating a description that can be understood by their partner (H. H. Clark & Fox Tree, 2002; H. H. Clark & Wasow, 1998; Fox Tree & Clark, 1997). They have also been shown to reflect the formation and use of common ground in RCTs, such as in the context of multiparty conversation (see Yoon & Brown-Schmidt, 2014). The word count did not include any information pertaining to location (e.g., "the next one is," "to the right of that is"), chatter that was unrelated to the task (e.g., inside jokes, side conversations), or any unfilled pauses between words (e.g., "this... bird is yellow" would be counted as four words; McKinley et al., 2017).

The strength of common ground formed between dyads for each image during the RCT was determined using the following equation:

$$\text{Common ground} = \frac{\text{T1} - \text{T3}}{\text{T1} + \text{T3}} \quad (1)$$

where T1 and T3 refer to the number of words used by the Director to describe an image in trial 1 and trial 3, respectively. This measure was initially proposed by Repp (1976) to index right- and left-ear advantages in dichotic listening while accounting for listeners' overall perceptual performance. Here, the measure was used to determine the relative level of common ground formed between dyads for each image in relation to the total number of words used by the Director to describe the image in trial 1 and trial 3 of the RCT (i.e., T1 + T3). Common ground formation for each image was thus normalized to Directors' overall performance in describing the image to their partner. Previous studies (e.g., H. H. Clark & Wilkes-Gibbs, 1986; McKinley et al., 2017) have used difference scores (i.e., T1−T3) to measure common ground formation during RCTs. However, difference scores do not control for individual variability in description lengths.

Recall Memory Performance. Word-for-word similarity and semantic similarity analyses were carried out to measure participants' performance on the recall memory task. The two measures examined the similarity between two sets of words: the verbal descriptions used by Directors to describe images in the final trial of matching in the RCT and Directors' and Matchers' recall memory for those descriptions about a week later ($M = 9.3$ days, min = 7 days, max = 14 days).

Word-for-word similarity scores were derived by using the word intelligibility scoring software Autoscore (Borrie et al., 2019). It is common in recall memory for conversation and speech intelligibility studies to use "loose" criteria when determining correct responses (e.g., Bamford & Wilson, 1979; Benoit & Benoit, 1988; Rosen & Corcoran, 1982; Thorndyke, 1977). In this method, words are scored as correct if the root is the same in target and response. Inflections are ignored because errors of agreement can occur as a response bias. Autoscore allows one to specify specific grammar and spelling rules to apply in determining loose correspondence between two sets of words. All default spelling and grammar rules were applied in determining accuracy scores (see Borrie et al., 2019 for a detailed description of each rule). Four additional spelling rules (colour/color, spikey/spiky/spikes/, leaf/leaves, sphere/spherical) were added to the default acceptable spelling list to improve the accuracy of Autoscore in scoring frequently used words. Autoscore provides the number of words correctly recalled for each description and this was converted to a proportion correct score by dividing the similarity score by the total number of words used by the Director to describe the image during the final trial of the RCT. The proportion correct score served as the dependent variable in mixed models to examine recall memory for conversation (see below).

Recent NLP developments have expanded the ability of computer models to obtain objective metrics of semantic textual similarity between two bodies of text. These methods make use of the concept of word and sentence embeddings, a numerical representation of text in a high-dimensional vector space in which words and phrases with similar meanings are "embedded" closely together. In computational linguistics, words can be represented as vectors. The use of words in communication differs with context (i.e., sentential, social, etc.). Word embedding is a computational technique to capture the context that the word is used in. Classical embedding methods such as latent semantic analysis (Landauer & Dumais, 1997) have used statistical methods based on the usage frequency of words to determine their embeddings. In recent years, embedding models based on artificial neural networks, such as word2vec (Mikolov et al., 2013) and GloVe (Global Vectors for Word Representation; Pennington et al., 2014), have gained widespread usage due to their increased performance when compared with human evaluations of several standard textual data sets (Conneau & Kiela, 2018). Most recently, models based on the deep neural network transformer architecture (Vaswani et al., 2017), such as BERT from Google research (Devlin et al., 2018), have shown breakthrough performance on a variety of NLP tasks, including assessing textual similarity.

In the current set of experiments, we used a pre-trained RoBERTa model (Liu et al., 2019) to obtain a measure of semantic recall memory performance. The RoBERTa model is an iterative improvement of the BERT model and has displayed very strong performance on a variety of semantic similarity tasks (Yang et al., 2020). Here, it was used to compute the level of semantic similarity between Directors' image descriptions during the final trial of the RCT and the Directors' and Matchers' recall memory for those same descriptions. Each description is represented numerically as a high-dimensional vector "embedding" with the pre-trained RoBERTa model. Cosine similarity is the measure most often used in NLP models to assess the degree to which words and phrases are semantically similar. It represents the angle (θ) between two distinct sentence vector embeddings (s_1 and s_2), and is defined using the following equation:

$$\text{CosSim}\,(s_1,\ s_2) = \cos(\theta) = \frac{s_1 \cdot s_2}{\|s_1\|\|s_2\|} \quad (2)$$

where the cosine similarity has a value of 1 for two identical vectors ($\theta = 0^{\circ}$) and a value of 0 for perpendicularly oriented vectors ($\theta = 90^{\circ}$). In the context of the current set of experiments, cosine similarity values closer to 1 indicate stronger semantic recall memory performance, while cosine similarity values closer to 0 indicate poorer semantic recall memory performance. The analysis was implemented using the sentence-transformers package in Python (Reimers & Gurevych, 2019).

Statistical Analyses

The primary analyses for both experiments involved linear mixed-effects modeling. All models were implemented using the lme4 package (V1.1-27; Bates, Mächler, et al., 2015) in R (R Core Team, 2020). This statistical approach allowed for variance

among participants and images to be entered as random-effects terms, and for the nesting of participants within groups to be considered. The maximal random-effects structure for each model was initially specified based on the set of rules proposed by Barr et al. (2013). Random intercepts were included for participants and images causing nonindependence in the data, and random slopes were included for within-unit predictors (Barr et al., 2013).

For each analysis, we refer to the model with the best fit to the data as the best-fit model. Best-fit models were determined via a "backward-fitting" model selection approach (Bates, Kliegl, et al., 2015). This involved first testing a model that included the maximal random-effects structure and all fixed-effects terms of interest based on the experimental design and research question. In successive models, fixed-effects terms were removed one at a time and models were compared for goodness of fit using likelihood ratio tests (LRTs). If a model did not converge, or if there was a singularity error, the maximal random effects structure for that model was simplified by removing random effects terms one at a time that explained minimal (or zero) variance (Barr et al., 2013). The random effect term that explained the least amount of variance was always removed first. This process always resulted in a best-fit model that satisfied convergence criteria and that outperformed all other models.

Best-fit models were established for each experiment to predict: (a) description lengths in number of words used by Directors to describe images in the RCT, (b) word-for-word recall memory, and (c) semantic recall memory for the referential labels used by Directors to describe images in the last trial of the RCT.

Control Experiment

One of the issues with using recall memory for image descriptions as an index of recall memory for conversation is that there are only a limited number of plausible words and phrases that can be used to describe each image. Hence, some baseline level of verbatim and semantic similarity will likely be observed between RCT image descriptions and recall memory descriptions elicited in a memory task independent of actual memory processes. We conducted two Control Experiments (one for Experiment 1; one for Experiment 2) to directly address this issue. Our aim was to determine whether there was evidence of true verbatim recall memory in the present experiments rather than just a similarity caused by the pictures evoking a description that was similar.

In the Control Experiment for Experiment 1, 12 fluent English-speaking participants (10 females; 2 males; $M_{age} = 21.58$, $SD_{age} = 2.39$) who did not participate in Experiment 1 were recruited from a local Facebook group. Participants individually joined a Zoom call with an experimenter and were asked to efficiently describe each of the 32 basic object category images that were presented to participants during the RCT and recall memory task in Experiment 1. Participants were also instructed to describe each image in the context of the other images presented to them. To mimic the Zoom setup used in Experiment 1, participants were asked to disable their self-view. The experimenter ensured that all participants were seated in a relatively quiet physical location and disabled any sound notifications on their cell phones and laptops to reduce distraction during the experiment. The Control Experiment was programmed using jsPsych (De Leeuw, 2015) and hosted online by Pavlovia (Peirce et al., 2019).

A visual depiction of the Control Experiment setup is shown in Figure 2. On one half of their computer screen, naïve participants were randomly presented with one image at a time and were provided with a text box to type in their description. On the other half of their computer screen, participants were shown the 4×4 matrix of images that were presented to dyads during round 1 and round 2 of the RCT in Experiment 1. This setup was used to evoke naïve image descriptions that would be most comparable to image descriptions used by Directors during the RCT. Naïve participants described each of the 16 images from round 1 first, followed by each of the 16 images from round 2.

Using Autoscore, we then computed the proportion of verbatim similarity between each of the naïve participants' image descriptions and each of the Directors' and Matchers' recall memory descriptions in Experiment 1. The idea is that the descriptions produced by the naïve participants act as an independent standard to compare the recalled memories against. The control descriptions share with the original Director descriptions that they were descriptions stimulated by each image and were made with all of the images visible. However, the Control Experiment descriptions are not shaped by a communicative process and thus, there is nothing unique to a particular conversation in the description. Thus, each recall memory description should be more similar to the one produced in the conversation a participant took part in than to generic picture descriptions.

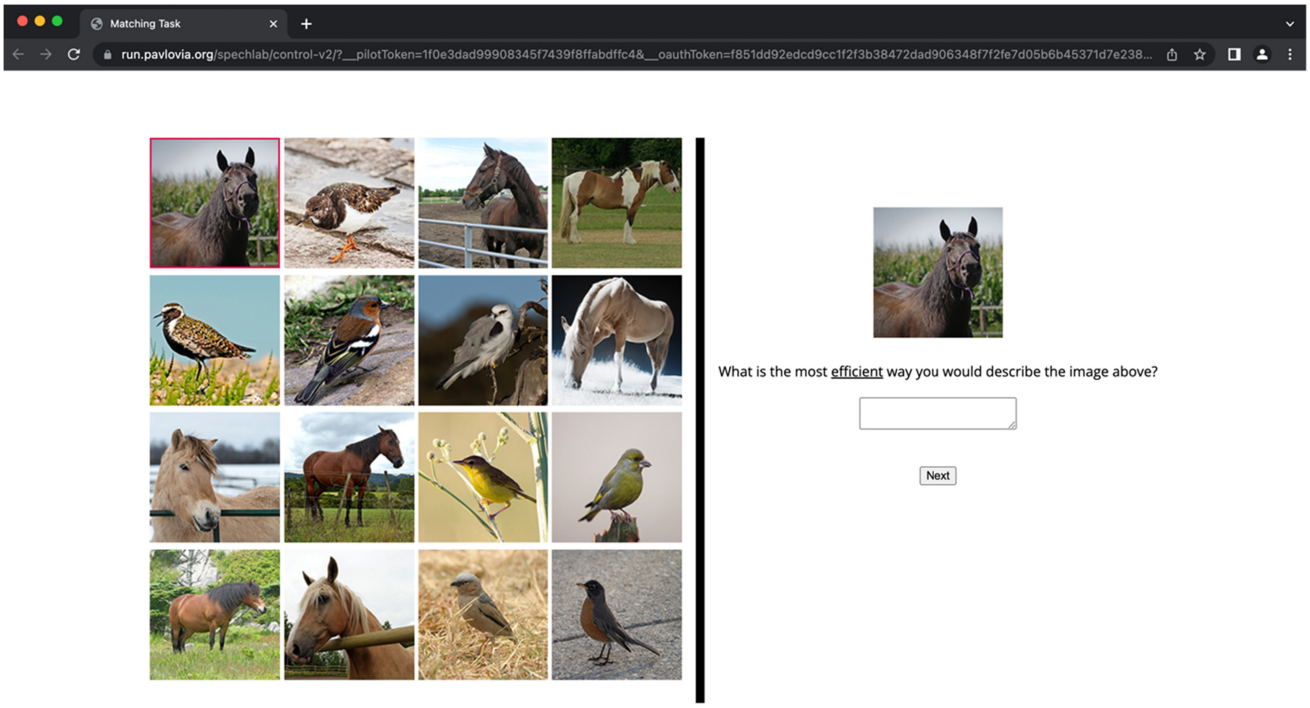
Due to there being no inherent match of a control description to a particular participant's recall, we created two separate distributions of similarity scores between the control data and the recall statements (i.e., Control vs. Directors; Control vs. Matchers). In essence, these were permutation tests with all 12 control descriptions for each image being compared with each of the Directors' and Matchers' recall memory descriptions for proportion of verbatim similarity (e.g., 12 control participants $\times$ 368 recall descriptions = 4,416 for the Director test). As with the recall memory analysis, the Autoscore value for each description comparison was converted to a proportion correct score by dividing the Autoscore value by the total number of words used by the control participant to describe the particular image. This was done to eliminate any differences in verbatim similarity caused by differences in description lengths. With this type of analysis, we were able to directly compare the means and confidence intervals for each control distribution of verbatim similarity scores to the means and confidence intervals of the true verbatim recall memory data from Experiment 1. This allows for a statement to be made about the probability that factors beyond image description influenced the verbatim recall among participants in Experiment 1.

