INVESTIGATING THE INFLUENCE OF LOCAL AND PERSONAL COMMON GROUND ON MEMORY FOR CONVERSATION USING AN ONLINE REFERENTIAL COMMUNICATION TASK

by

Daniel Robert Nault

A thesis submitted to the Graduate Program in Psychology

in conformity with the requirements for

the degree of Master of Science

Queen's University

Kingston, Ontario, Canada

(September 2021)

Abstract

A shared understanding or common ground is known to be an essential aspect of efficient conversation. Interlocutors require some level of mutual knowledge to build their conversation together without restating redundant information. Recently, common ground has also been shown to be related to recognition memory for conversation (McKinley, Brown-Schmidt, & Benjamin, 2017). Here, we investigated the influence of two forms of common ground between conversational dyads on their ability to recall verbatim and semantic conversational content about a week later. Semantic recall memory was measured using a natural language processing approach. In Experiment 1, we examined whether the strength of local common ground formed between dyads for images during an online referential communication task (RCT) predicted their ability to recall image descriptions used during the RCT. In Experiment 2, we varied the level of pre-existing personal common ground between dyads participating in the RCT and examined their recall memory performance. We did this by recruiting pairs of friends and strangers. In both experiments, there was a significant association between the strength of local common ground formed between dyads for images during the RCT and their verbatim (but not semantic) recall memory for image descriptions. These findings provide additional evidence that individuals can remember some verbatim words and phrases used during conversations, and partially support the view that common ground and memory are intricately linked. However, the null findings with regards 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. Participants who generated the image descriptions during the RCTs also tended to show superior recall memory performance. In Experiment 2, friends used significantly less numbers of words to describe images during the RCT than strangers, providing evidence that conversational

efficiency was afforded by their pre-existing personal common ground. However, contrary to our hypothesis, results suggest that strangers exhibited better verbatim and semantic recall memory performance than friends. These findings are discussed in relation to the multidimensional nature of common ground and the importance of more natural conversational tasks.

Keywords: common ground, conversation, memory

Acknowledgements

This project would not have been possible without the unwavering support and expertise of my supervisor and mentor, Dr. Kevin Munhall. Through all the extenuating circumstances of the COVID-19 pandemic and the many ups and downs that were experienced transitioning this work online, Dr. Munhall was there to coach and support me. I will be forever grateful for everything he has done for me. That includes the countless edits he made to this paper over the past several months. I also want to thank my committee members, Dr. Jeffrey Wammes and Dr. Valerie Kuhlmeier, for their support, guidance, and advice throughout the development of my thesis. I am also grateful for Matthew Nicastro, who played a major role in getting the experiments presented in this thesis up and running online. This project would not have been possible without his hard work and dedication during what was a very uncertain time for everyone. I also want to thank Rohit Voleti for his contributions to this project, particularly in helping me to explore and apply a natural language processing approach to my thesis. Natalie Hanna and Hope Mitchell also did a lot of work behind the scenes to make this project a success. I want to thank them for their hard work on this project, and for the many hours they put it in over the past year. I also want to thank one of my biggest supporters, Yifei Yin. I would not have produced this piece of work without his endless positivity and support. Finally, I will always be grateful for the love and compassion provided by my family. I owe it all to them.

Table of Contents

Abstract……...ii

Acknowledgements…….iv

Table of Contents……v

List of Tables…….....vii

List of Figures……...viii

List of Abbreviations……...……...ix

Chapter 1: Introduction…….....1

The Current Research……..8

Chapter 2: Experiment 1…….....9

Method……...……11

Results……19

Discussion……..27

Chapter 3: Experiment 2 ……...30

Method……...31

Control Experiment……33

Results……36

Discussion……..……44

Chapter 4: General Discussion……...……48

Conclusions……57

References……..59

Appendix A: List of Referential Communication Task (RCT) Stimuli…….68

Appendix B: List of Foil Stimuli for the Recognition Memory Task……...…….72

Appendix C: Descriptive Statistics for the Referential Communication Task (RCT)……...76

Appendix D: Descriptive Statistics for the Recognition Memory Task……78

Appendix E: Descriptive Statistics for the Recall Memory Task……..……79

Appendix F: Ethics Clearance Letter……..…80

List of Tables

Table 1. Coefficients for the Best Fit Linear Mixed-Effects Model Predicting Directors' Description Lengths During the RCT in Experiment 1…….22

Table 2. Coefficients for the Best Fit Linear Mixed-Effects Model Predicting Word-For-Word Recall Memory Performance in Experiment 1……..24

Table 3. Coefficients for the Best Fit Linear Mixed-Effects Model Predicting Semantic Recall Memory Performance in Experiment 1……...……...……26

Table 4. Coefficients for the Best Fit Linear Mixed-Effects Model Predicting Directors' Description Lengths During the RCT in Experiment 2…….39

Table 5. Coefficients for the Best Fit Linear Mixed-Effects Model Predicting Word-for-Word Recall Memory Performance in Experiment 2……..42

Table 6. Coefficients for the Best Fit Linear Mixed-Effects Model Predicting Semantic Recall Memory Performance in Experiment 2……..……43

Table 7. Descriptive Statistics from the Control Distributions vs. Descriptive Statistics from the Verbatim Recall Exhibited by Participants in Experiment 2……..…44

Table S1. Descriptive Statistics for the RCT in Experiment 1……..76

Table S2. Descriptive Statistics for the RCT in Experiment 2……..77

Table S3. Descriptive Statistics for the Recognition Memory Task in Experiment 1……...78

Table S4. Descriptive Statistics for the Recognition Memory Task in Experiment 2……...……78

Table S5. Descriptive Statistics for Recall Memory Performance in Experiment 1…….79

Table S6. Descriptive Statistics for Recall Memory Performance in Experiment 1…….79

List of Figures

Figure 1. Computer Screen Setup for the RCT via Zoom…….……14

Figure 2. Average Number of Words Used by Directors in Trials 1, 2, and 3 to Describe Images in the RCT in Experiment 1……...21

Figure 3. Best Fit Linear-Mixed Effects Model Predicting Word-For-Word Recall Memory Performance in Experiment 1……25

Figure 4. Best Fit Linear Mixed-Effects Model Predicting Semantic Recall Memory Performance in Experiment 1…….…….…….…….……..…27

Figure 5. Computer Screen Setup for the Control Experiment……..35

Figure 6. Average Number of Words Used by Directors in the Friends and Strangers Groups to Describe Images in Trials 1, 2, and 3 in the RCT in Experiment 2……...38

List of Abbreviations

RCT = referential communication task

NLP = natural language processing

Chapter 1

Introduction

Conversation is one of the most ubiquitous human activities. It is our main way of sharing stories, building and maintaining relationships, helping others, and passing on knowledge to future generations. Crucial to the success of any conversation is the ability for interlocutors to encode and form memory representations for information that is shared, and access memories formed in previous interactions (e.g., Horton & Gerrig, 2005, 2016). Speakers rely on their memories to plan and tailor their utterances to the shared knowledge they have accumulated with their partner over repeated exchanges (Clark & Marshall, 1981; Clark, 1996). Conversational topics are also 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. The aim of the current set of experiments was to examine whether the process by which conversational memories are formed influences the degree to which their representations can later be accessed.

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, Liptak, Warren, & Cohen, 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 classic 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 subjects 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 they had heard 80 syllables beyond the sentence being tested. Hence, while the meaning of sentences was retained quite well in memory over time, memory for the original surface form of the sentences was forgotten very quickly (Sachs, 1967). In another study, Bransford and Franks (1971) presented subjects with simple sentences that contained either one, two, or three semantically related ideas. In a recognition memory task, subjects were then presented with the same sentences, along with several new complex ones that contained different combinations of all of 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 the 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 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 are able to 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). A number of factors have also been shown to influence the amount of verbatim information that individuals can recall from discourse. For example, Keenan, MacWhinney, and Mayhew (1977) examined individuals' ability to recognize verbatim statements from a lunchroom discussion 30 hours after the discussion had ended. The results showed that individuals were three times more likely to accurately recognize verbatim statements when they contained information about a speakers' beliefs, intentions, and attitudes towards the listener (i.e., high interactional content) as opposed to when the information contained emotionally neutral information (i.e., low interactional

content; 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 conversational partners 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 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 (Clark, 2015). Thus, common ground is an essential aspect of conversation that allows interlocutors to build their communication together indefinitely without restating redundant information.

In the conversation literature, 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, 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 orientation in their matrix. Over the course of six trials of matching with the same images, 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., Wilkes-Gibbes & Clark, 1992; Brennan & Clark, 1996; Van der Wege, 2009; McKinley, Brown-Schmidt, & Benjamin, 2017) and is taken as evidence for the formation of common ground in conversation.

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 and colleagues (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. Specifically, for every reduction in one word used by Directors to describe an image in the matching task, 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 are discussed in light of the generation effect in memory (see Slamecka & Graf, 1978) and 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 (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. It also begs the question of whether the memory representations formed by interlocutors during single interactions with different degrees of previously existing common ground are structurally different.

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, Anderson, and Newlands (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 Cartenson (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. While these mixed findings could be due to differences in the methodologies used to evoke conversation, further research is needed to better understand how shared knowledge influences the grounding of new information, as well as the impact 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 Current Research

The aim for this paper was to provide a comprehensive investigation of the association between different types of common ground and memory for conversation. This was carried out by conducting two separate experiments using an online version of the RCT (Krauss & Weinheimer, 1964, 1966) that we adapted for use during the COVID-19 pandemic. In Experiment 1, we directly examined whether the strength of local common ground formed between dyads for basic category object images during a RCT could predict their ability to individually recall verbatim and semantic conversational content (i.e., image descriptions) from the RCT about a week later. This work was primarily motivated by a recent study by McKinley et al. (2017), that showed that image recognition memory was significantly enhanced when interlocutors formed stronger common ground for the same images with their partner during a RCT.