Results

The primary dataset for Experiment 1 consists of 1,152 trials (24 Directors $\times$ 3 trials $\times$ 16 images per trial = 1,152) wherein the Director was tasked with describing a target image to the Matcher. On several of those trials, Matchers communicated to the Director whether or not they understood which image was being described. These Matcher utterances were not analyzed here. A total of 12 Director trials were eliminated as outliers (i.e., the number of words used by the Director to describe an image in these trials was greater than three standard deviations above the overall mean number of words used by Directors).

Figure 2

Computer Screen Setup for the Control Experiment in Experiment 1



Note. Each image that naïve participants were asked to describe (on the right) was pointed out to them with a red box surrounding it in the 4 × 4 matrix of images (on the left). Shown here are the 16 basic category object images from round 1 of the RCT in Experiment 1. Images retrieved from Version 6 of the Open Images Dataset, described in more detail in "The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale," by Kuznetsova et al., 2018., arXiv preprint, 1811 (https://doi.org/10.1007/s11263-020-01316-z). CC BY 2.0. See the online article for the color version of this figure.

Description Lengths

As shown in Figure 3, the average number of words used by Directors to describe images to Matchers in round 1 and round 2 decreased over the course of three trials of matching with the same images. Descriptive statistics for description lengths used by Directors and Matchers in the RCT in Experiment 1 are provided in Table 1.

The best-fit model predicting Directors' description lengths in the RCT produced the best fit to the data and included a random-effects structure with random intercepts for participants and images and correlated random slopes for participants by trial. It also included the fixed effects of trial and round, but not their interaction term. Fixed factor coefficients for the best-fit model were reliable and including the fixed effects terms significantly improved the best-fit model relative to a null model that only included the random effects structure, $\chi^2(2) = 44.75$, $p < .001$. An alternative model that included the interaction between trial and round did not significantly outperform the best-fit model, $\chi^2(1) = 3.48$, $p = .062$. The best-fit model was a significantly better fit to the data than alternative models that only included the fixed effect of trial, $\chi^2(1) = 7.54$, $p = .006$, or round, $\chi^2(3) = 422.46$, $p < .001$.

The best-fit model showed a significant trial effect, such that Directors used shorter description lengths over repeated trials of describing the same images. The best-fit model also yielded a significant round effect. Directors used longer labels to describe images in round 1 than in round 2 of the RCT. The results from the best-fit model predicting Directors' description lengths from the RCT in Experiment 1 are shown in Table 2.

Recognition Memory

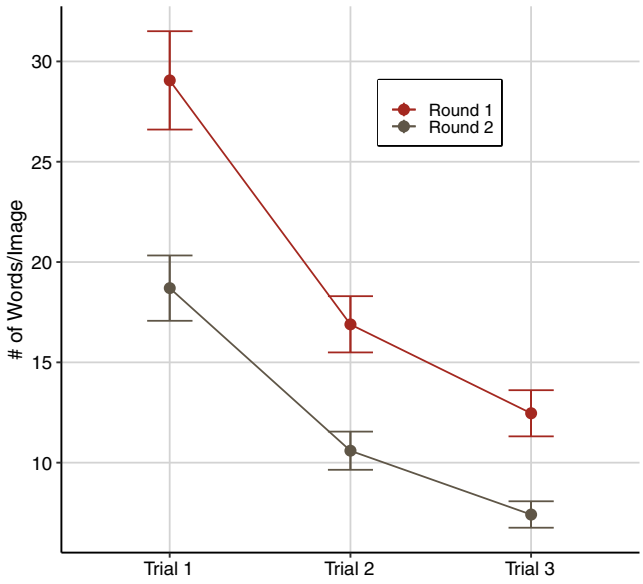
Ceiling levels of performance were observed in the recognition memory task. Irrespective of role, participants performed at an overall average of 98.4% accuracy (63/64; $SD = 1.8\%$). Participants correctly recognized an average of 97.4% (31.2/32; $SD = 3.5\%$) of the 32 images shown to them during the RCT (i.e., true stimuli), and correctly denied an average of 99.5% (31.8/32; $SD = 1.5\%$) of the 32 images not shown to them during the RCT (i.e., foil stimuli). There was no significant difference in participants' ability to recognize true versus foil stimuli, and there were no significant effects of round or role on recognition memory performance, all $ps > .05$. The interaction between round and role was also not significant, $p > .05$. Descriptive statistics for recognition memory performance are presented in Table 3.

Recall Memory

Two mixed models (one for verbatim proportion correct scores based on Autoscore, one for semantic similarity using RoBERTa)

Figure 3

Average Number of Words Used by Directors in Trials 1, 2, and 3 to Describe Basic Category Object Images in the RCT in Experiment 1



Note. Round 1 and round 2 are shown in light and dark, respectively. Error bars represent 95% confidence intervals. RCT = referential communication task. See the online article for the color version of this figure.

were built to examine the influence of role (i.e., Director/Matcher) and relative strength of common ground formation during the RCT on participants' recall memory performance. The relative strength of the common ground formation variable was grand-mean centered. The average reduction in the number of words used by directors to describe images from T1 to T3 across both rounds was 14.13 ($SD = 14.60$, min $= -21$, max $= 73$). The average level of common ground formed between dyads in relation to the maximum possible level of common ground that could have been achieved was 0.37 ($SD = 0.34$, min $= -2.33$, max $= 0.95$).

The best-fit model for verbatim similarity produced the best fit to the data and included a maximal random-effects structure with random intercepts and slopes for participants and images. It also included fixed effects of relative strength of common ground formation and role, but not their interaction term. Fixed factor coefficients for the best-fit model were reliable, and the best-fit model significantly outperformed a null model that only included the maximal random effects structure, $\chi^2(2) = 34.97$, $p < .001$. An alternative model that included the interaction term between the fixed effects did not yield a significantly better model fit than the best-fit model, $\chi^2(1) = 0.002$ $p = .96$. The best-fit model also significantly outperformed alternative models that did not include the fixed effect of role, $\chi^2(1) = 7.87$ $p = .005$, or relative strength of common ground formation, $\chi^2(1) = 26.87$, $p < .001$.

The best-fit model had a significant effect on relative strength of common ground formation. This effect indicates that verbatim recall memory performance was enhanced when stronger levels of common ground were formed between dyads for images during the RCT.[1] The effect of role was also significant (see Figure 4A). On average, Directors recalled 3.8% more of the same words they themselves used to describe images during the RCT as compared to Matchers. Best-fit model coefficients for verbatim recall memory performance are presented in Table 4.

Linear mixed-effects modeling for semantic recall memory performance revealed different results. The best-fit model again included a maximal random-effects structure with random intercepts and slopes for participants and images. However, it only included the fixed effect of role and did not include the centered fixed effect of relative strength of common ground formation. The best-fit model factor coefficients were reliable, and the best-fit model significantly outperformed a null model that only included the maximal random-effects structure, $\chi^2(1) = 22.51$, $p < .001$. An alternative model that included the centered fixed effects of role, relative strength of common ground formation, and their interaction term did not significantly outperform the best-fit model, $\chi^2(2) = 0.66$, $p = .719$. The best-fit model was also a significantly better fit to the data than an alternative model that only included the fixed effect of relative strength of common ground formation, $\chi^2(1) = 22.51$, $p < .001$.

The best-fit model yielded a significant effect on role. On average, Directors' recall memories had 6.2% higher semantic similarity to their own descriptions during the final trials of the RCT as compared to Matchers. Best-fit model coefficients for semantic recall memory performance are presented in Table 5, and a visual depiction of the model is shown in Figure 4B. Descriptive statistics for verbatim and semantic recall memory performance are presented in Table 6.

Table 1

Directors' and Matchers' Image Description Lengths During the RCT in Experiment 1

Round (trial)	Length	*SD*	*n*
Round 1 (trial 1)			
Director A	29.06	17.14	190
Matcher A	5.95	6.26	192
Round 1 (trial 2)			
Director A	16.89	9.81	190
Matcher A	3.40	4.78	192
Round 1 (trial 3)			
Director A	12.46	8.03	190
Matcher A	2.22	3.86	192
Round 2 (trial 1)			
Director B	18.70	11.38	190
Matcher B	6.74	7.95	192
Round 2 (trial 2)			
Director B	10.59	6.65	190
Matcher B	2.78	4.07	192
Round 2 (trial 3)			
Director B	7.41	4.61	190
Matcher B	1.99	2.58	192

Note. Length = average number of words used to describe each image; RCT = referential communication task.

[1] Correlation analyses determined that participants tended to exhibit superior verbatim recall memory for image descriptions when Directors used shorter image description lengths in trial 3 of the RCT. The correlation in Round 1 was −.595 and in Round 2 was −.783. Such correlations indicate that verbatim recall was influenced by description lengths. However, it does not preclude the explanation that underlying sociolinguistic processes that allow two talkers to efficiently accomplish the RCT produced a more robust memory trace. Disentangling the relative degree to which verbatim recall was impacted by description lengths and common ground formation would require more participants than we tested here, and a more formal theoretical proposal on the generating structure.

Table 2
Coefficients for the Best-Fit Linear Mixed-Effects Model Predicting Directors' Description Lengths During the RCT in Experiment 1

Fixed effects	Estimate (*SE*)	95% CI	*t*	η_p^2	*p*	Random effects	Variance (*SD*)
Intercept	32.74 (2.39)	[27.88, 37.61]	13.68			Participant	
Trial	−6.99 (0.76)	[−8.55, −5.42]	−9.24	.79	**<.001**	Intercept	95.63 (6.78)
Round	−5.66 (1.92)	[−9.55, −1.78]	−2.95	.18	**.005**	Trial	11.23 (3.35)
						Image	
						Intercept	10.76 (3.28)
						Residual	79.17 (8.90)

Note. Number of observations = 1,140; number of images = 32; number of participants = 24. RCT = referential communication task; CI = confidence interval. Bolded values indicate significance at the $p < .05$ level.

Image Analysis

Given the significant round effect in Directors' description lengths, and our inability to objectively control for image similarity in stimulus selection, we examined the association between image saliency (i.e., variability in description difficulty) and the relative strength of common ground formed among dyads during the RCT. Our reasoning was that, if image saliency had a significant impact on relative common ground formation, large positive correlations should be observed between the number of words that Directors used to describe images for the first time in trial 1 and the relative strength of common ground they formed with their partner for those same images.

The correlation between the average description lengths used across all Directors for each image in trial 1 and the average relative strength of common ground formed for those same images during the RCT in round 1 was not significant, $r(16) = -.043$, $p = .874$. There was a significant correlation between description lengths in trial 1 and relative common ground formation in round 2, $r(16) = .301$, $p = .022$. For comparison, we also computed the correlation between average description lengths in trial 1 and the common ground formation measure often used in previous studies (i.e., a simple difference score between trial 1 and trial 3). The correlations in round 1, $r(16) = .863$, and round 2, $r(16) = .828$, were highly significant, both $ps < .001$.

Control Experiment

As noted above, one participant did not complete the recall memory task. The two datasets from the permutation tests used to determine the proportion of verbatim similarity between naïve participants' image descriptions and Directors' and Matchers' recall memory descriptions in Experiment 1 thus contained a total of 4,416 proportion similarity scores (i.e., 23 Directors/Matchers × 16 images = 368 recall memory descriptions × 12 naïve participant combinations = 4,416). The means, standard deviations, and 95% confidence intervals for the two control distributions of proportion similarity scores are presented in Table 7. The descriptive statistics from the recall memory task in Experiment 1 are also shown in Table 7 for comparison.