In Experiment 2, we varied the level of pre-existing personal common ground between dyads participating in the RCT. This was achieved by recruiting two groups of dyads–friends who had known each other for at least six months prior to participating in the experiment and strangers who had never previously met. Our aim was to investigate whether the collection of shared knowledge and previous experiences among pairs of friends would facilitate their ability to form local common ground for referential labels of facial images during a RCT and later recall them as compared to strangers without personal common ground. While 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., Isaacs & Clark, 1987; Boyle et al., 1994), little is known about what influence this efficiency may have on the memory representations that are formed during conversation and that can later be retrieved (e.g., Samp & Humphreys, 2007).

Our prediction was that local common ground formation during the RCT and previously existing personal common ground among friends would lead to stronger memory for conversational content at follow-up.

In both experiments, we opted to include a measure of both verbatim and semantic recall memory performance. As previously noted, humans are known to have a limited capacity to recall verbatim words and phrases that they encounter during connected discourse (Sachs, 1967; Bransford & Franks, 1971; Stafford & Daly, 1984; Gernsbacher, 1985; Potter & Lombardi, 1990). We directly addressed the possibility that participants would not remember verbatim information from the RCT very well by 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.

Chapter 2

Experiment 1

In Experiment 1, we sought to examine the relationship between common ground formation and memory using an online RCT (Krauss & Weinheimer, 1964, 1966). Based on consistent findings from several previous RCT experiments (e.g., Krauss & Weinheimer, 1964, 1966; Wilkes-Gibbes, & Clark, 1992; McKinley et al., 2017), we expected that conversational dyads, albeit in a virtual format, would develop shared referential labels to describe basic category object images. We further predicted that dyads would use shorter referential labels over time, thus providing evidence of common ground formation (Wilkes-Gibbes, & Clark, 1992; McKinley et al., 2017).

McKinley et al. (2017) showed that participants' image recognition memory was enhanced when they formed stronger levels of common ground for the same images during a

RCT. We extended this work by testing whether the level of common ground formed between dyads in a RCT could predict their verbatim and semantic recall memory for image descriptions in a follow-up recall memory task. Our prediction was that stronger common ground formation would lead to superior word-for-word and semantic recall memory performance among participants about a week later. Our design also included a recognition memory task that was presented immediately following the RCT. We used this task as a manipulation check to ensure participants were paying attention and forming memories for images in an online environment.

As noted above, previous research has indicated that humans have a limited capacity to recognize and recall verbatim conversational content, even after short time delays (e.g., Kintsch & Bates, 1977; Ross & Sicoly, 1979; Stafford & Daly, 1984). This is thought to be due to linguistic material being encoded and represented in memory in terms of its overall meaning or "gist" rather than its original surface structure (e.g., Sachs, 1967). With this in mind, we opted to include a measure of semantic recall memory. This type of 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).

To test verbatim recall, we used word accuracy methods similar to approaches used in speech intelligibility research (e.g., Bamford & Wilson, 1979; Rosen & Corcoran, 1982). For the more semantic aspects of memory, we used NLP methods. 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.

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, Chen, Corrado, & Dean, 2013) and GloVe (Pennington, Socher, & Manning, 2014), have gained widespread usage due to their increased performance when compared with human evaluations of several standard textual data sets (e.g., 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, Chang, Lee, & Toutanova, 2018) have shown breakthrough performance on a variety of NLP tasks, including textual similarity. In the current set of experiments, we used the NLP model RoBERTa (Liu et al., 2019), an iterative improvement of the BERT model, which has displayed very strong performance on a variety of semantic similarity tasks (Yang et al., 2020).

Method

Participants

Fifteen pairs of participants (30 participants) were recruited from a university Facebook group. Three pairs (six participants) were excluded from the study, one due to an internet connection issue, and the other two due to data saving issues. The remaining 24 participants (16 females) ranged in age from 20-30 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; See Appendix A for the full list of stimuli). Images within the same basic object category were selected in an attempt to ensure that no single image was more distinctive than another, and each image in the dataset was cropped to be equal in size and resolution (200x200 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 (8 additional images in each of the 4 different basic object categories) were only shown to participants during the forced-choice recognition memory task (see Appendix B for the full list of foil stimuli). 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

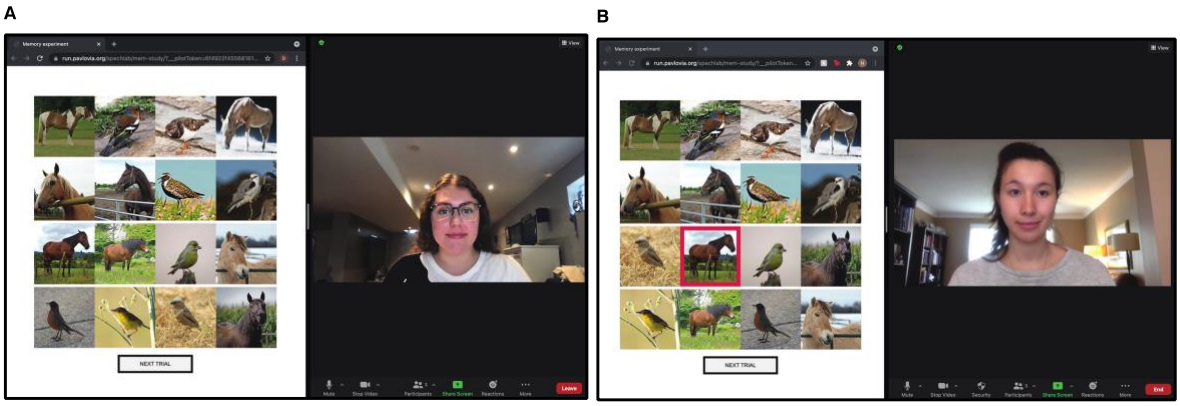
The RCT and the recognition memory task were both hosted by Pavlovia (Peirce et al., 2019), a website that can be used to host and run experiments online. The RCT was programmed using jsPsych (De Leeuw, 2015), a JavaScript-based web browser experiment builder. The recognition memory task was built using 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 (Zoom Video Communications Inc., 2016) 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 4x4 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 (A) describes each image from top to bottom and from left to right to the Matcher (B) who clicks on the respective image in their matrix and swaps it in the correct position to match their partner.

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 4x4 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, each round consisting 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 4x4 matrix again. Upon completion of the RCT, pairs of participants remained on Zoom with the experimenter and were provided with instructions for the forced-choice 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 and 32 'new' that had not been seen in the experiment). 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 seven 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, Burggraf, & Sharkey, 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., Yoon & Brown-Schmidt, 2014; McKinley et al., 2017), 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 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"). 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:

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

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., Clark & Wilkes-Gibbes, 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 (i.e., Directors' overall accuracy in describing images). Thus, a difference score measure of common ground formation is determined in part by overall levels of verbosity.

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, Barrett, & Yoho, 2019). It is common in recall memory for conversation and speech intelligibility studies to use 'loose' criteria when determining correct responses (e.g., Benoit & Benoit, 1988; Thorndyke, 1977; Bamford & Wilson, 1979; Rosen & Corcoran, 1982). 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).

A pre-trained RoBERTa model (Liu et al., 2019) was used to obtain a measure of semantic recall memory performance. This was carried out by computing the level of semantic (i.e., cosine) similarity between Directors' image descriptions during the final trial of the RCT and the Directors' and Matchers' recall memory for those same descriptions. 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 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 ($\theta = 0^{\text{o}}$) and a value of 0 for perpendicularly oriented vectors ($\theta = 90^{\text{o}}$). 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 involving linear mixed-effects modelling. All models were implemented using the lme4 package (v1.1-27; Bates, Mächler, Bolker, & Walker, 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 specified based on the set of rules proposed by Barr, Levy, Scheepers, and Tily (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, Vasishth, & Baayen, 2015). This involved first testing a model that included the optimal 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). The Best Fit Model for each analysis always outperformed all other models and satisfied convergence criteria.

Best Fit Models were established for each experiment to predict: (1) description lengths in number of words used by Directors to describe images in the RCT, (2) word-for-word recall memory and (3) semantic recall memory for the referential labels used by Directors to describe images in the last trial of the RCT.

Results

The primary dataset for Experiment 1 consists of 1152 trials wherein the Director was tasked with describing a target image to the Matcher (24 Directors * 3 trials * 16 images per trial

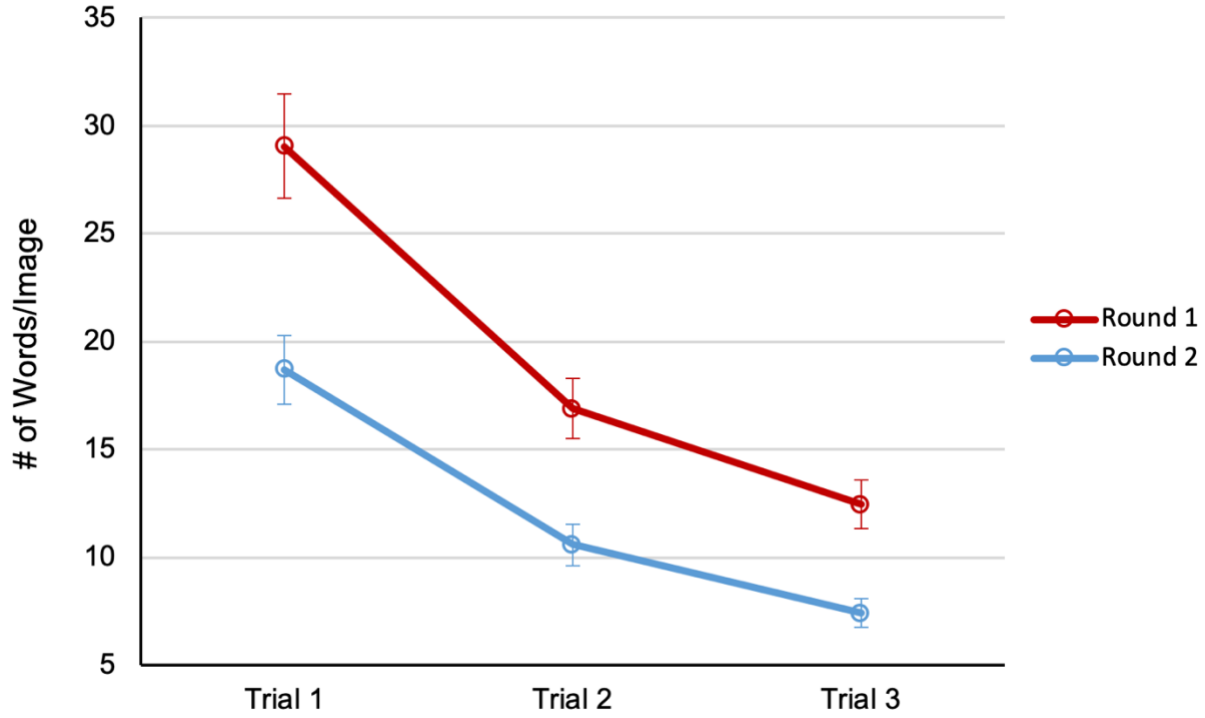
= 1152). On several of those trials, Matchers communicated to the Director whether or not they understood which image was being described. These Matcher utterances were not analyzed here. A total of 12 Director trials were eliminated as outliers (i.e., the number of words used by the Director to describe an image in these trials was greater than 3.29 standard deviations above the overall mean number of words used by Directors).