Table 3
Descriptive Statistics for the Recognition Memory Task in Experiment 1

Variable	No. of images	Average % (*SD*)
Total	64	98.37 (1.81)
True stimuli	32	97.66 (3.23)
Foil stimuli	32	99.09 (2.35)
Director	16	98.18 (2.42)
Matcher	16	98.57 (2.25)

As can be seen, the overall mean of each control distribution of verbatim similarity scores is considerably lower than the overall mean of its respective comparison group from Experiment 1. The 95% confidence intervals associated with the mean of each control distribution are also well outside of the range of scores obtained from the recall memory task in Experiment 1. For example, the overall mean of the distribution comparing the level of verbatim similarity between control descriptions and Directors' recall memory descriptions was 0.224 ($SD = 0.221$). In comparison, the level of verbatim similarity observed in Experiment 1 between image descriptions used by Directors in the final trial of matching in the RCT and their recall memory descriptions was 0.414 ($SD = 0.244$). There is thus a mean difference of 0.19 (or 19%) between the two sets of verbatim similarity scores. A similar mean difference (0.16) exists in the Matcher comparison as well.

Discussion

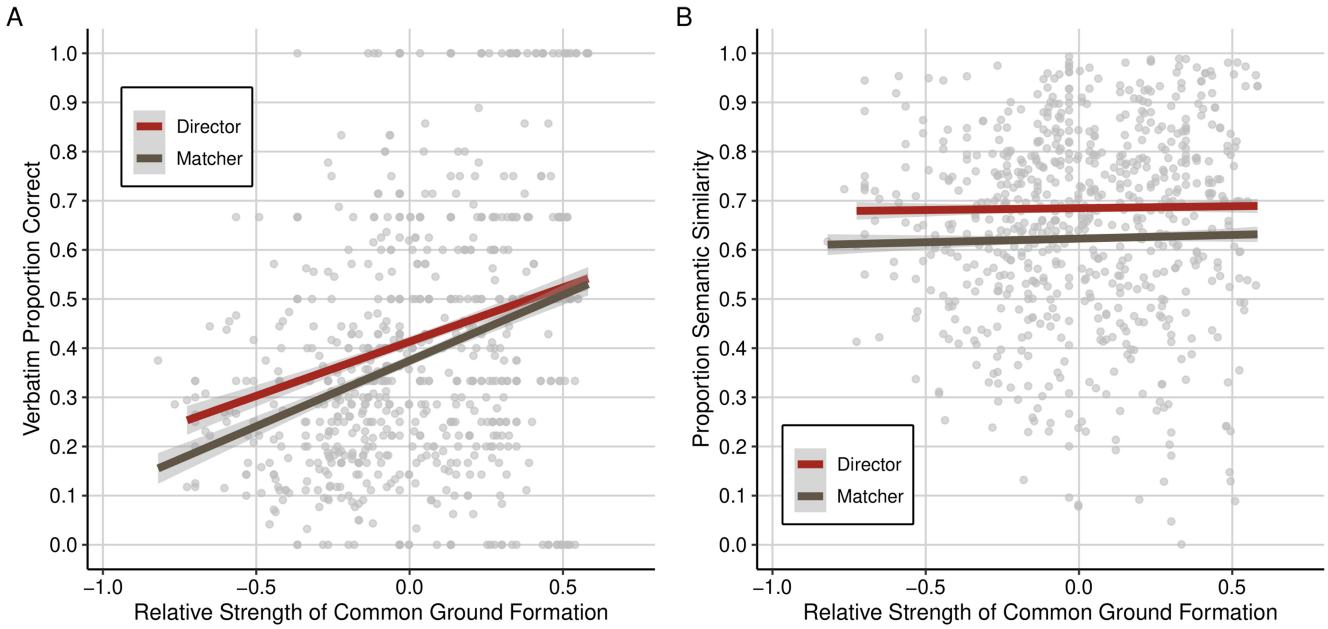
The results from the recognition memory task in Experiment 1 showed that participants performed with near-perfect accuracy. This is unsurprising given that humans have an exceptional capacity for image recognition generally (e.g., Standing, 1973) and particularly immediately following exposure to images in RCTs (McKinley et al., 2017). Nonetheless, these results provide evidence that participants were engaged during the online RCT and formed memories for images presented to them.

In the virtual RCT, dyads were shown to form common ground for basic category object images. Directors used shorter description lengths over the course of three trials of describing the same images to their partner over Zoom. These findings are consistent with several previous laboratory studies that have used RCTs to experimentally investigate the development of common ground in conversation (e.g., Brennan & Clark, 1996; H. H. Clark & Wilkes-Gibbs, 1986; Krauss & Weinheimer, 1964, 1966; McKinley et al., 2017; Schober & Clark, 1989). To our knowledge, this is the first study to provide evidence of common ground formation over videoconference technology using a RCT.

During the RCT, Directors were also shown to describe images with significantly less numbers of words in round two than in round one. There are several possible explanations for the round effect that cannot be distinguished by the present experiment. These explanations include individual differences in the strategies used by Directors in the two rounds to describe images to their

Figure 4

Best-Fit Linear Mixed-Effects Models Predicting Verbatim (A) and Semantic (B) Recall Memory Performance in Experiment 1



Note. Light and dark regression lines represent Directors and Matchers, respectively. The grey zone around each regression line indicates 95% confidence intervals. Grey dots indicate individual data points. For visualization purposes, four individual data points on the far left of the relative strength of common ground formation distribution are not shown. See the online article for the color version of this figure.

partner, task-specific learning of the RCT from being the Matcher in round one, and stimulus set differences between the two rounds. However, results from our image analysis revealed that the relative strength of common ground formation measure used in the present experiment limited the impact of stimulus set differences (i.e., image saliency, variability in description difficulty) on the reported results. While in round 2, there was a significant positive correlation between average description lengths used by Directors in trial 1 and the relative strength of common ground they formed with their partner, the correlation was not significant in round 1.

In the present experiment, participants' ability to recall word-for-word image descriptions used by Directors in the RCT was significantly enhanced when they formed relatively stronger common ground for images during the RCT. Importantly, permutation tests from the Control Experiment data indicated that Directors' and Matchers' verbatim recall memory scores were considerably higher than what would be expected if they had been simply redescribing the same images during cued recall. These results are consistent with McKinley et al. (2017), who reported that stronger common ground formation during a RCT led to significantly enhanced item and context recognition memory. However, surprisingly, the relationship between common ground formation during the RCT and semantic recall memory performance as assessed using RoBERTa was not significant. One possible explanation for this null finding is that RCTs are highly structured and place major limitations on the ability of dyads to engage in free-flowing, naturalistic conversation about a range of topics with varying semantic content. Participants' memory representations from the RCT are confined to specific images presented to them and described by Directors. Thus, their recall memory for those descriptions may not have included enough variation in semantic information that could be predicted by their relative strength of common ground

Table 4

Coefficients for the Best-Fit Linear Mixed-Effects Model Predicting Verbatim Recall Memory Performance in Experiment 1

Fixed effects	Estimate (*SE*)	95% CI	*t*	η_p^2	*p*	Random effects	Variance (*SD*)
Intercept	0.412 (0.025)	[0.36, 0.46]	16.77			Image	
Common ground	0.209 (0.034)	[0.14, 0.28]	6.16	.61	**< .001**	Intercept	0.0049 (0.070)
Role (Matcher)	−0.045 (0.016)	[−0.08, −0.01]	−2.81	.01	**.005**	Common ground	0.0071 (0.085)
						Participant	
						Intercept	0.0074 (0.086)
						Common ground	0.0050 (0.071)
						Residual	0.045 (0.211)

Note. Number of observations = 712; number of images = 32; number of participants = 23. CI = confidence interval. Bolded values indicate significance at the $p < .05$ level.

Table 5
Coefficients for the Best-Fit Linear Mixed-Effects Model Predicting Semantic Recall Memory Performance in Experiment 1

Fixed effects	Estimate (*SE*)	95% CI	*t*	η_p^2	*p*	Random effects	Variance (*SD*)
Intercept	0.682 (0.017)	[0.65, 0.72]	40.04			Image	
Role (Matcher)	−0.065 (0.013)	[−0.09, −0.04]	−4.80	.04	< **.001**	Intercept	0.0032 (0.056)
						Common Ground	0.019 (0.109)
						Participant	
						Intercept	0.0023 (0.048)
						Common Ground	0.000 (0.009)
						Residual	0.032 (0.179)

Note. Number of observations = 712; number of images = 32; number of participants = 23. CI = confidence interval. Bolded value indicates significance at the $p < .05$ level.

formation during the RCT. It is also possible that the pre-trained RoBERTa model used in the present study was not optimal in measuring levels of semantic similarity between image descriptions and recall memory for those descriptions. However, as previously noted, RoBERTa models have been shown to sensitively capture levels of semantic similarity in various textual datasets (e.g., clinical notes; see Yang et al., 2020).

The results from both the verbatim and semantic recall memory analyses in Experiment 1 revealed a significant conversational role effect. On average, Directors recalled significantly higher proportions of the same words they themselves used to describe images during the RCT than Matchers who did not generate the image descriptions. This trend in recall memory performance is suggestive of the generation/production effect in memory (Slamecka & Graf, 1978) and is consistent with previous findings in RCT recognition memory (McKinley et al., 2017).

In Experiment 2, more calibrated stimuli were used to address stimulus inequalities and task-specific learning that may have led to the round effect observed in Experiment 1. We return to some of the issues raised here in greater detail in our General Discussion.

Experiment 2

In Experiment 2, we extended our investigation of common ground and its influence on recall memory for conversation by introducing a task-irrelevant personal common ground manipulation. Using the same online RCT paradigm as Experiment 1 but with different stimuli, our aim was to test whether pairs of friends would establish common ground for images more efficiently and remember image descriptions more accurately than pairs of strangers. We hypothesized that the collection of shared lived experiences among pairs of friends (i.e., their personal common ground) would facilitate their capacity to form local common ground during the RCT. That is, we predicted that pairs of friends would describe images in the RCT more efficiently than pairs of strangers. While mixed findings have been reported in the literature (Boyle et al., 1994; Isaacs & Clark, 1987; Pollmann & Krahmer, 2018; Schober & Carstensen, 2009), it has been proposed that local conversational efficiency may be afforded by the collection of shared knowledge and beliefs that exist among conversational partners (e.g., Boyle et al., 1994; H. H. Clark, 1996; Isaacs & Clark, 1987; Rodrigues et al., 2021). Moreover, previous mixed findings may have been impacted by the use of conversational tasks that limited the opportunity for interlocutors to use their personal common ground effectively to their advantage (e.g., Boyle et al., 1994). Here, we predicted that pairs of friends would benefit from their shared experiences and common knowledge of each other when forming local common ground for facial images during a RCT. Given the paucity of research on the influence of conversational efficiency on memory representation, formation, and recall, we made no a priori predictions about whether friends or strangers would exhibit superior recall memory for RCT descriptions at one-week follow-up.

For Experiment 2, we used facial images from the open-source Glasgow Unfamiliar Face Database (GUFD; Burton et al., 2010) rather than images from basic object categories. This decision was primarily made to improve control over stimuli shown during the RCT. The GUFD includes similarity data that quantifies the average perceived similarity between any two identities in the database, which allowed for a more systematic stimulus selection process (see Methods section below) than in Experiment 1. We predicted that having better control over the level of similarity between stimuli in Experiment 2 would eliminate the round effect on Directors' description lengths, thus providing enhanced statistical power to detect any potential group differences in common ground formation

Table 6
Descriptive Statistics for the Recall Memory Task in Experiment 1

Type of recall	Role	*n*	*M* (*SD*)
Verbatim similarity (Autoscore)	Director	356	0.414 (0.244)
	Matcher	356	0.375 (0.261)
Semantic similarity (RoBERTa)	Director	356	0.685 (0.178)
	Matcher	356	0.620 (0.214)

Table 7
Descriptive Statistics From the Control Experiment Distributions Versus Descriptive Statistics From the Verbatim Recall Exhibited by Participants in Experiment 1

Control distributions	
Control vs. Director recall memory	Control vs. Matcher recall memory
$M = 0.224$	$M = 0.215$
$SD = 0.221$	$SD = 0.215$
95% CI [0.217, 0.230]	95% CI [0.209, 0.221]
$N = 4{,}416$	$N = 4{,}416$
Verbatim recall memory from Experiment 1	
Director recall memory	Matcher recall memory
$M = 0.414$	$M = 0.375$
$SD = 0.244$	$SD = 0.261$
95% CI [0.388, 0.439]	95% CI [0.348, 0.403]
$N = 356$	$N = 358$

and recall memory. We also chose facial images with the aim of providing friends with more salient opportunities during the RCT to benefit from using their pre-existing personal common ground. We reasoned that human faces would facilitate the elicitation of memories of shared experiences among pairs of friends that could reasonably improve their capacity to form local common ground.

Method

The methods for Experiment 2 generally match those of Experiment 1 and thus, only differences will be described.

Participants

Fifty-two participants (i.e., 26 pairs) that did not participate in Experiment 1 were recruited from a local Facebook group. Four participants (i.e., two pairs) were excluded because of technical issues with the Zoom audio recording. The remaining 48 participants ranged in age from 18 to 28. Twenty-four of them (12 pairs; 22 females; 2 males; $M_{age} = 22.71$; $SD_{age} = 2.01$) were recruited as part of the Friends group and were required to have known each other for at least six months prior to participating. The average friendship length among the 12 pairs of friends was five years ($SD = 2.68$ years, min = 1.5 years, max = 10 years). The other 24 participants (12 pairs; 18 females; 6 males; $M_{age} = 21.58$, $SD_{age} = 1.77$) were recruited as part of the Strangers group. They signed up individually and were randomly paired together by the experimenter with a partner they had never previously met. All participants reported speaking Canadian English as their first language, although 28 of them reported being able to speak fluently in at least one language other than English. All participants passed the same screening criteria as reported above (see Experiment 1). They also provided informed consent online prior to participating and were financially compensated for their time. All experimental procedures were approved by the Institutional Review Board at Queen's University.

Materials

The stimuli for Experiment 2 were selected from the GUFD (Burton et al., 2010; downloadable here: http://www.facevar.com/glasgow-unfamiliar-face-database). The GUFD was used to develop the Glasgow Face Matching Test to study unfamiliar face perception, recognition, and memory. It consists of multiple facial images of 304 individuals (132 females; 172 males) taken with two separate cameras at different angles. It also includes similarity data that quantifies the average perceived similarity between any two identities in the database. Similarity scores were obtained by Burton et al. (2010) by asking 30 participants (12 males; 18 females) to sort all 304 facial identities into piles according to their perceived similarity (see Bruce et al., 1999 for a full description of these methods). Scores range from 0 to 1 and represent the average frequency with which participants paired any two facial identities together into the same pile.

For this experiment, 64 different facial photos (32 female; 32 male) taken from the same angle (0 degrees) and camera (Olympus Camedia C-350 Zoom, 3 megapixel) were selected from the GUFD (see https://osf.io/hg7d3/files/ for the full list of stimuli numbers). Photos in the GUFD were edited by Burton et al. (2010) to remove all clothing and background cues. As in Experiment 1, only half of the stimulus set (i.e., 32 images) was shown to participants during the RCT. These 32 images were selected in four groups of eight (i.e., two groups of eight males; two groups of eight females) based on similarity scores. Within each group, faces were selected to be from a restricted range of perceived similarity to each other (.4 to .7). Sixteen images (eight male and eight female) were presented to the dyads during each round of the RCT with different sets being used in round 1 (trials 1–3) and round 2 (trials 4–6). The remaining 32 facial images were only shown to participants during the recognition memory task. These foil images were selected to be highly similar to those shown in the RCT. Each foil image had a 0.8 (i.e., 80%) perceived similarity to one of the images presented in the RCT. All images were cropped to be equal in size and resolution (200 × 200 pixels). During the recall memory task (as in Experiment 1), participants were randomly shown and asked to describe from memory the same 32 facial images that they had previously seen during the RCT.

Software

The RCT and the recognition memory task were programmed and hosted online in the same way as reported above for Experiment 1. To eliminate the possibility of data loss and to make data analysis more efficient, the recall memory task was also programmed using jsPsych (De Leeuw, 2015) and hosted by Pavlovia (Peirce et al., 2019).

Procedure

In Experiment 2, participants completed the recall memory task over Zoom with an experimenter present on the call, rather than on their own time via Qualtrics. This decision was made to eliminate the possibility of participants (particularly those in the Friends group) collaborating on the task and to have increased control over the amount of time that passed between when participants completed the RCT and recognition memory task in phase one and the recall memory task in phase two. While completing the recall memory task, participants were asked to mute their audio on Zoom and eliminate any possible distractions (e.g., sound notifications) in their environment.

Control Experiment

As in Experiment 1, we conducted a Control Experiment to investigate the extent to which verbatim similarity between RCT image descriptions and recall memory descriptions was due to true recall memory processes. Our expectation was that there would be some baseline level of verbatim similarity between image descriptions in our recall memory analysis resulting from images evoking similar descriptions and being describable in a limited number of ways. We recruited an additional twelve fluent English-speaking participants (10 females; 2 males; $M_{age} = 25.08$, $SD_{age} = 3.75$) who did not participate in Experiment 1 or Experiment 2 from a local Facebook group. The procedures and set up for the Control Experiment were the exact same as reported above for the Control Experiment in Experiment 1. The only difference was that participants were presented with the 32 facial images from the RCT in Experiment 2. Naïve participants described each of the 16 facial images from round 1 first, followed by each of the 16 facial images from round 2. We used Autoscore to compare each of the naïve participants' image descriptions to

each of the Directors' and Matchers' recall memory descriptions from the Friends and Strangers groups in Experiment 2.

Results

Directors described target images to Matchers on 2,304 trials (48 Directors × 3 trials × 16 images per trial = 2,304). Half of the images were described by Directors in the Friends group, while the other half were described by Directors in the Strangers group. A total of 46 Director trials were eliminated from the dataset, 27 of which were outliers (i.e., image descriptions longer than three standard deviations above the overall mean). An additional 19 trials were eliminated due to technical issues with Zoom audio recordings. Description lengths served as the primary analysis tool for measuring common ground formation and the same criteria were used to determine the total number of words used to describe each image by Directors and Matchers (see Experiment 1 Methods).

Description Lengths

As in Experiment 1, the average number of words used to describe images (in this case, faces) by Directors decreased over the course of three trials of matching with the same images (see Figure 5). This trend was observed across both rounds in the Friends and Strangers groups. Descriptive statistics for the description lengths used by Directors and Matchers in the RCT in Experiment 2 are provided in Table 8.

The random-effects structure for the best-fit model used to predict Directors' description lengths in the RCT had random intercepts for participants and images, and random correlated slopes for participants and images by trial. The best-fit model also included the fixed effects of group (Friends/Strangers) and trial, along with their interaction terms. Coefficients for the best-fit model were reliable, and the best-fit model significantly outperformed a null model that only included the random-effects structure, $\chi^2(3) = 89.39$, $p < .001$. Including the fixed effect of round and the three-way interaction between group, trial, and round in an alternative model did not lead to a significantly better model fit, $\chi^2(4) = 2.73$, $p = .60$. The minimal variance explained by the round effect likely reflects the increased level of control over the selection of stimuli in Experiment 2 as compared to Experiment 1. The best-fit model also had a significantly better fit to the data than alternative models that only included the fixed effect of trial, $\chi^2(2) = 9.49$, $p = .009$, or group, $\chi^2(3) = 210.08$, $p < .001$.

The results from the best-fit model showed a significant trial effect, providing evidence that common ground was being formed between dyads for facial images shown during the RCT. There was also a significant group effect. Directors in the Strangers group described images with more words than Directors in the Friends group. Across all trials, Directors in the Strangers group used an overall average of 8.13 more words than Directors in the Friends group to describe images in the RCT. The interaction between trial and group was not significant. Best-fit model coefficients are presented in Table 9.

Given the group effect in the primary analysis, we wanted to determine whether there was a relationship between friendship duration and the number of words used by Directors in the Friends group to describe images in the RCT. There was no significant correlation between friendship length and Directors' description lengths in Trial 1: $r(23) = .230$, $p = .291$, Trial 2: $r(23) = -.016$, $p = .942$, or Trial 3: $r(23) = .085$, $p = .693$.

Recognition Memory

Results from the recognition memory task in Experiment 2 were similar to Experiment 1 in that participants performed at ceiling

Table 8

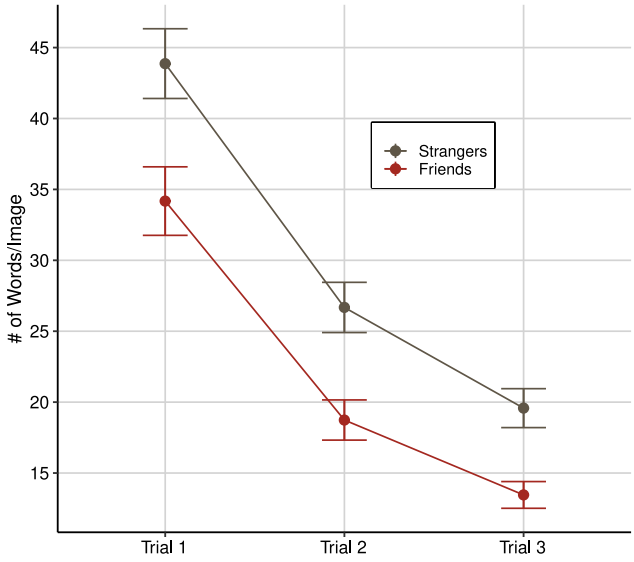
Directors' and Matchers' Image Description Lengths for the Friends and Strangers Groups in the RCT in Experiment 2

Round (trial)	Friends' length (*SD*)	*n*	Strangers' length (*SD*)	*n*
Round 1 (trial 1)				
Director A	34.92 (23.28)	173	41.32 (23.91)	191
Matcher A	8.89 (9.62)	174	8.48 (9.57)	189
Round 1 (trial 2)				
Director A	18.68 (13.10)	188	24.29 (16.37)	188
Matcher A	3.79 (4.77)	189	3.92 (5.84)	189
Round 1 (trial 3)				
Director A	12.92 (9.10)	187	17.27 (11.00)	192
Matcher A	1.99 (2.48)	188	1.59 (2.30)	186
Round 2 (trial 1)				
Director B	33.50 (23.56)	190	46.40 (24.82)	192
Matcher B	11.39 (11.87)	189	9.91 (11.09)	190
Round 2 (trial 2)				
Director B	18.43 (14.07)	190	29.04 (18.30)	190
Matcher B	4.81 (7.12)	190	4.43 (6.07)	189
Round 2 (trial 3)				
Director B	13.98 (9.50)	190	21.96 (15.46)	186
Matcher B	3.15 (4.32)	188	2.65 (3.64)	185

Note. Length = average number of words used to describe each image. RCT = referential communication task.

Figure 5

Average Number of Words Used by Directors in the Friends (Dark) and Strangers (Light) Groups to Describe Facial Images in Trials 1, 2, and 3 of the RCT in Experiment 2



Note. Data have been averaged across both rounds. Error bars indicate 95% confidence intervals. RCT = referential communication task. See the online article for the color version of this figure.

Table 9
Coefficients for the Best-Fit Linear Mixed-Effects Model Predicting Directors' Description Lengths During the RCT in Experiment 2

Fixed effects	Estimate (*SE*)	95% CI	*t*	η_p^2	*p*	Random effects	Variance (*SD*)
Intercept	43.21 (3.21)	[36.80, 49.63]	13.46			Participant	
Trial	−10.45 (0.94)	[−12.34, −8.56]	−11.06	.84	**<.001**	Intercept	153.01 (12.37)
Group (Strangers)	11.14 (3.97)	[3.15, 19.12]	2.81	.15	**.007**	Trial	9.16 (3.03)
Group × Trial	−1.61 (1.18)	[−3.98, 0.77]	−1.36	.04	.180	Image	
						Intercept	76.09 (8.72)
						Trial	5.96 (2.44)
						Residual	232.46 (15.25)

Note. Number of observations = 2,258; number of images = 32; number of participants = 48; number of groups = 2. RCT = referential communication task; CI = confidence interval. Bolded values indicate significance at the $p < .05$ level.

level. Irrespective of group membership, participants performed at an overall average of 98% accuracy (62.7/64; $SD = 2.4\%$). Given the minimal variance and near-perfect performance of participants on the recognition memory task, no further analyses were conducted. Descriptive statistics for the recognition memory performance are presented in Table 10.