Description Lengths

As shown in Figure 2, 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 Appendix C.

Figure 2

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



Note: Round 1 and round 2 are shown in red and blue, respectively. Error bars represent 95% confidence intervals.

The Best Fit Model predicting Directors' description lengths in the RCT produced the best fit to the data and included a maximal random-effects structure with random intercepts and correlated slopes for participants and images. It also included fixed effects of trial, round, and 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 maximal random-effects structure, $\chi^2(5) = 49.48$, $p < .001$. The Best Fit Model performed significantly better than alternative models that only included the fixed effect of trial ($\chi^2(2) = 10.88$, $p = 0.012$) or round, $\chi^2(4) = 40.92$ $p < .001$.

The Best Fit Model revealed 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 effect of round. Directors used longer labels to describe images in round 1 than in round 2 of the RCT. The interaction between trial and round was not significant. The results from the Best Fit Model predicting Directors' description lengths from the RCT in Experiment 1 are shown in Table 1.

Table 1

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

Fixed effects	**Estimate**	**SE**	***t*-value**	***p*-value**	**Random Effects**	**Variance**	***SD***
Intercept	29.20	2.46	11.87		Groups		
Trial 2	-12.20	1.95	**-6.26**	***p* < .001**	*Image*		
Trial 3	-16.68	2.16	**-7.73**	***p* < .001**	Intercept	21.69	4.66
Round	-10.41	3.48	**-2.99**	***p* = 0.005**	Trial 2	2.41	1.55
Trial 2*Round	4.04	2.76	1.47	$p = 0.156$	Trial 3	6.30	2.51
Trial 3*Round	5.36	3.05	1.76	$p = 0.091$	Participant		
					Intercept	51.72	7.19
					Trial 2	34.42	5.87
					Trial 3	41.76	6.46
					Residual	74.00	8.60

Note: Number of observations = 1140; number of images = 32; number of participants = 24.

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 Appendix D.

Recall Memory

Two mixed-models (one for word-for-word proportion correct scores based on Autoscore, one for semantic similarity using RoBERTa) 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-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 word-for-word similarity produced the best fit to the data and included a maximal random-effects structure with random intercepts and correlated slopes for participants and images. It also included fixed effects of relative strength of common ground formation and role, and 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, $\chi^2(5) = 46.39$, $p < .001$. The Best Fit Model was also a significantly better fit to the data than an alternative model that did not include the fixed effect of role, $\chi^2(2) = 7.87$, $p = .0195$. The Best Fit Model significantly outperformed another alternative model that did not include the fixed effect of relative strength of common ground formation, $\chi^2(2) = 26.88$, $p < .001$.

As shown in Figure 3, there was a significant effect of relative strength of common ground formation. This effect indicates that word-for-word recall memory performance was enhanced when stronger levels of common ground formation for images were achieved between dyads during the RCT. The effect of role was also significant. On average, Directors recalled 3.9% more of the same words they themselves used to describe images during the RCT as compared to Matchers. Best Fit Model coefficients for word-for-word recall memory performance are presented in Table 2.

Table 2

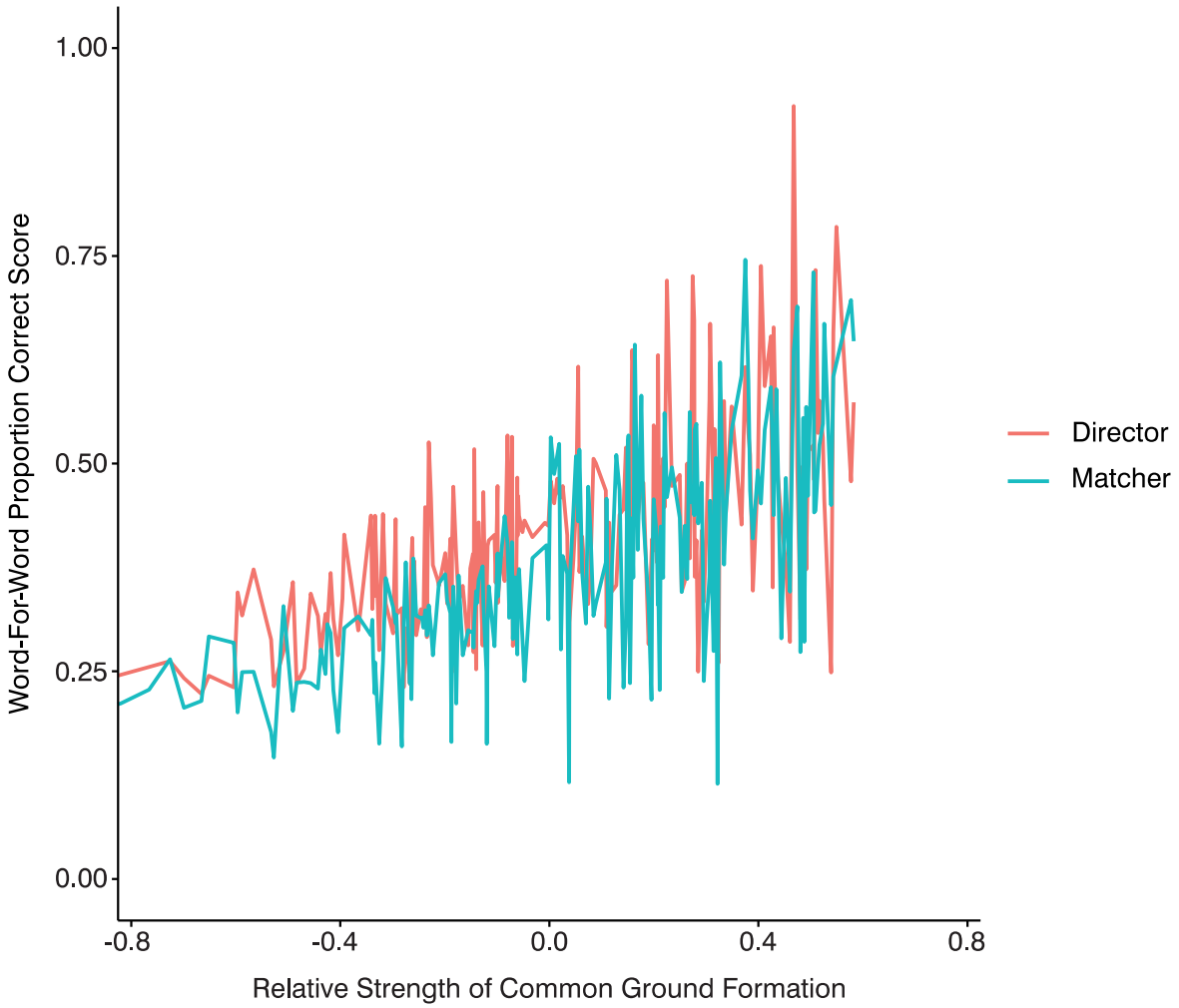
Coefficients for the Best Fit Linear Mixed-Effects Model Predicting Word-For-Word Recall Memory Performance in Experiment 1

	Estimate	**SE**	***t*-value**	***p*-value**	**Random Effects**	**Variance**	***SD***
Intercept	0.411	0.025	16.76		Groups		
Common Ground	0.211	0.042	**5.08**	***p* < .001**	*Image*		
Role	-0.045	0.016	**-2.81**	***p* = 0.005**	Intercept	0.0049	0.070
Common Ground*Role	-0.0025	0.049	-0.051	$p = 0.960$	Common Ground	0.0072	0.085
					Participant		
					Intercept	0.0074	0.086
					Common Ground	0.0052	0.072
					Residual	0.045	0.212

Note: Number of observations = 712; number of images = 32; number of participants = 23.

Figure 3

Best Fit Linear-Mixed Effects Model Predicting Word-For-Word Recall Memory Performance in Experiment 1



Linear mixed-effects modelling for semantic recall memory performance revealed different results. The Best Fit Model again included a maximal random-effects structure with random intercepts and correlated 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 effect of relative strength of common ground formation and the interaction between relative strength of common ground formation and role did not significantly outperform the Best Fit Model, $\chi^2(2) = .66$, $p = .719$. The Best Fit Model yielded a significant effect of role. On average, Directors' recall memories had 6.5% 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 3 and a visual depiction of the model is shown in Figure 4. Descriptive statistics for word-for-word and semantic recall memory performance are presented Appendix E.

Table 3

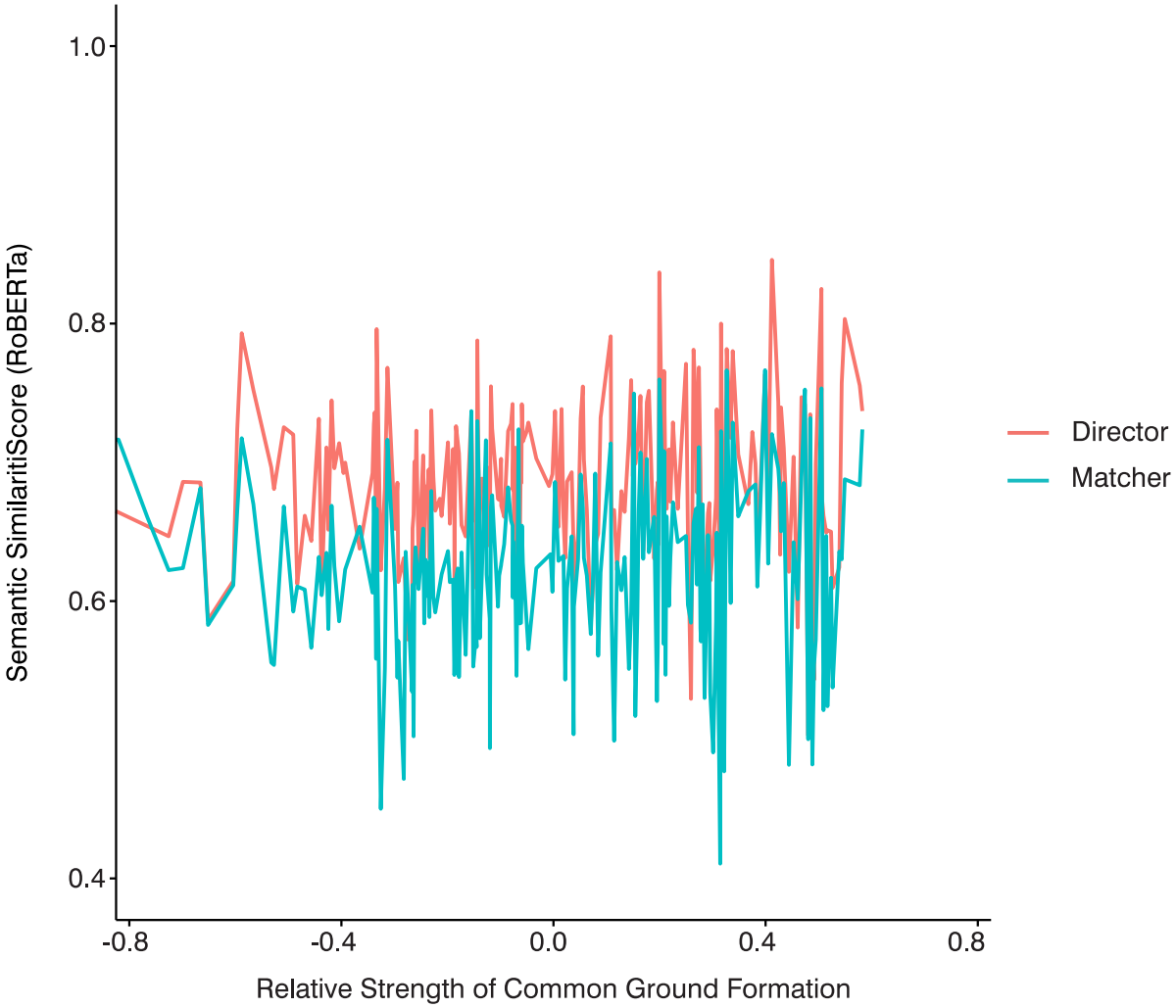
Coefficients for the Best Fit Linear Mixed-Effects Model Predicting Semantic Recall Memory Performance in Experiment 1

Fixed effects	**Estimate**	**SE**	**t-value**	**p-value**	**Random Effects**	**Variance**	**SD**
Intercept	0.682	0.017	40.04		*Image*		
Role	-0.065	0.013	**-4.80**	***p* < .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.

Figure 4

Best Fit Linear Mixed-Effects Model Predicting Semantic Recall Memory Performance in Experiment 1



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., Krauss & Weinheimer, 1964, 1966; Clark & Wilkes-Gibbes, 1986; Schober & Clark, 1989; Brennan & Clark, 1996; McKinley et al., 2017). 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 partner, task-specific learning of the RCT from being the Matcher in round one, and stimulus set differences between the two rounds.

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. 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 in the present experiment. One possible explanation for this

null finding is that RCTs are highly structured and place major limitations on the ability for 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 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 also enhanced the level of control over the amount of time elapsed between participants' participation in the RCT and their involvement in the recall memory task. This was carried out to obtain a more accurate estimate of the potential effects of common ground formation and conversational

role on recall memory for conversation. We return to the some of the issues raised here in greater detail in our General Discussion.

Chapter 3

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 determine 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 and later remember image descriptions during the recall memory task as compared to pairs of strangers who did not have previous shared experiences to draw upon. A few studies have directly examined the influence of friendship on local common ground formation (e.g., Boyle et al., 1994), but only one study that we know of has directly examined its influence on memory for conversation (Samp & Humphreys, 2007).

For Experiment 2, we used facial images from the open-source Glasgow Unfamiliar Face Database (GUFD; Burton, White, & McNeill, 2010) rather than images from basic object categories. This decision was 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 and recall memory.

Method

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

Participants

Twenty-six pairs of participants (52 participants) that did not participate in Experiment 1 were recruited from a local Facebook group. Two pairs of participants (four participants) were excluded because of technical issues with the Zoom audio recording. The remaining 48 participants (40 females) ranged in age from 18-28. Twenty-four of them (12 pairs; 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; 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 Glasgow Unfamiliar Face Database (GUFD; Burton et al., 2010). 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 303 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 (see Appendix A for the full list of stimuli) from the GUFD. 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 forced-choice 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 (see Appendix B for the full list of foil

stimuli). All images were cropped to be equal in size and resolution (200x200 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

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 possible 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 independent of recall memory processes. We conducted a Control Experiment to

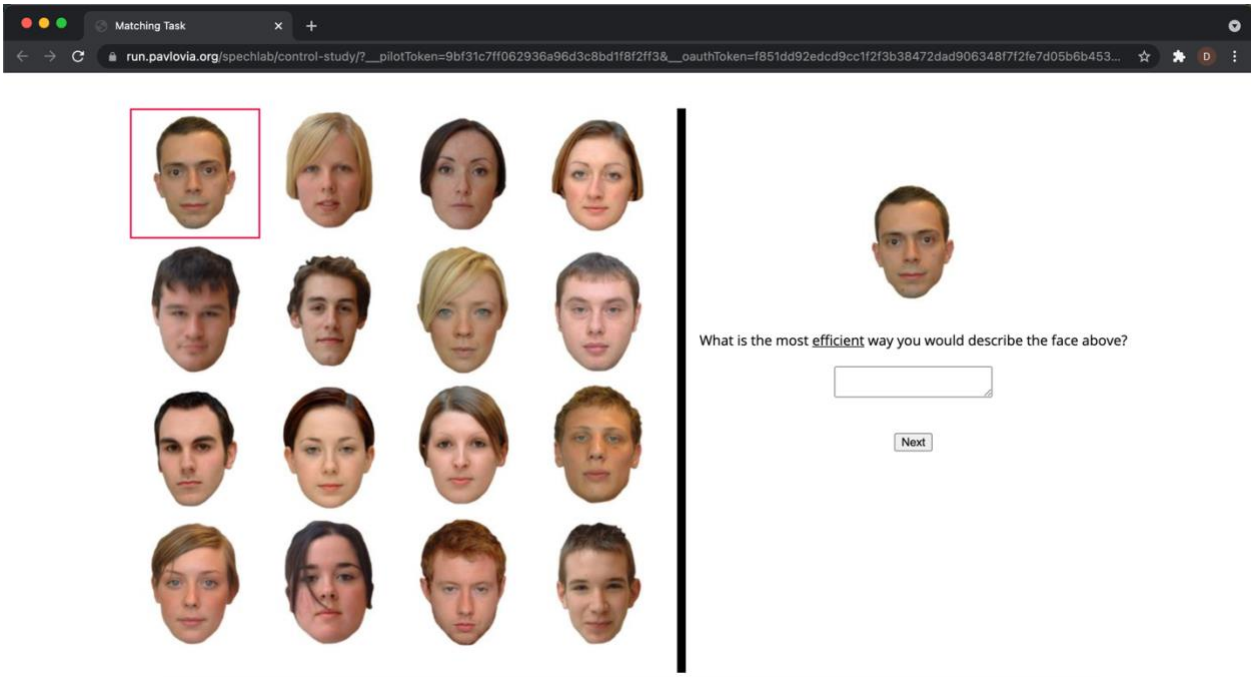
directly address this issue. Our aim was to determine whether there was evidence of true verbatim recall memory in Experiment 2 rather than just a similarity caused by the pictures evoking a description that was similar.

Twelve fluent English-speaking participants (9 females; M_{age} = 25.08, SD_{age} = 3.75) who did not participate in Experiment 1 or Experiment 2 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 facial images that were presented to participants during the RCT and recall memory task in Experiment 2. 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 the two previous experiments, 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 5. On one half of their computer screen, 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 4x4 matrix of images that were presented to dyads during round 1 and round 2 the RCT in Experiment 2. 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.

Figure 5

Computer Screen Setup for the Control Experiment



Note: Each image that participants were asked to describe (on the right) was pointed out to them with a red box surrounding it in the 4x4 matrix of images (on the left). Shown here are the 16 facial images from round 1 of the RCT in Experiment 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 from the Friends and Strangers groups in Experiment 2. The idea is that the descriptions produced by the naïve control subjects act as an independent standard to compare the recalled memories against. The control descriptions have in common 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 descriptions are not shaped by a communicative process and thus, there is nothing unique to a particular conversation in the description. If the

recall memory contains traces from the original conversation rather than simply being an evoked picture description, the verbatim similarity scores from Experiment 2 should be higher than the verbatim similarity scores derived from comparing control descriptions to recall memory descriptions.

Due to there being no inherent match of a control description to a particular participant's recall, we created four separate distributions of similarity scores between the control data and the recall statements (i.e., Control vs. Directors in Strangers group; Control vs. Matchers in Strangers group; Control vs. Directors in Friends group; Control vs. Matchers in Friends group). 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 X 384 recall descriptions = 4,608 for the Stranger 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 2. This allows for a statement to be made about the probability that factors beyond image description influenced the verbatim recall among participants in Experiment 2.