Recall Memory

Two linear mixed-models (one for verbatim proportion correct scores based on Autoscore, one for semantic similarity using RoBERTa) were constructed to examine the influence of group (i.e., Friends/Strangers), role (i.e., Director/Matcher), and relative strength of common ground formation during the RCT on participants' recall memory performance. Descriptive statistics for verbatim and semantic recall memory performance are presented in Table 11. Note that here we are testing the effect of the change in common ground from trial 1 to trial 3 of the RCT. Across both rounds, there was no statistically significant mean difference in the *relative* strength of common ground formed between dyads in the Friends ($M = 0.38$, $SD = 0.34$) versus Strangers ($M = 0.36$, $SD = 0.29$) groups, $t(1,470) = 1.20$, $p = .230$.

The best-fit model for verbatim recall memory performance had a maximal random-effects structure that included random intercepts and slopes for participants and images. It also had the centered fixed effect of relative strength of common ground formation, along with the fixed effects of group and role, without their interaction terms. Coefficients for the best-fit model were reliable, and the best-fit model significantly outperformed a null model that only included the maximal random-effects structure, $\chi^2(3) = 32.28$, $p < .001$. An alternative model that included the interaction terms between all three fixed effects was not a significantly better fit to the data than the best-fit model, $\chi^2(4) = 4.94$, $p = .294$. The best-fit model significantly outperformed an alternative model that did not include the fixed effect of role, $\chi^2(1) = 19.82$, $p < .001$, and was marginally better than another model that did not include fixed effect of group, $\chi^2(1) = 3.12$, $p = .077$.

Results from the best-fit model indicate a significant effect of relative strength of common ground formation during the RCT. Dyads who established relatively stronger common ground for images during the RCT again exhibited superior verbatim recall memory performance at follow-up.[2] There was also a significant role effect (see Figure 6A). On average, Directors recalled 3.7% more of the same words they used to describe images in the final trials of matching than Matchers. However, the best-fit model did not show that personal common ground was significantly related to better verbatim recall (see Figure 6B). Best-fit model coefficients for verbatim recall memory performance in Experiment 2 are presented in Table 12.

The best-fit model predicting semantic recall memory performance in Experiment 2 had reliable coefficients but had minimal predictive capacity. It included random intercepts and correlated slopes for participants and images, and the fixed effect of group. The best-fit model was only marginally better than a null model that included just the maximal random-effects structure, $\chi^2(1) = 3.75$, $p = .053$. All other alternative models did not significantly outperform the best-fit model, all $ps > .05$. Results from the best-fit model showed a marginally significant effect of group (see Figure 7). On average, participants in the Strangers group recalled Directors' descriptions from the final trial of the RCT with 4.1% more semantic similarity than participants in the Friends group. Best-fit model coefficients for semantic recall memory performance are presented in Table 13.

The best-fit model used to predict Directors' raw image description lengths during the RCT revealed that there was a significant group effect. Directors in the Friends group were significantly more efficient in their use of words to describe facial images during the RCT than Directors in the Strangers group (see Figure 5). However, the relative strength of common ground formation variable used to predict participants' recall memory performance controls for these differences in raw image description lengths. Thus, two

Table 10
Descriptive Statistics (With Standard Deviations) for the Recognition Memory Task in Experiment 2

Variable	No. of images	Friends average % (*SD*)	Strangers average % (*SD*)
Total	64	97.79 (2.48)	98.30 (2.44)
True stimuli	32	97.66 (2.95)	98.37 (3.25)
Foil stimuli	32	97.92 (3.28)	98.37 (2.64)
Director	16	97.92 (3.53)	97.83 (4.46)
Matcher	16	97.40 (3.65)	98.91 (2.42)

[2] As in Experiment 1, correlation analyses determined that participants tended to exhibit superior verbatim recall memory performance when Directors used shorter image description lengths in trial 3 of the RCT. The correlation in Round 1 was −.465 and in Round 2 was −.290.

Table 11
Descriptive Statistics for the Recall Memory Task in Experiment 2

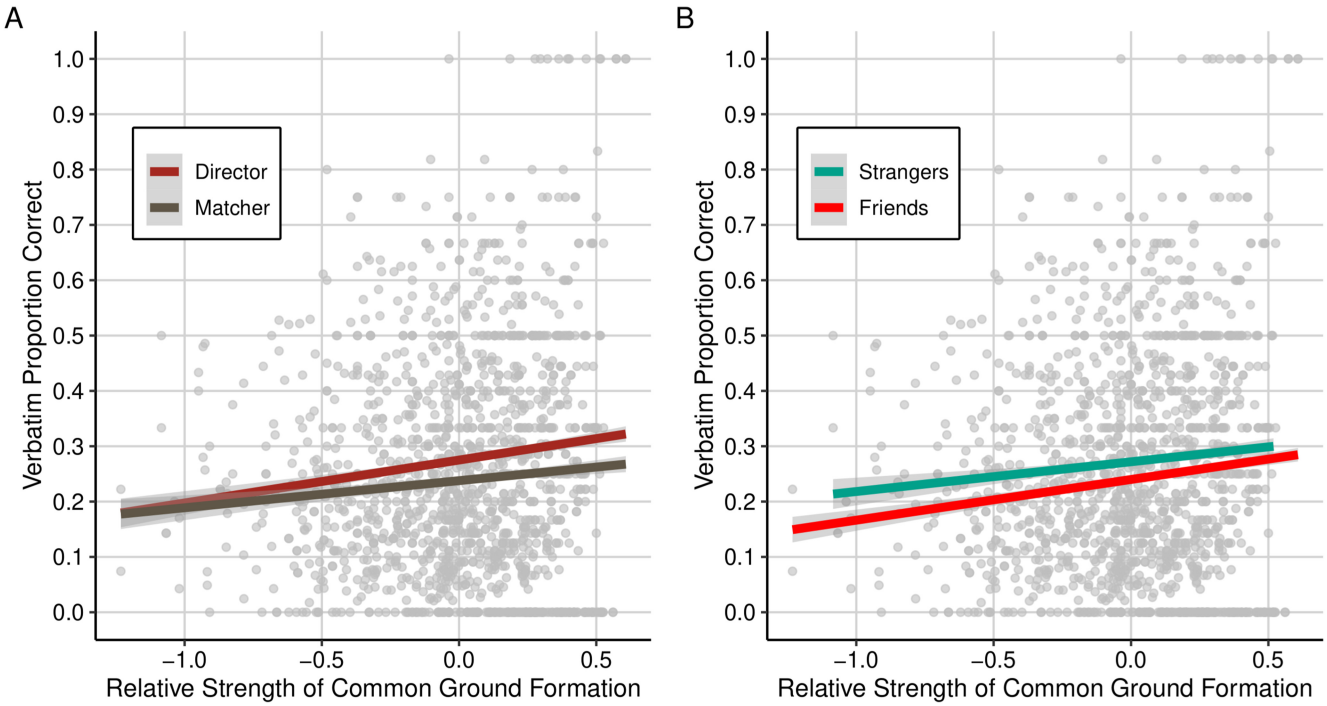
Type of recall	Group	Role	*n*	Mean (*SD*)
Verbatim similarity (Autoscore)	Friends	Director	359	0.261 (0.211)
		Matcher	359	0.221 (0.200)
	Strangers	Director	377	0.289 (0.194)
		Matcher	377	0.254 (0.190)
Semantic similarity (RoBERTa)	Friends	Director	359	0.529 (0.198)
		Matcher	359	0.511 (0.217)
	Strangers	Director	377	0.566 (0.179)
		Matcher	377	0.561 (0.197)

multiple linear regressions (one for verbatim recall, one for semantic recall) were carried out to determine whether there was an association between Directors' raw image description lengths during the RCT and participants' recall memory performance in the Friends and Strangers groups. In both regression models, the independent variables were (a) Directors' average description lengths for each image across all three trials of the RCT and (b) group membership (dummy coded variable with the Friends and Strangers groups coded as 1 and 0, respectively). Adding the interaction term between Directors' average description lengths and group membership to both regression models led to a high degree of multicollinearity (Variance Inflation Factor > 10). The interaction term was thus removed from both regression models. Participants' average verbatim and semantic recall memory performance for each image was entered as the dependent variable in each analysis, respectively. The dataset used to construct each regression model consisted of 64 data points (32 images described and recalled by participants in the Friends and Strangers groups).

A significant regression equation was found in the multiple regression analysis used to predict participants' average verbatim recall memory performance, $F(2, 63) = 11.19$, $p < .001$. Overall, the model accounted for 26.8% of the variance in verbatim recall memory performance. There was a significant negative association between Directors' average description lengths during the RCT and participants' average verbatim recall memory performance, $b = -0.006$, $p < .001$. For every one-word reduction in Directors' average description lengths during the RCT, verbatim recall memory performance increased by approximately 0.6%. There was also a significant negative association between group membership and verbatim recall memory performance, $b = -0.077$, $p < .001$. Participants in the Friends group demonstrated 7.7% lower verbatim recall memory performance than participants in the Strangers group for every one-word reduction in Directors' average description lengths during the RCT.

There was also a significant regression equation in the multiple regression analysis used to predict participants' average semantic recall memory performance, $F(2, 63) = 10.43$, $p < .001$. The model accounted for 25.5% of the overall variance in semantic recall memory performance. There was a significant negative association between Directors' average description lengths during the RCT and participants' semantic recall memory performance, $b = -0.003$, $p = .005$. Participants' semantic recall memory performance

Figure 6
Best-Fit Linear Mixed-Effects Model Predicting Verbatim Recall Memory Performance in Experiment 2



Note. Panel A shows the significant role effect. Directors and Matchers are plotted as light and dark regression lines, respectively. Panel B shows the nonsignificant group effect. Strangers and Friends are shown in light and dark regression lines, respectively. The grey zone around each regression line indicates 95% confidence intervals. Grey dots indicate individual data points. See the online article for the color version of this figure.

Table 12
Coefficients for the Best-Fit Linear Mixed-Effects Model Predicting Verbatim Recall Memory Performance in Experiment 2

Fixed effects	Estimate (*SE*)	95% CI	*t*	η_p^2	*p*	Random effects	Variance (*SD*)
Intercept	0.254 (0.020)	[0.21, 0.29]	12.41			Image	
Common ground	0.065 (0.020)	[0.02, 0.11]	3.18	.27	**.004**	Intercept	0.002 (0.045)
Group (Strangers)	0.045	[−0.06, −0.02]	1.81	.01	.077	Common ground	0.004 (0.062)
Role (Matcher)	−0.041	[−0.01, 0.10]	−4.50	.06	< **.001**	Participant	
						Intercept	0.008 (0.087)
						Common ground	0.002 (0.048)
						Residual	0.030 (0.172)

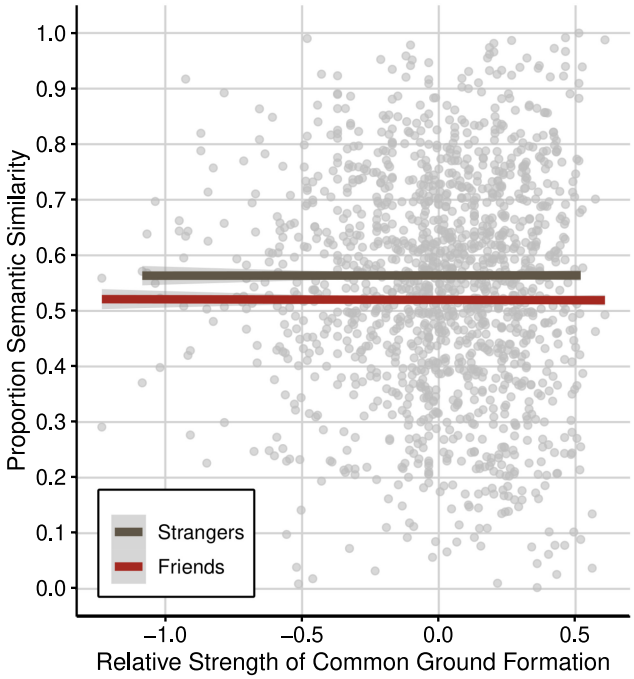
Note. Number of observations = 1,472; number of images = 32; number of participants = 48; number of groups = 2. CI = confidence interval. Bolded values indicate significance at the $p < .05$ level.

increased by approximately 0.3% for every one-word reduction in Directors' average description lengths during the RCT. There was also a significant association between group membership and semantic recall memory performance, $b = -0.072$, $p < .001$. For every one-word reduction in Directors' average description lengths during the RCT, participants in the Friends group demonstrated 7.2% lower semantic recall memory performance than participants in the Strangers group.