Results

Directors described target images to Matchers on 2304 trials (48 Directors * 3 trials * 16 images per trial = 2304). Half of the images were described by Directors in the Friends group,

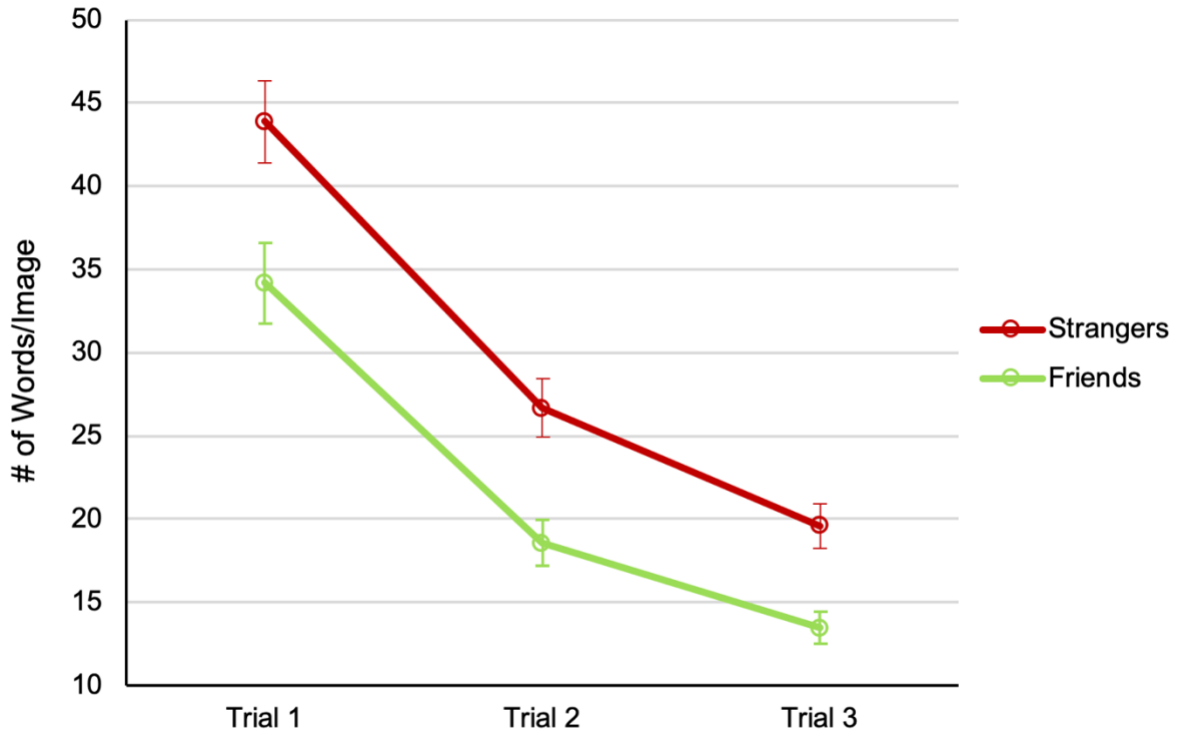
while the other half were described by Directors in the Strangers group. A total of 47 Director trials were eliminated from the dataset, 13 of which were outliers (i.e., image descriptions longer than 3.29 standard deviations above the overall mean). An additional 19 trials were eliminated due to technical issues with Zoom audio recordings and 15 trials were eliminated due to Directors using only gestures to describe a facial image to their partner (e.g., using hand gestures to specify a distinctive type of hairstyle). 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 6). 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 Appendix C.

Figure 6

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



Note: Data have been averaged across both rounds. Error bars indicate 95% confidence intervals.

The maximal random-effects structure for the Best Fit Model used to predict Directors' description lengths in the RCT had random intercepts and correlated slopes for participants and images. The Best Fit Model also included fixed effects of group (Friends/Strangers), trial, and their interaction term. 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, $\chi^2(5) = 87.17$, $p < .001$. Including the fixed effect of round in an alternative model did not lead to a significantly better model fit, $\chi^2(6) = 3.34$, $p = 0.765$. 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 did not include the fixed effect of group ($\chi^2(3) = 9.27$, $p = .026$) or trial number, $\chi^2(3) = 79.95$, $p < .001$.

The results from the Best Fit Model revealed 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 9.25 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 4.

Table 4.

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

Fixed effects	**Estimate**	**SE**	**t-value**	**p-value**	**Random Effects**	**Variance**	**SD**
Intercept	34.67	2.58	13.43		Groups		
Group	9.25	3.19	**2.90**	***p* = 0.006**	*Image*		
Trial 2	-15.82	1.81	**-8.76**	***p* < .001**	Intercept	48.99	7.00
Trial 3	-21.06	1.97	**-10.67**	***p* < .001**	Trial 2	25.20	5.02
Group*Trial 2	-1.14	2.20	-0.52	*p* = 0.608	Trial 3	29.33	5.42
Group*Trial 3	-3.04	2.42	-1.25	*p* = 0.216	*Participant*		
					Intercept	106.63	10.33
					Trial 2	28.82	5.37
					Trial 3	40.93	6.40
					Residual	222.43	14.91

Note: Number of observations = 2257; number of images = 32; number of participants = 48; number of groups = 2.

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 level. Irrespective of group membership, participants performed at an overall average of 98% accuracy (62.7/64; *SD* = 1.6%). 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 Appendix D.

Recall Memory

Two linear mixed-models (one for word-for-word 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 word-for-word and semantic recall memory performance are presented in Table 7 and Appendix E. The average relative strength of common ground formed between dyads in the Friends group across both rounds was 0.38 (*SD* = 0.34, min = -0.86, max = 0.98). The average relative strength of common ground formed between dyads in the Strangers group across both rounds was 0.36 (*SD* = 0.29, min = -0.71, max = 0.89).

The Best Fit Model for word-for-word recall memory performance had a maximal random-effects structure that included random intercepts and correlated 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, $\chi^2(4) = 4.94$, $p = .294$. Another alternative model that did not include the fixed effect of relative strength of common ground formation did not converge. 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. Dyads who established relatively stronger common ground formation for images during the RCT again exhibited superior word-for-word recall memory performance at follow-up. There was also a significant effect of role. On average, Directors recalled 3.7% more of the same words they used to describe images in the final trials of matching than Matchers. The effect of group in the Best Fit Model fell just short of the conventional level of significance. Best Fit Model coefficients for word-for-word recall memory performance in Experiment 2 are presented in Table 5.

Table 5

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

	Estimate	**SE**	***t*-value**	***p*-value**	**Random Effects**	**Variance**	***SD***
Intercept	0.254	0.020	12.41		Groups		
Common Ground	0.065	0.020	**3.18**	***p* = 0.004**	*Image*		
Group	0.045	0.025	1.81	$p = 0.077$	Intercept	0.002	0.045
Role	-0.041	0.009	**-4.50**	***p* < .001**	Common Ground	0.004	0.062
					Participant		
					Intercept	0.008	0.087
					Common Ground	0.002	0.048
					Residual	0.030	0.172

Note: Number of observations = 1472; number of images = 32; number of participants = 48; number of groups = 2.

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 *p*s > .05. The effect of group in the Best Fit Model fell just short of the conventional level of significance. On average, participants in the Strangers group recalled Directors' descriptions from the final trial of the RCT with 4.4% more semantic similarity than participants in the Friends group. Best Fit Model coefficients for semantic recall memory performance are presented in Table 6.

Table 6

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

	Estimate	**SE**	***t*-value**	***p*-value**	**Random Effects**	**Variance**	***SD***
Intercept	0.524	0.016	33.06		Groups		
Group	0.041	0.021	1.96	$p = 0.057$	*Image*		
					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 = 1472; number of images = 32; number of participants = 48; number of groups = 2.

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 4608 proportion similarity scores (e.g., 24 Friends Directors * 16 images = 384 recall memory descriptions * 12 naïve participant combinations = 4608). Descriptive statistics for each control distribution of the proportion similarity scores are presented in Table 7. For comparison, descriptive statistics from the recall memory task in Experiment 2 are also presented in Table 7.

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 2. The 95% confidence intervals associated with the mean of each control distribution are also outside of the range of scores obtained from the recall memory task in Experiment 2. For example, the overall mean of the distribution comparing the level of verbatim similarity between

control descriptions and Directors' recall memory descriptions from the Friends group was 0.116 (*SD* = 0.117). In comparison, the level of verbatim similarity observed in Experiment 2 between image descriptions used by Directors in the Friends group in the final trial of matching in the RCT and their recall memory descriptions was 0.261 (*SD* = 0.211). There is thus a mean difference of 0.145 (or 14.5%) between the two sets of verbatim similarity scores. Similar mean differences exist among the three other comparisons as well.

Table 7

Descriptive Statistics from the Control Distributions vs. Descriptive Statistics from the Verbatim Recall Exhibited by Participants in Experiment 2

Control Distributions			
Control vs. Director Recall Memory (Friends)	Control vs. Matcher Recall Memory (Friends)	Control vs. Director Recall Memory (Strangers)	Control vs. 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 to 0.12	95% CI =0.108 to 0.114	95% CI = 0.142 to 0.150	95% CI = 0.137 to 0.144
N = 4608	N = 4608	N = 4608	N = 4608
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 to 0.282	95% CI =0.200 to 0.241	95% CI = 0.269 to 0.308	95% CI = 0.235 to 0.273
N = 359	N = 359	N = 377	N = 377

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

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.

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 6, 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 from the rate by which dyads in the Friends group formed local common ground for facial images over time 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 in an attempt 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 of 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 verbatim and semantic recall memory models did not reveal a 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, a non-significant trend was surprisingly observed in the opposite direction. 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. The lack of a significant group effect in both types of recall memory performance may reflect a lack of power or the fact that both groups of dyads formed similar, relative levels of common ground for images during the RCT.

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. This is due to the fact that 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 a 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. Again, this is due to the fact that images can only be described in so many ways based on their distinctive characteristics, and that 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.

Chapter 4

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., Krauss & Weinheimer, 1964, 1966; Clark & Wilkes-Gibbes, 1986; Schober & Clark, 1989; Brennan & Clark, 1996; Ven der Wege, 2009; McKinley et al., 2017; Knutsen & Le

Bigot, 2018). They also 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 goal for the current set of studies was not merely to provide an online replication of previous RCT findings. Rather, our primary aim was to examine the influence of different forms of common ground on verbatim and semantic recall memory for conversation. In both experiments, this was carried out by computing a relative measure of the strength of local common ground formed between dyads for images during a RCT (Repp, 1976). Using linear mixed modelling, we then used this measure to predict participants' ability to recall image descriptions used by Directors during the final trials of the RCT about a week later. In Experiment 2, we directly tested whether pairs of friends with pre-existing personal common ground could form relatively stronger local common ground for facial images during a RCT and later exhibit superior recall memory performance for image descriptions used by Directors than strangers without personal common ground. This work was primarily motivated by a recent study by McKinley et al. (2017), who found a significant relationship between the strength of local common ground formed between dyads for images during a RCT and their immediate image recognition memory performance.

The results from the present experiments partially support the view that common ground and memory are intricately linked (e.g., 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 for 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. This was our expectation because 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). For example, in one of the first and only experiments to directly test verbatim recall memory for conversation 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 (Sachs, 1967; Bock & Brewer, 1974). 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. 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, Bautista, & McCarron, 2017). 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, Lo, & Cohan, 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 most 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.

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, Nastase, & Goldstein, 2020; Nastase, Goldstein, & Hasson, 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., 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. No significant group effects were observed between friends and strangers in recall memory performance.

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, Buck, & Dreyer, 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., Binet & Henri, 1894; Bartlett, 1932; Sachs, 1967; Potter & Lombardi, 1998). 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, Johnson, and Goldberg (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 and colleagues (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., Ross & Sicily, 1979; Stafford & Daly, 1984; Knutsen, Ros, & Le Bigot, 2016; McKinley et al., 2017), others have found the opposite effect (Stafford, Burggrad, & Sharkey, 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.

References

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

Bamford, J., & Wilson, I. (1979). Methodological considerations and practical aspects of the BKB sentence lists. *Speech-Hearing Tests and the Spoken Language of Hearing-Impaired Children*, 148-187.

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.

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

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Benoit, P. J., & Benoit, W. L. (1988). Conversational memory employing cued and free recall. *Communication Studies, 39*(1), 18-27.

Binet, A., & Henri, V. (1894). La mémoire des phrases. *L'annee Psychologique, 1*, 24-59.

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.

Bock, J. K., & Brewer, W. F. (1974). Reconstructive recall in sentences with alternative surface structures. *Journal of Experimental Psychology, 103,* 837-843.

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.

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,* 1-20.

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology, 2*(4), 331-350.

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.

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.

Burton, A.M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*(1), 286-291.

Clark, E. V. (2015). Common ground. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp.328-353). London: Wiley-Blackwell.

Clark, H. H. (1996). *Using language*. New York: Cambridge University Press.

Clark, H. H., & Brennan, S. A. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp.127-149). Washington, DC: APA Books.

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, England: Cambridge University Press.

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

Conneau, A., & Kiela, D. (2018). Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv: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.

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

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.

Gurevich, O., Johnson, M. A., & Goldberg, A. E. (2010). Incidental verbatim memory for language. *Language and Cognition*, *2*(1), 45-78.

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.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127-142.

Horton, W. S., & Gerrig, R. J. (2016). Revisiting the memory-based processing approach to common ground. *Topics in Cognitive Science*, *8*(4), 780-795.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General*, *116*, 26-37.

Johnson-Laird, P. N., & Stevenson, R. (1970). Memory for syntax. *Nature*, *227*(5256), 412.

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.

Kintsch, W., & Bates, E. (1977). Recognition memory for statements from a classroom lecture. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(2), 150-159.

Knutsen, D., & Le Bigot, L. (2014). Capturing egocentric biases in reference reuse during collaborative dialogue. *Psychonomic Bulletin & Review*, *21*(6), 1590-1599.

Knutsen, D., & Le Bigot, L. (2017). Conceptual match as a determinant of reference reuse in dialogue. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(3), 350-368.

Knutsen, D., & Le Bigot, L. (2020). The influence of conceptual (mis) match on collaborative referring in dialogue. *Psychological Research*, *84*(2), 514-527.

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.

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*(1), 113-114.

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.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., ... & Ferrari, V. (2018). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.

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.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

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.

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

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, *222*, 117254.

Neisser, U. (1981). John Dean's memory: A case study. *Cognition*, *9*(1), 1-22.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195-203.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543).

Pollmann, M. M., & 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.

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*(6), 633-654.

Potter, M. C., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language*, *38*(3), 265-282.

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

Redeker, G. (1984). On differences between spoken and written language. *Discourse Processes*, *7*(1), 43-55.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Repp, B. H. (1976). Identification of dichotic fusions. *The Journal of the Acoustical Society of America*, *60*(2), 456-469.

Rosen, S., & Corcoran, T. (1982). A video-recorded test of lipreading for British English. *British Journal of Audiology*, *16*(4), 245-254.

Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology*, *37*(3), 322-336.

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.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, *2*(9), 437-442.

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.

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). Mahwah, NJ: Lawrence Erlbaum

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

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.

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.

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.

Standing, L. (1973). Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology*, *25*(2), 207-222.

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology, 9*(1), 77-110.

Tulving, E. (1972). Episodic and semantic memory. Organization of memory. In Tulving, E., & Donaldson, W. (Eds.), pp. 381-402. New York: 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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998-6008).

Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*, 183-194.

Wilson, S. M., Bautista, A., & McCarron, A. (2018). Convergence of spoken and written language processing in the superior temporal sulcus. *Neuroimage, 171*, 62-74.

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), e19735.

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.

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.

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).

Appendix A: List of Referential Communication Task (RCT) Stimuli

Experiment 1

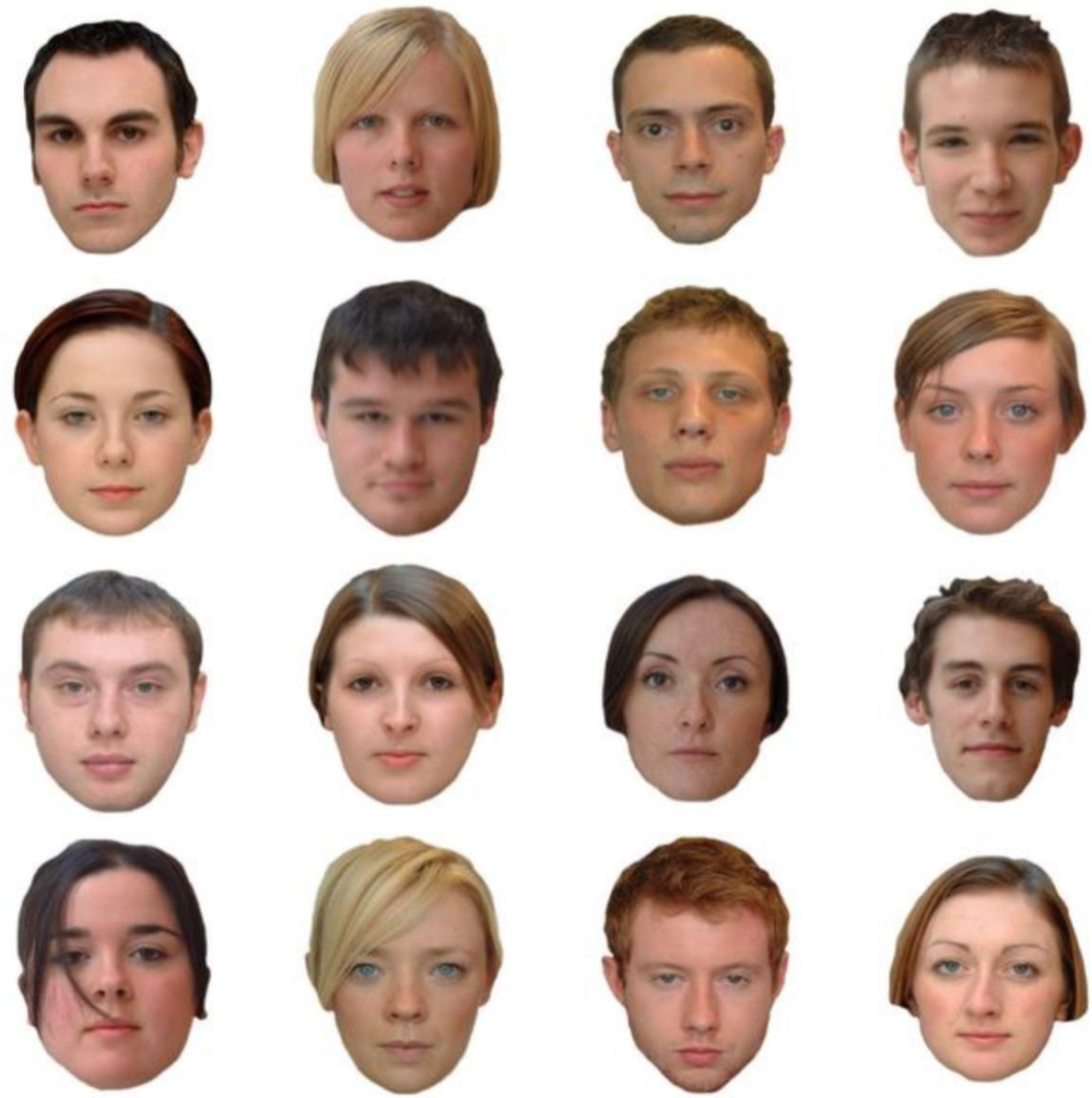
Round 1 (Eight Birds; Eight Horses)

Experiment 1

Round 2 (Eight Bowls; Eight Flowers)