Image Analysis

Despite the elimination of the round effect and our objective control over the selection of stimuli in Experiment 2, we again conducted an image analysis to quantify the potential impact of image saliency on common ground formation during the RCT. The correlation between Directors' average description lengths in trial 1 and the relative strength of common ground formed among dyads for each image in round 1 was not significant, $r(16) = .249$, $p = .353$. The correlation in round 2 was significant, $r(16) = .568$, $p = .022$. For comparison, the correlation between Directors' average description lengths in trial 1 and a simple difference score measure of common ground formation was highly significant in trial 1, $r(16) = .870$ and trial 2, $r(16) = .897$, both $ps < .001$.

Figure 7
Best-Fit Linear Mixed-Effects Model Predicting Semantic Recall Memory Performance in Experiment 2



Note. Strangers and Friends are plotted as light and dark regression lines, respectively. The grey zone around each regression line indicates 95% confidence intervals. Grey dots indicate individual data points. See the online article for the color version of this figure.

Control Experiment

The datasets from the four permutation tests used to determine the proportion of verbatim similarity between naïve participants' image descriptions and Directors' and Matchers' recall memory descriptions from the Friends and Strangers groups in Experiment 2 each contained a total of 4,608 proportion similarity scores (e.g., 24 Friends Directors × 16 images = 384 recall memory descriptions × 12 naïve participant combinations = 4,608). Descriptive statistics for each control distribution of the proportion similarity scores are presented in Table 13. For comparison, descriptive statistics from the recall memory task in Experiment 2 are also presented in Table 14.

As was shown in the control data for Experiment 1, the mean verbatim proportion similarity scores in each of the control distributions are substantially lower than those produced by Directors and Matchers in the recall memory task in Experiment 2. This is true for both the Friends and Strangers groups. For instance, the mean verbatim similarity score in the control distribution for Directors in the Strangers group was 0.146 ($SD = 0.125$). In contrast, the mean verbatim similarity score among Directors in the Strangers group in Experiment 2 was 0.289 ($SD = 0.194$). This amounts to a difference of 0.143 (or 14.3%). Similar mean differences are also evident in the Matcher data. These results indicate that, on average, participants demonstrated a considerable degree of verbatim recall memory for the image descriptions used by Directors during the final trial of the RCT. The verbatim similarity scores obtained by participants in Experiment 2 were not simply a result of them re-describing the same images upon cued recall.

Discussion

As previously discussed in Experiment 1, the recognition memory results from Experiment 2 suggest that participants paid sufficient attention during the online RCT and formed memories for the presented stimuli. Regardless of group membership or the role that

Table 13
Coefficients for the Best-Fit Linear Mixed-Effects Model Predicting Semantic Recall Memory Performance in Experiment 2

Fixed effects	Estimate (*SE*)	95% CI	*t*	η_p^2	*p*	Random effects	Variance (*SD*)
Intercept	0.524 (0.016)	[0.49, 0.56]	33.06			Image	
Group (Strangers)	0.041 (0.021)	[0.00, 0.08]	1.96	.08	.057	Intercept	0.001 (0.032)
						Common Ground	0.007 (0.086)
						Participant	
						Intercept	0.004 (0.065)
						Common Ground	0.001 (0.027)
						Residual	0.033 (0.183)

Note. Number of observations = 1,472; number of images = 32; number of participants = 48; number of groups = 2. CI = confidence interval.

participants played while interacting with the facial stimuli, they correctly recognized nearly all the images they were presented during the RCT in an immediate recognition memory task.

The results from the virtual RCT in Experiment 2 indicated that dyads formed common ground for facial images. There was a significant trial effect across both the Friends and Strangers groups, meaning that Directors used shorter referential labels to describe faces to their partner over three trials. Interestingly, there was also a significant group effect. Directors in the Friends group described facial images with significantly less numbers of words than Directors in the Strangers group in all three trials across both rounds of the RCT. In other words, Directors in the Friends group began and ended each round by describing faces more concisely than Directors in the Strangers group. One possible explanation for the group effect is that Directors in the Friends group may have been able to describe faces more concisely to their partner by drawing connections between some of the faces presented to them and individuals they both knew outside the experimental context. Directors in the Friends group may also have had an enhanced ability to read their partner's non-verbal cues (e.g., facial expressions) and thus, detect when their partner had been given sufficient information to identify faces in their matrix. We discuss this further in our General Discussion section.

Table 14
Descriptive Statistics From the Control Experiment Distributions Versus Descriptive Statistics From the Verbatim Recall Exhibited by Participants in Experiment 2

Control distributions			
Control versus Director Recall Memory (Friends)	Control versus Matcher Recall Memory (Friends)	Control versus Director Recall Memory (Strangers)	Control versus Matcher Recall Memory (Strangers)
$M = 0.116$	$M = 0.111$	$M = 0.146$	$M = 0.140$
$SD = 0.117$	$SD = 0.101$	$SD = 0.125$	$SD = 0.126$
95% CI [0.113, 0.12]	95% CI [0.108, 0.114]	95% CI [0.142, 0.150]	95% CI [0.137, 0.144]
$N = 4{,}608$	$N = 4{,}608$	$N = 4{,}608$	$N = 4{,}608$
Verbatim recall memory from Experiment 2			
Director Recall Memory (Friends)	Matcher Recall Memory (Friends)	Director Recall Memory (Strangers)	Matcher Recall Memory (Strangers)
$M = 0.261$	$M = 0.221$	$M = 0.289$	$M = 0.254$
$SD = 0.211$	$SD = 0.200$	$SD = 0.194$	$SD = 0.190$
95% CI [0.239, 0.282]	95% CI [0.200, 0.241]	95% CI [0.269, 0.308]	95% CI [0.235, 0.273]
$N = 359$	$N = 359$	$N = 377$	$N = 377$

The best-fit model used to predict Directors' description lengths during the RCT did not reveal any significant interactions between group membership and trial number. In other words, there was no significant difference in the relative degree to which Directors in the Friends and Strangers groups shortened their descriptions for facial images over time. This can be clearly seen in Figure 5, where the slopes pertaining to the average reduction in the number of words used by Directors in both groups to describe images over three trials of matching in the RCT are nearly identical. One interpretation of this pattern of results is that the description efficiency supposedly afforded by friendship (i.e., personal common ground) was independent of the rate by which dyads in the Friends group formed local common ground for facial images overtime during the RCT. This underscores the multidimensional nature of common ground and its ability to influence discourse memory in a variety of ways. Personal and local common ground seemed to have differential impacts on the ability for dyads to form referential labels for images and recall those labels from memory to use in subsequent trials.

In Experiment 2, we systematically controlled for the level of similarity among presented stimuli. This was carried out to eliminate a possible round effect in Experiment 2 and provide a better explanation for the significant round effect observed in Experiment 1. In the present experiment, Directors were shown to describe facial images with similar numbers of words in both rounds of the RCT. This suggests that the round effect shown in Experiment 1 was likely due to there being stimulus differences in the image sets that were presented to dyads in the two rounds.

The findings from the recall memory analyses in Experiment 2 again revealed a significant influence of common ground formation on verbatim—but not semantic—recall memory performance. The greater the relative strength of common ground formed between dyads in either group for facial images during the RCT, the greater their ability to recall word-for-word image descriptions 6–7 days later. This finding serves as a replication of the effect observed in Experiment 1. However, the semantic recall memory model did not show a significant effect on relative strength of common ground formation. This again highlights the potential drawbacks of the structured nature of the RCT as compared to naturalistic conversation and/or the NLP model RoBERTa used to measure semantic similarity in the current set of experiments.

The linear mixed models used to predict verbatim and semantic recall memory performance did not reveal a statistically significant group effect. Contrary to our hypothesis, participants in the Friends group did not exhibit superior recall memory performance than participants in the Strangers group. In fact, participants in the

Strangers group tended to remember more verbatim and semantic information from the descriptions that Directors used in the final trials of matching in the RCT than participants in the Friends group. Secondary analyses also revealed that, on average, as Directors' description lengths decreased, participants in the Strangers group exhibited significantly higher verbatim and semantic recall memory performance than participants in the Friends group. This surprising finding suggests that the number of words used in a description is a complex measure of common ground. While changes in description length within the task positively predict better memory, a more efficient description based on familiarity does not translate into enhanced recall. Potential reasons for this are discussed below.

The permutation tests conducted using the Control Experiment data furnished convincing evidence that the image descriptions provided by participants in the recall memory task in Experiment 2 contained true recall memory components for the specific conversations. There were considerably lower mean levels of verbatim similarity between the control descriptions and Directors' and Matchers' recall memory descriptions than there was between the Directors' image descriptions during the final trials of matching in the RCT and Directors' and Matchers' recall memory descriptions (see Table 7). The 95% confidence intervals associated with the mean of the four control distributions were also well below the range of observed similarity scores among participants in Experiment 2. Thus, it is very unlikely that participants in Experiment 2 were only redescribing facial images based on their physical characteristics. Rather, their descriptions seemed to have contained a portion of verbatim content that was influenced by conversational forces during the RCT. In fact, mean difference scores between the control distributions and true recall memory distributions indicate that participants in Experiment 2 recalled anywhere between 11% and 14.5% of true verbatim conversational content. These performance levels are similar to previous studies that have investigated verbatim recall memory for conversation (e.g., Stafford & Daly, 1984 who showed that participants recalled an average of 10% of information from social interaction with a partner).

It is important to acknowledge, however, that there were considerable levels of verbatim similarity observed among the four control distributions. This suggests that only a portion of the verbatim similarity being observed in the current set of experiments is reflective of true recall memory for conversation. Images can only be described in so many ways based on their distinctive characteristics, and participants viewed each of the images while describing them during the RCT and while they were asked to recall image descriptions at follow-up. This highlights the importance of running control experiments in investigations such as these to delineate true verbatim recall memory from verbatim similarity that is evoked from other task constraints.

General Discussion

In the present experiments, conversational dyads formed common ground for basic category object images (Experiment 1) and facial images (Experiment 2) in a virtual RCT conducted over Zoom. Over the course of three trials of describing the same images to their partner (i.e., the Matcher), Directors in both experiments were shown to use progressively shorter referential labels to refer to images in their matrix. These general findings serve as an online replication of several previous laboratory studies that have used RCTs to experimentally investigate the use of referential expressions and the development of common ground in conversation (e.g., Brennan & Clark, 1996; H. H. Clark & Wilkes-Gibbs, 1986; Knutsen et al., 2018; Krauss & Weinheimer, 1964, 1966; McKinley et al., 2017; Schober & Clark, 1989; Van der Wege, 2009). They provide evidence that the RCT can be reliably administered over videoconference technology and suggest that conversational partners form local common ground for images in a similar fashion in virtual environments as compared to in-person settings.

The results from the present experiments partially support the view that common ground and memory are intricately linked (e.g., H. H. Clark & Marshall, 1981; Horton & Gerrig, 2005, 2016; McKinley et al., 2017). Participants who formed relatively stronger levels of local common ground for images with their partner during a RCT were shown to exhibit superior verbatim recall memory for image descriptions used by the Director about a week later. This was true both when participants were asked to form local common ground and recall image descriptions for basic category object images in Experiment 1 and for facial images in Experiment 2. These significant findings serve as an extension of previous work in recognition memory (McKinley et al., 2017) and suggest that the strength of local common ground formed between interlocutors during conversation is an important predictor of their ability to access verbatim conversational content from memory.