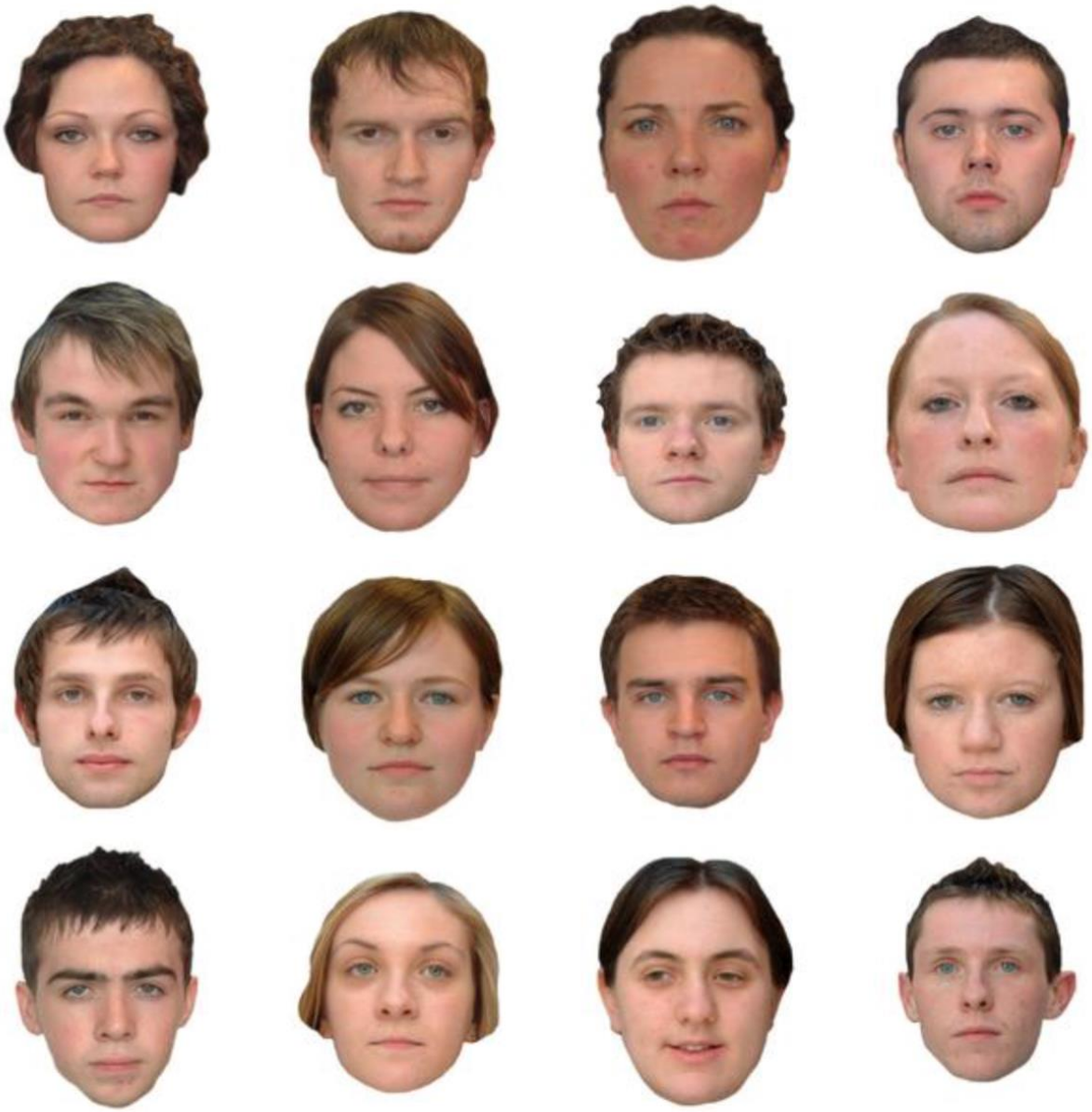
Experiment 2

Round 1 (Two Cliques of Four Women; Two Cliques of Four Men)



Experiment 2

Round 2 (Two Cliques of Four Women; Two Cliques of Four Men)



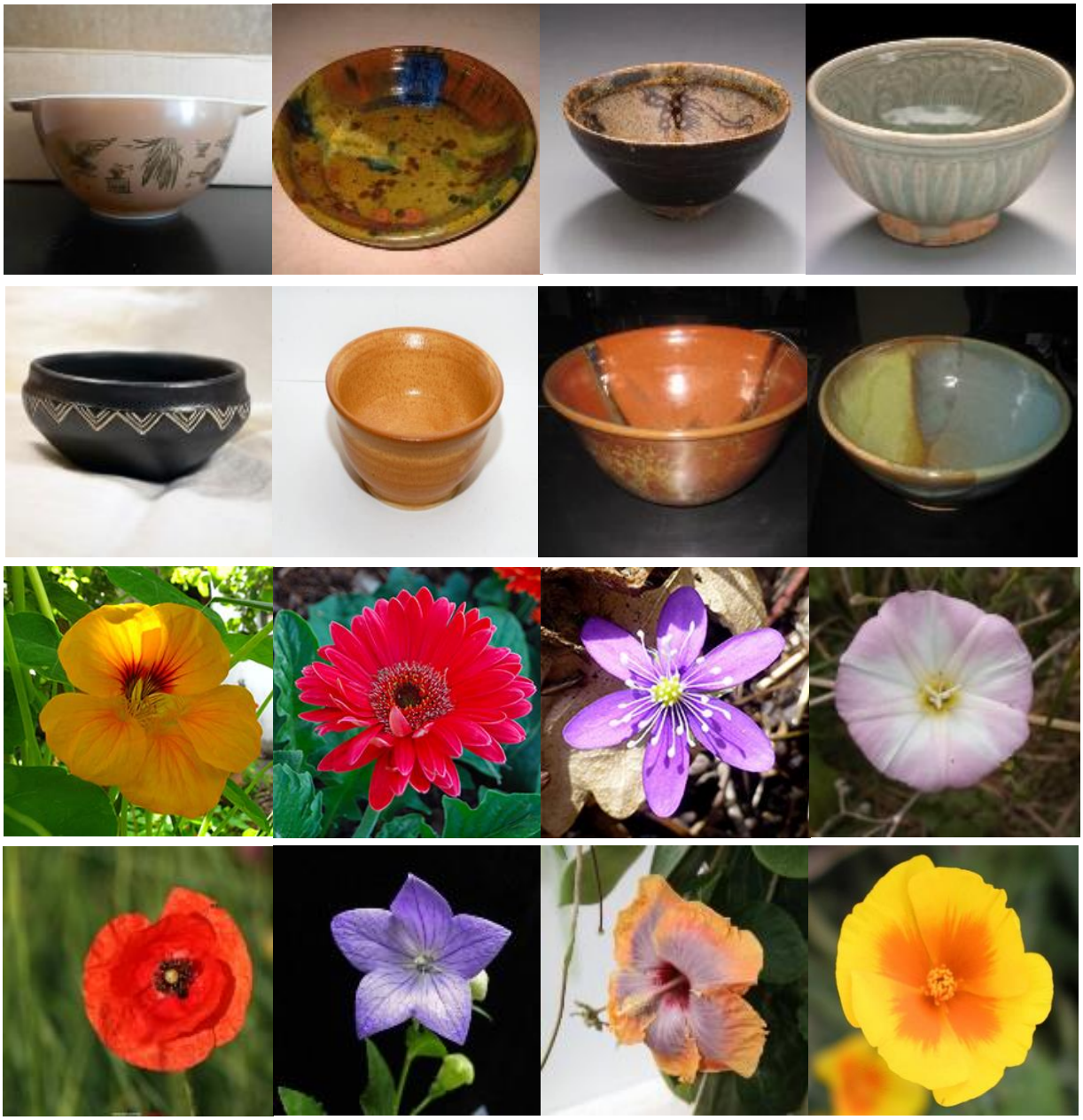
Appendix B: List of Foil Stimuli for the Recognition Memory Task