Our hypothesis was that we would observe an even stronger relationship between local common ground formation and semantic recall memory for conversation. It has previously been shown that humans have a limited ability to recall verbatim content from previous conversational interactions (e.g., Neisser, 1981; Stafford & Daly, 1984). Using a free recall method, Stafford and Daly (1984) reported that subjects only remembered an average of 10% of what was said in a conversational exchange five minutes after. It is also thought that memory representations of discourse more strongly reflect the overall meaning or "gist" of the discourse rather than verbatim words and phrases (Bock & Brewer, 1974; Sachs, 1967). However, our findings from the present experiments did not support our hypothesis. No significant associations were observed in either experiment between the relative strength of local common ground formed between dyads for images during a RCT and their semantic recall memory performance at follow-up. As previously noted, we believe this null finding likely reflects the structured nature of the RCT as compared to naturalistic conversation. Another possible explanation is that recall memory was influenced by modality differences in encoding and retrieval. In the present experiments, participants encoded memories for image descriptions based on spoken language. At follow-up, participants recalled their memories for the same descriptions by typing/writing their responses. While this design facilitated our ability to conduct the experiments online and analyze the recall memory data in an efficient manner, modality effects may have played a role. Spoken language has more disfluencies, more partial sentences, and is less formal than written language. Written language has the benefit of editing and is thus less constrained by time in production. However, the vast majority of memory studies use written response for ease of scoring (Putnam & Roediger, 2013). Putnam and Roediger (2013) also directly addressed the issue of response modality in the context of the "testing effect" by manipulating whether responses were typed or spoken in either the first or last memory tests. They found no consistent advantage for either mode of response (Putnam & Roediger, 2013).

While less likely, it is also possible that there was a shortcoming with the NLP model RoBERTa that was used in the present set of experiments to obtain a measure of semantic recall memory. The RoBERTa model is pre-trained on the English language using textual information from sources like Wikipedia and BooksCorpus (Zhu et al., 2015), a dataset involving over 10,000 open-sourced online books (see Liu et al., 2019 for more detailed information). While this training set is comprehensive, it does not include a database of spoken language, which differs from written language in many ways (e.g., Redeker, 1984). For instance, spoken language is usually less formal and includes more repetitions, corrections, and dysfluencies than written language. Written language is more planned, less interactive, and designed for a wider audience. Spoken language is more spontaneous and intended for smaller, specific audiences. Given that we used RoBERTa to compare verbal (i.e., Directors' RCT descriptions) and textual (i.e., recall memory descriptions) image descriptions for semantic similarity, it is possible that the model did not fully capture the semantic overlap between the two types of descriptions. Two counter arguments can be raised to this concern. First, the verbatim similarity was assessed here using a comparison between written and spoken language and a relationship was demonstrated between common ground and recall. Second, it is believed that spoken and written language processing converge for higher linguistic and semantic processing (e.g., Wilson et al., 2018). One potential solution to this problem may be the development of a specialized NLP model that is pre-trained exclusively on information from spontaneous spoken language. Recent developments in NLP have allowed researchers to pre-train models on various sorts of information to solve different types of specialized tasks (e.g., scientific text; see Beltagy et al., 2019).

The structured nature of the RCT has the advantage of a high degree of experimental control to empirically examine conversational mechanisms like common ground formation. However, one likely consequence of this control is that it constrains the nature of the conversations and thus, the memory representations that individuals encode from the interaction. In natural conversations, participants coordinate their dialogue locally in adjacency pairs. In each turn, a speaker links their contribution to their partner's turn to maintain coherence. Across a series of turns, the collocutors cooperate to maintain one or more topics. The experimental task used here is a type of conversation, but one without key attributes of spontaneous conversation. For example, there is no topic per se. The analysis is based on only the Director's turns and there is no necessary link between the adjacency pairs in the analysis. During the RCT, Directors are tasked with repeatedly describing the same series of images with a limited number of characteristics to their partner (i.e., the Matcher). With practice, dyads learn that the optimal way to complete the task is for the Director to refer to images using similar words and phrases in each trial. As a result, participants are likely to pay close attention to and form memory representations for the verbatim words and phrases being used to describe each image. There is little semantic structure for the NLP model to represent. This may help to explain why in the current experiments the relative strength of local common ground formed between dyads for images during the RCT did not significantly predict their semantic recall memory performance. Unless Directors opted to describe an image using information outside of the experimental context, the memory representations encoded by participants for image descriptions would not have been very rich or meaningful. Rather, they would mostly have contained verbatim words or phrases that differentiated each image from the rest. It is also important to consider that the measure of common ground formation used here and in other RCT studies focuses heavily on the efficiency of conversation, and that is only one part of the evidence that people share a common understanding. The use of this measure thus underestimates the potential richness of the mutual understandings that can be formed, even in structured tasks like the RCT.

Our findings thus highlight the need to use more naturalistic paradigms to fully understand the influence of local common ground formation on semantic recall memory for conversation. Structured tasks like the RCT may not encourage participants to discuss topics that vary enough in conversational content to allow for rich semantic representations to be built that can later be probed and measured using methods like NLP. As some have recently argued, ecologically valid experiments should be used to develop theories in the first place, rather than be used as an afterthought to validate findings from highly controlled experiments (Hasson et al., 2020; Nastase et al., 2020). Given that the association between common ground formation and memory for conversation has only been tested using highly structured communication tasks like the RCT, more work needs to be done to disentangle this possible relationship in real-world contexts. This is especially true in the current line of work given that the RCT has been proposed to involve different types of cognitive and linguistic skills than a typical social interaction. For example, Bishop and Adams (1991) reported that there was no significant association between the receptive and expressive language skills of children and their performance on a RCT. They concluded by suggesting that extralinguistic skills such as one's ability to visually scan images among highly similar alternatives may have been more important for task performance than true conversational ability (Bishop & Adams, 1991).

Our findings from Experiment 2 are in line with the view that common ground is a multidimensional construct (e.g., E. V. Clark, 2015) that influences conversational behavior in a variety of ways. If the number of words used to describe an image is a meaningful proxy for common ground, our results indicate two distinct main effects on Directors' RCT description lengths that reflect two different forms of common ground. One is the trial effect that has been reported in several previous studies, which represents local common ground formation. Directors in the Friends and Strangers groups used significantly shorter referential labels for facial images over time during the RCT. The other is the group effect, which represents the influence of pre-existing personal common ground on description efficiency. Directors in the Friends group described images with significantly less numbers of words than Directors in the Strangers group. Importantly, there was no significant interaction between these two main effects. Directors in both groups shortened their referential labels at similar rates over the course of repeatedly describing the same images to their partner. These findings suggest that local and personal common ground acted as separate conversational forces during the RCT, which is further supported by the fact that they each had different influences on participants' ability to recall verbatim conversational content about a week later. The relative strength of local common ground formed between dyads during the RCT was a significant predictor of their ability to individually recall verbatim conversational content. In contrast, personal common ground did not provide any memorial benefit for participants in the Friends group. Regression analyses

surprisingly showed that, on average, the image description efficiency presumably afforded by friends' familiarity with one another did not lead to significantly better recall memory performance. Friends exhibited significantly lower verbatim and semantic recall memory for image descriptions than strangers. One possible explanation for this finding is that the RCT was a more salient social experience for participants in the Strangers group than it was for participants in the Friends group (Samp & Humphreys, 2007). While friends have engaged in several previous conversations together and may have conversed multiple times between the RCT and the recall memory task, the RCT was the first-ever interaction between strangers. This could have made the RCT a more memorable experience for strangers. Participants in the Strangers group may also have been more focused, attentive, and/or expended greater effort during the RCT than participants in the Friends group to promote a more positive self-image and/or to avoid negative evaluation from a person they had never previously met.

As previously noted, there are several possible explanations for the descriptive efficiency supposedly afforded by the collection of shared experiences and common knowledge among friends in Experiment 2. One is that friends would have been afforded more opportunity than strangers to make connections between facial images presented to them and information they both knew outside of the experimental context. For example, a Director in the Friends group may have noticed that a facial image had similar features to a character in a movie that they and their partner had previously watched together. Given that this information was already established in their personal common ground, the Director may have opted to refer to that image using the character's name rather than listing a number of the image's physical characteristics. Thus, for some images, it may have been easier for the Director to recall previously grounded information from memory rather than trying to ground entirely new information. The communicative efficiency observed among friends may also have been due to their enhanced ability to read each other's nonverbal cues like emotions and facial expressions. Some evidence suggests that individuals in close interpersonal relationships have a heightened ability to interpret each other's facial expressions as compared to individuals who do not have a close relationship (Altman & Taylor, 1973; Sabatelli et al., 1982; Zhang & Parmley, 2011). In the context of the RCT, facial expressions used by the Matcher may offer useful information for the Director in helping them to determine whether they need to elaborate on their description of an image. When unable to find the image being described by a Director, for example, a Matcher may scrunch their forehead or raise their eyebrows in confusion. Alternatively, when a Matcher has been given enough information to find the correct image in their matrix, they might adopt a subtle smile or nod their head. With less sensitivity to these types of nonverbal cues used by their partner, Directors in the Strangers group may have been led to describe images less efficiently than Directors in the Friends group.

The results from the current set of experiments provide additional evidence that humans do recall some verbatim components of past conversations, even a week after the interaction has passed. For over a century, it was generally thought that individuals could not remember verbatim content from discourse much better than chance (or even at all), except under certain limited circumstances (e.g., Bartlett, 1932; Binet & Henri, 1894; Potter & Lombardi, 1998; Sachs, 1967). For instance, when told in advance that their memory would be tested (Johnson-Laird & Stevenson, 1970), or if asked to recall isolated content that was not part of a coherent discourse (Gernsbacher, 1985). More recent investigations, however, have called this view into question by showing that verbatim recognition and recall for discourse is above chance level. For example, in a series of experiments conducted by Gurevich et al. (2010), subjects were shown to recognize and recall verbatim content from discourse presented in naturalistic contexts (i.e., short stories) much greater than chance. This was despite participants not having been told that their memory would be tested in advance, and the stories being over 300 words in length (Gurevich et al., 2010). Gurevich et al. (2010) concluded by suggesting that while semantic memory outperforms verbatim memory, humans do have the capacity to remember specific words and phrases they encounter during discourse. Our results from the present experiments are in line with this view. In Experiment 2, for example, participants recalled between 11% and 14.5% of verbatim content used by Directors to describe images during the RCT about a week later. This level of cued verbatim recall was observed to be above and beyond the constraints imposed by the images used during the experiment to examine verbatim recall memory.

Finally, our results also add to a mixed set of findings regarding the existence of the generation/production effect (Slamecka & Graf, 1978) in conversational memory. The generation effect proposes that the person who generates language during conversation may have superior memory for that information than a person who was on the receiving end. In both experiments, we found evidence for the generation/production effect, such that Directors were shown to recall significantly more verbatim and semantic content from RCT image descriptions that they themselves produced than Matchers who did not generate the image descriptions. While some studies have reported the same effect (e.g., Knutsen et al., 2016; McKinley et al., 2017; Ross & Sicoly, 1979; Stafford & Daly, 1984), others have found the opposite effect (Stafford et al., 1987) or no significant difference at all (Knutsen & Le Bigot, 2014). One possible reason for these mixed findings may be that different types of dialogue are more or less susceptible to the generation/production effect. The RCT used in the current set of experiments, for example, encourages Directors to produce most of the conversational content. Thus, there is a wide gap in the amount of language processing being carried out by the Director as compared to the Matcher. Compare this to a more naturalistic conversation, where collocutors generate and share more equal amounts of information. In this case, both individuals may benefit more equally from the generation/production effect, leading to smaller differences to be observed in recall memory performance.

Conclusions

The current set of experiments offer additional insight into the influence of common ground on memory for conversation. In two experiments, we showed that the relative strength of local common ground formed between dyads for basic category object images (Experiment 1) and facial images (Experiment 2) during a RCT significantly enhanced their ability to recall verbatim conversational content from the RCT about a week later. These findings serve as an extension of previous work in recognition memory and suggest that local common ground formation is an important predictor of verbatim recall memory for conversation. In providing evidence of this

relationship, the current findings are in support of the view that individuals can remember some specific words and phrases that are used during a conversational interaction (e.g., Gurevich et al., 2010). Our results are also important in highlighting the multidimensionality of common ground, as well as the limitations that are imposed by highly controlled communication tasks on the memory representations that are formed by individuals during the interaction. We showed that local and personal common ground exerted separate influences on conversational behavior and verbatim recall memory, but no significant associations were observed in either experiment between local common ground formation and semantic recall memory performance. These null findings highlight the importance of the need for future research to focus on using more naturalistic communication paradigms in better understanding how conversational mechanisms like common ground formation influence semantic recall memory for conversation.

Context of the Research

The ideas presented in this paper originated from the identification of important gaps in the literature on common ground formation and memory for conversation that could feasibly and reliably be addressed in a series of online studies during the COVID-19 pandemic. Common ground is thought to be a multidimensional force that influences conversation in various ways. To date, however, its multidimensionality has received minimal experimental attention. Moreover, at the time of devising the present research project, the link between common ground and memory for conversation had almost exclusively been tied to recognition memory data. Extending this work into the realm of recall memory was deemed to be an important future direction.

The findings and issues presented in this paper spark several possibilities for future studies to extend this work. For example, researchers may wish to examine how task-relevant personal common ground influences the grounding of new information and memory representations. This could be carried out by adding in a social manipulation prior to the RCT, such as a thin-slicing social judgment task, where participants make collective judgments about some of stimuli. The overarching goal for this line of work is to implement the use of more naturalistic conversational paradigms to uncover the various ways in which common ground influences language and memory processes.

References

Altman, S., & Taylor, D. A. (1973). *Social penetration: The development of interpersonal relationships*. Holt, Rinehart & Winston.

Bamford, J., & Wilson, I. (1979). Methodological considerations and practical aspects of the BKB sentence lists. In J. Bench & J. Bamford (Eds.), *Speech-hearing tests and the spoken language of hearing-impaired children* (pp. 148–187). Academic Press. https://doi.org/10.3109/03005367909078884

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. arXiv preprint. https://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. arXiv preprint. https://arxiv.org/abs/1903.10676

Benoit, P. J., & Benoit, W. L. (1988). Conversational memory employing cued and free recall. *Central States Speech Journal*, *39*(1), 18–27. https://doi.org/10.1080/10510978809363231

Binet, A., & Henri, V. (1894). La mémoire des phrases. *L'annee Psychologique*, *1*(1), 24–59. https://doi.org/10.3406/psy.1894.1045

Bishop, D. V., & Adams, C. (1991). What do referential communication tasks measure? A study of children with specific language impairment. *Applied Psycholinguistics*, *12*(2), 199–215. https://doi.org/10.1017/S0142716400009140

Bock, J. K., & Brewer, W. F. (1974). Reconstructive recall in sentences with alternative surface structures. *Journal of Experimental Psychology*, *103*(5), 837–843. https://doi.org/10.1037/h0037391

Borrie, S. A., Barrett, T. S., & Yoho, S. E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *The Journal of the Acoustical Society of America*, *145*(1), 392–399. https://doi.org/10.1121/1.5087276

Boyle, E. A., Anderson, A. H., & Newlands, A. (1994). The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and Speech*, *37*(1), 1–20. https://doi.org/10.1177/002383099403700101

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology*, *2*(4), 331–350. https://doi.org/10.1016/0010-0285(71)90019-3

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493. https://doi.org/10.1037/0278-7393.22.6.1482

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, *5*(4), 339–360. https://doi.org/10.1037/1076-898X.5.4.339

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286–291. https://doi.org/10.3758/BRM.42.1.286

Clark, E. V. (2015). Common ground. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 328–353). Wiley-Blackwell. https://doi.org/10.1002/9781118346136.ch15

Clark, H. H. (1996). *Using language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620539

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). APA Books. https://doi.org/10.1037/10096-006

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Clark, H. H., & Marshall, C. (1981). Definite reference and mutual knowledge. In A. Joshi, B. Webber, & J. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge University Press.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, *37*(3), 201–242. https://doi.org/10.1006/cogp.1998.0693

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Conneau, A., & Kiela, D. (2018). Senteval: An evaluation toolkit for universal sentence representations. arXiv preprint. https://arxiv.org/abs/1803.05449

De Leeuw, J. R. (2015). Jspsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint. https://arxiv.org/abs/1810.04805

Fisher, J. S., & Radvansky, G. A. (2018). Patterns of forgetting. *Journal of Memory and Language*, *102*, 130–141. https://doi.org/10.1016/j.jml.2018.05.008

Fox Tree, J. E., & Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, *62*(2), 151–167. https://doi.org/10.1016/S0010-0277(96)00781-0

Gangel, J., Liptak, K., Warren, M., & Cohen, M. (2021, February 12). *New details about Trump-McCarthy shouting match show Trump refused to call off the rioters*. *Cable News Network (CNN)*. https://www.cnn.com/2021/02/12/politics/trump-mccarthy-shouting-match-details/index.html

Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology*, *17*(3), 324–363. https://doi.org/10.1016/0010-0285(85)90012-X

Gurevich, O., Johnson, M. A., & Goldberg, A. E. (2010). Incidental verbatim memory for language. *Language and Cognition*, *2*(1), 45–78. https://doi.org/10.1515/langcog.2010.003

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, *105*(3), 416–434. https://doi.org/10.1016/j.neuron.2019.12.002

Hicks, J. L., & Marsh, R. L. (2000). Toward specifying the attentional demands of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1483–1498. https://doi.org/10.1037/0278-7393.26.6.1483

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–142. https://doi.org/10.1016/j.cognition.2004.07.001

Horton, W. S., & Gerrig, R. J. (2016). Revisiting the memory-based processing approach to common ground. *Topics in Cognitive Science*, *8*(4), 780–795. https://doi.org/10.1111/tops.12216

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, *116*(1), 26–37. https://doi.org/10.1037/0096-3445.116.1.26

Johnson-Laird, P. N., & Stevenson, R. (1970). Memory for syntax. *Nature*, *227*(5256), 412. https://doi.org/10.1038/227412a0

Keenan, J. M., MacWhinney, B., & Mayhew, D. (1977). Pragmatics in memory: A study of natural conversation. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 549–560. https://doi.org/10.1016/S0022-5371(77)80018-2

Knutsen, D., & Le Bigot, L. (2014). Capturing egocentric biases in reference reuse during collaborative dialogue. *Psychonomic Bulletin & Review*, *21*(6), 1590–1599. https://doi.org/10.3758/s13423-014-0620-7

Knutsen, D., Ros, C., & Le Bigot, L. (2016). Generating references in naturalistic face-to-face and phone-mediated dialog settings. *Topics in Cognitive Science*, *8*(4), 796–818. https://doi.org/10.1111/tops.12218

Knutsen, D., Ros, C., & Le Bigot, L. (2018). Spoilt for choice: Initially considering several referential expressions affects subsequent referential decisions. *Language, Cognition and Neuroscience*, *33*(5), 618–632. https://doi.org/10.1080/23273798.2017.1400080

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*(5), 113–114. https://doi.org/10.3758/BF03342817

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*(3), 343–346. https://doi.org/10.1037/h0023705

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., & Ferrari, V. (2018). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint. https://arxiv.org/abs/1811.00982. https://doi.org/10.1007/s11263-020-01316-z

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized Bert pretraining approach. arXiv preprint. https://arxiv.org/abs/1907.11692

McKinley, G. L., Brown-Schmidt, S., & Benjamin, A. S. (2017). Memory for conversation and the development of common ground. *Memory & Cognition*, *45*(8), 1281–1294. https://doi.org/10.3758/s13421-017-0730-3

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint. https://arxiv.org/abs/1301.3781

Miller, J. B., & de Winstanley, P. A. (2002). The role of interpersonal competence in memory for conversation. *Personality and Social Psychology Bulletin*, *28*(1), 78–89. https://doi.org/10.1177/0146167202281007

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, Article 117254. https://doi.org/10.1016/j.neuroimage.2020.117254

Neisser, U. (1981). John Dean's memory: A case study. *Cognition*, *9*(1), 1–22. https://doi.org/10.1016/0010-0277(81)90011-1

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543). https://doi.org/10.3115/v1/D14-1162

Pollmann, M. M. H., & Krahmer, E. J. (2018). How do friends and strangers play the game taboo? A study of accuracy, efficiency, motivation, and the use of shared knowledge. *Journal of Language and Social Psychology*, *37*(4), 497–517. https://doi.org/10.1177/0261927X17736084

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*(6), 633–654. https://doi.org/10.1016/0749-596X(90)90042-X

Potter, M. C., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language*, *38*(3), 265–282. https://doi.org/10.1006/jmla.1997.2546

Putnam, A. L., & Roediger, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition*, *41*(1), 36–48. https://doi.org/10.3758/s13421-012-0245-x

Radvansky, G. A., Doolen, A. C., Pettijohn, K. A., & Ritchey, M. (2022). A new look at memory retention and forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(11), 1698–1723. https://doi.org/10.1037/xlm0001110

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

Redeker, G. (1984). On differences between spoken and written language. *Discourse Processes*, *7*(1), 43–55. https://doi.org/10.1080/01638538409544580

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using Siamese Bert-networks. arXiv preprint. https://arxiv.org/abs/1908.10084

Repp, B. H. (1976). Identification of dichotic fusions. *The Journal of the Acoustical Society of America*, *60*(2), 456–469. https://doi.org/10.1121/1.381103

Rodrigues, M. A., Yoon, S. O., Clancy, K. B., & Stine-Morrow, E. A. (2021). What are friends for? The impact of friendship on communicative

efficiency and cortisol response during collaborative problem solving among younger and older women. *Journal of Women & Aging*, *33*(4), 411–427. https://doi.org/10.1080/08952841.2021.1915686

Rosen, S., & Corcoran, T. (1982). A video-recorded test of lipreading for British English. *British Journal of Audiology*, *16*(4), 245–254. https://doi.org/10.3109/03005368209081469

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology*, *37*(3), 322–336. https://doi.org/10.1037/0022-3514.37.3.322

Sabatelli, R. M., Buck, R., & Dreyer, A. (1982). Nonverbal communication accuracy in married couples: Relationship with marital complaints. *Journal of Personality and Social Psychology*, *43*(5), 1088–1097. https://doi.org/10.1037/0022-3514.43.5.1088

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, *2*(9), 437–442. https://doi.org/10.3758/BF03208784

Samp, J. A., & Humphreys, L. R. (2007). "I said what?" Partner familiarity, resistance, and the accuracy of conversational recall. *Communication Monographs*, *74*(4), 561–581. https://doi.org/10.1080/03637750701716610

Schober, M. F., & Carstensen, L. L. (2009). Does being together for years help comprehension. In E. Morsella (Ed.), *Expressing oneself/expressing one's self: Communication, cognition, language, and identity* (pp. 107–124). Lawrence Erlbaum.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232. https://doi.org/10.1016/0010-0285(89)90008-X

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Stafford, L., Burggraf, C. S., & Sharkey, W. F. (1987). Conversational memory: The effects of time, recall, mode, and memory expectancies on remembrances of natural conversations. *Human Communication Research*, *14*(2), 203–229. https://doi.org/10.1111/j.1468-2958.1987.tb00127.x

Stafford, L., & Daly, J. A. (1984). Conversational memory: The effects of recall mode and memory expectancies on remembrances of natural conversations. *Human Communication Research*, *10*(3), 379–402. https://doi.org/10.1111/j.1468-2958.1984.tb00024.x

Standing, L. (1973). Learning 10,000 pictures. *The Quarterly Journal of Experimental Psychology*, *25*(2), 207–222. https://doi.org/10.1080/14640747308400340

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology*, *9*(1), 77–110. https://doi.org/10.1016/0010-0285(77)90005-6

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–402). Academic Press.

Van der Wege, M. M. (2009). Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, *60*(4), 448–463. https://doi.org/10.1016/j.jml.2008.12.003

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008). https://arxiv.org/abs/1706.03762

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*(2), 183–194. https://doi.org/10.1016/0749-596X(92)90010-U

Wilson, S. M., Bautista, A., & McCarron, A. (2018). Convergence of spoken and written language processing in the superior temporal sulcus. *Neuroimage*, *171*, 62–74. https://doi.org/10.1016/j.neuroimage.2017.12.068

Yang, X., He, X., Zhang, H., Ma, Y., Bian, J., & Wu, Y. (2020). Measurement of semantic textual similarity in clinical texts: Comparison of transformer-based models. *JMIR Medical Informatics*, *8*(11), Article e19735. https://doi.org/10.2196/19735

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

Zhang, F., & Parmley, M. (2011). What your best friend sees that I don't see: Comparing female close friends and casual acquaintances on the perception of emotional facial expressions of varying intensities. *Personality and Social Psychology Bulletin*, *37*(1), 28–39. https://doi.org/10.1177/0146167210388194

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE International Conference on Computer Vision (pp. 19–27). https://doi.org/10.1109/ICCV.2015.11

Received January 8, 2022
Revision received October 21, 2022
Accepted November 9, 2022 ■