Experiment 1

Eight Birds; Eight Horses

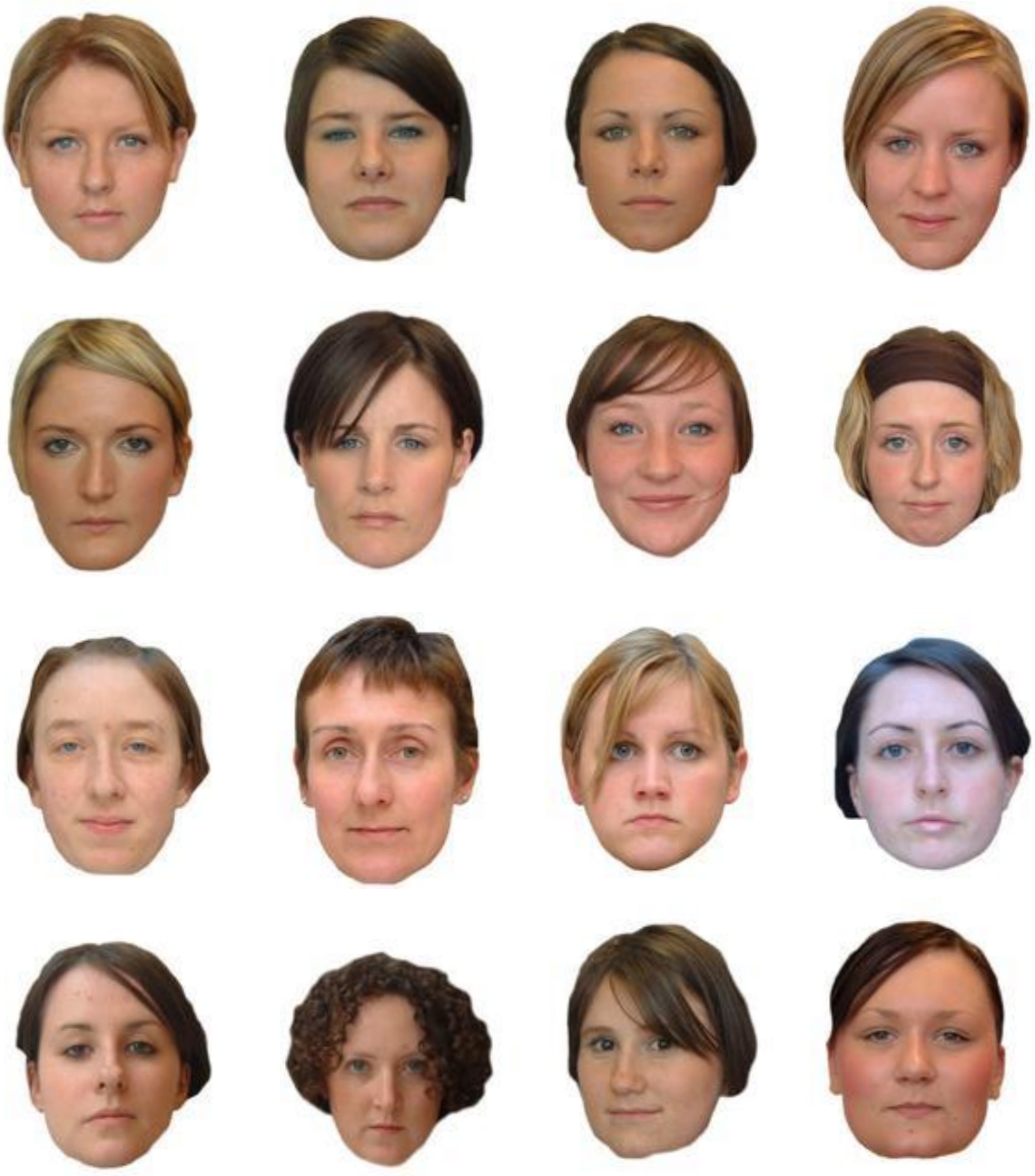
Experiment 1

Eight Bowls; Eight Flowers



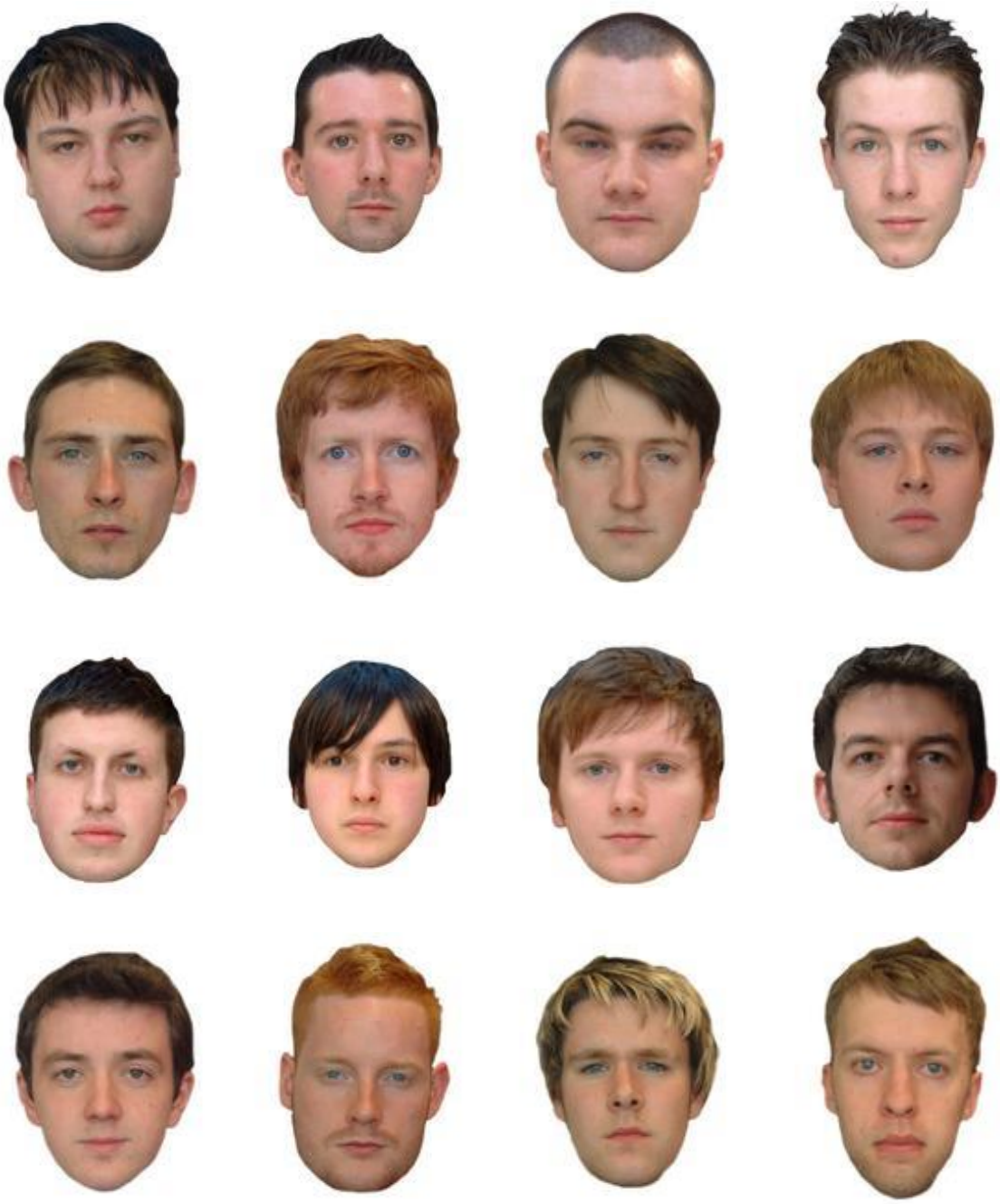
Experiment 2

Sixteen Women (Each Face Being Highly Similar to One Other Face in True Stimuli)



Experiment 2

Sixteen Men (Each Face Being Highly Similar to One Other Face in True Stimuli)



Appendix C: Descriptive Statistics for the Referential Communication Task (RCT)

Experiment 1

Table S1

Descriptive Statistics for 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; SD = standard deviation; n = number of trials.

Experiment 2

Table S2

Descriptive Statistics for 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; SD = standard deviation;

n = number of trials.

Appendix D: Descriptive Statistics for the Recognition Memory Task

Experiment 1

Table S3

Descriptive Statistics for the Recognition Memory Task in Experiment 1

	# of Images	Average % (SD)
Total	64	98.37 (1.81)
True Stimuli	32	97.40 (3.53)
Foil Stimuli	32	99.48 (1.50)
Director	16	98.18 (2.42)
Matcher	16	98.57 (2.25)

Experiment 2

Table S4

Descriptive Statistics for the Recognition Memory Task in Experiment 2

	# 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)	98.23 (2.64)
Director	16	97.92 (3.53)	98.10 (4.39)
Matcher	16	97.40 (3.65)	98.64 (2.64)

Appendix E: Descriptive Statistics for the Recall Memory Task

Experiment 1

Table S5

Descriptive Statistics for Recall Memory Performance in Experiment 1

Type of Recall	Role	n	Mean (SD)
Verbatim (Autoscore)	Director	356	0.414 (0.244)
	Matcher	356	0.375 (0.261)
Semantic (RoBERTa)	Director	356	0.685 (0.178)
	Matcher	356	0.620 (0.214)

Note: SD = standard deviation; n = number of trials.

Experiment 2

Table S6

Descriptive Statistics for Recall Memory Performance in Experiment 2

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

Note: SD = standard deviation; n = number of trials.

Appendix F: Ethics Clearance Letter

May 28, 2020

Dr. Kevin Munhall
Professor
Department of Psychology
Queen's University
Humphrey Hall
Kingston, ON, K7L 3N6

Dear Dr. Munhall:

GREB TRAQ #: 6026802
Title: "GPSYC-929-19 Multisensory cues for conversational coordination"

The General Research Ethics Board (GREB) has reviewed and cleared your request for renewal of ethics clearance for the above-named study. This renewal is valid for one year from June 6, 2020. Prior to the next renewal date, you will be sent a reminder memo and the link to ROMEO to renew for another year. You are reminded of your obligation to submit an Annual Renewal/Closure Form prior to the annual renewal due date (access this form at http://www.queensu.ca/traq/signon.html/ click on "Events;" under "Create New Event" click on "General Research Ethics Board Annual Renewal/Closure Form for Cleared Studies"). Please note that when your research project is completed, you need to submit an Annual Renewal/Completed Form in Romeo/traq indicating that the project is 'completed' so that the file can be closed. This should be submitted at the time of completion; there is no need to wait until the annual renewal due date.

You are reminded of your obligation to advise the GREB of any adverse event(s) that occur during this one-year period. An adverse event includes, but is not limited to, a complaint, a change or unexpected event that alters the level of risk for the researcher or participants or situation that requires a substantial change in approach to a participant(s). You are also advised that all adverse events must be reported to the GREB within 48 hours. To submit an adverse event report, access the application at http://www.queensu.ca/traq/signon.html/ click on "Events;" under "Create New Event" click on "General Research Ethics Board Adverse Event Form."

You are also reminded, that all changes that might affect human participants must be cleared by the GREB. For example, you must report changes in study procedures or implementation of new aspects into the study procedures. Your request for protocol changes will be forwarded to the appropriate GREB reviewers and/or the GREB Chair. To submit an amendment form, access the application at http://www.queensu.ca/traq/signon.html/ click on "Events;" under "Create New Event" click on "General Research Ethics Board Request for the Amendment of Approved Studies."

Note: Due to COVID-19, human participant research policies, in relation to hospital and non-hospital based research, are being continually updated. Many restrictions are now in place with respect to in-person research. For the most current information on the COVID-19 impact on research, please visit https://www.queensu.ca/vpr/covid-19. For information directly related to GREB please visit the Research Ethics FAQs.

On behalf of the General Research Ethics Board, I wish you continued success in your research.

Yours sincerely,

Chair, General Research Ethics Board (GREB)
Professor Dean A. Tripp, PhD
Departments of Psychology, Anesthesiology & Urology Queen's University

c.: Daniel Nault and Cynthia Sedlezky, Co-investigators
Dr. Luis Flores, Chair, Unit REB