UNIVERSITY OF CALIFORNIA
SANTA CRUZ

COLLABORATION AND POLITENESS IN THE CONVERSATIONS OF FRIENDS AND STRANGERS IN COMPUTER-MEDIATED TEXT-BASED CHAT OVER TIME

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PSYCHOLOGY

By

Kris Liu

December 2015

The Dissertation of Kris Liu is approved:

_____
Professor Jean Fox Tree

_____
Professor Travis Seymour

_____
Professor Steve Whittaker

_____
Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

Abstract

Collaboration and politeness in the conversations of friends and strangers in computer-mediated text-based chat over time

Kris Liu

This dissertation examines whether computer-mediated, text-based conversation (chat) between friends and strangers differ in efficiency and politeness over time. Experiment 1 is a longitudinal corpus collection task, which is analyzed to see if there are differences in how efficient friend and stranger dyads are at completing collaborative tasks and whether these differences persist across the three-week study. Experiment 1 participants did an online version of a traditional referential communication task (the tangram task) every week, which became practiced over time, as well as a novel puzzle task that changed every week. (Experiment 1 analysis only analyzes the tangram task data.) Experiment 2 uses stimuli taken from both tangram and puzzle tasks and looks at how third-party 'over-reader' judgments on politeness evolve across three weeks. Experiment 3 expands on this by systematically manipulating the dialogue taken from Experiment 1 to demonstrate that politeness is not purely determined by a speaker's intention, as suggested by the dominant theory of politeness by Brown and Levinson (1987). The results of Experiment 1 indicate that there were few differences between the number of unique descriptive words that friend and stranger dyads used, though friends tended to take more turns than strangers. The results of Experiment 2 show that third-party over-readers are bad at explicitly distinguishing between friend and stranger dyads in both the practiced and novel tasks, though they do rate strangers as being more polite than friends at first; this difference evaporates by the second week. The results of Experiment 3 suggest

that though an obviously impolite utterance will never be considered very polite (and vice versa), an interlocutor's response to an utterance can nonetheless influence how polite an utterance sounds. Furthermore, those dyads thought to be friends are granted more flexibility in their utterances than those thought to be strangers: impolite utterances can be judged as neutral when over-readers think they are reading the conversation of friends but remains impolite when they think they are reading the conversation of strangers.

Acknowledgments

I admit that I originally wanted to skip this optional page. But I would have been severely remiss to not thank the people who have helped me get to this point. I am more thankful than I know how to express for the unwavering support provided by my advisor, Dr. Jeannie Fox Tree: a graduate student could not ask for a better advisor. I would also like to thank Dr. Steve Whittaker for the time he has put in on my quals and dissertation committees and Dr. Travis Seymour for having been on every one of my committees. Thanks also to Dr. Doug Bonett, who provided advice on statistics, and Dr. Lyn Walker, who influenced the direction of my research.

I am indebted to many others, such as my many research assistants who have provided invaluable help; I would still be collecting data for this dissertation today if it were not for them. Members of the Fox Tree Lab – particularly Charlotte Zeamer, Natalia Blackwell and Jackson Tolins – who made lab meetings, lab gatherings and conferences extremely entertaining. Members of the Wellesley College alum community in academia, with whom I have commiserated and procrastinated through endless conversation online.

I am deeply grateful to Jeff Warshaw for believing in me and feeding me throughout graduate school. I apologize for all of the times I let dinner go cold while working and the many dubious looks I gave you during pep talks.

And, finally, to my parents, Josephine and Arthur: thank you for all your support and love. And thank you for not asking me too frequently when I would be done with my doctorate!

Introduction

Kris*: Sorry for bothering you about this but I'm running late and was hoping you could get a bottle of white wine for the table. Would it be possible for you to do that? No worries if not!*

Kris' Friend*: WTF? Why are you treating me like a stranger?*

The above text exchange took place at one of my college reunions. In my mind, I was imposing on this acquaintance by asking her to go wheedle a bottle of wine out of the lunch caterers. In her mind, I acted too deferential, used too many words, and treated her like she was someone I barely knew. I had treated her as if she were someone I was not supposed to impose upon and what are friends if not people you casually impose upon? She, like many others, had an intuitive idea that friends and strangers sound different when addressing each other. I did too: until the day before, I had not seen this person for five years.

Anecdotes aside, do friends and strangers actually talk to each other in quantitatively different ways? And how familiar does a pair of strangers have to be before they start relaxing into the quick, casual imposition that is thought to occur between friends? This dissertation is comprised of three studies that examine politeness over time in text-based, task-oriented conversations between friends and strangers. The first section will be a brief literature review on the following topics: discourse between friend and stranger dyads, politeness in face-to-face communication and politeness in computer-mediated communication. Experiment 1 is a longitudinal corpus collection task, which is analyzed to see if there are differences in how efficient friend and stranger dyads are at completing collaborative

tasks. Experiment 2 uses stimuli taken from Experiment 1 and looks at how third-party judgments on politeness evolve across the span of Experiment 1. Experiment 3 expands on this by systematically manipulating the dialogue taken from Experiment 1 to demonstrate that politeness is not purely determined by a speaker's intention.

Throughout the dissertation, I will show that it is surprisingly hard to distinguish between friends and strangers in certain types of text-based conversation. Not because typed dialogue is inherently less rich, but because the differences between friend and stranger conversations can be fairly subtle.

Conversation and Collaboration Between Friends and Strangers

Do people speak (or type) differently to friends than they do to strangers in conversation? The evidence is mixed, suggesting that differences can be subtle and possibly highly context-dependent. When naïve judges were played short clips of friends' and strangers' face-to-face conversations, the naïve judges had an approximately 80% accuracy rate discriminating between the two. Participants cited conversational content (e.g. knowledge about the other's life, how intimate the topics were), as well as style and manner (e.g. formality, relaxedness) as being most informative (Planalp & Benson, 1992). However, subsequent analyses showed that content intimacy was not a discriminating factor, though partner knowledge, a sense of relationship continuity (a mutually-shared past and future) and linguistic formality were (Planalp, 1993). The vast numbers of internet forums dedicated to the discussion of stigmatized topics such as mental illness are populated almost entirely by strangers

quickly lays to rest the idea that content intimacy (De Choudhury & De, 2014) has a strong correlation to acquaintanceship.

Even when you take away clear cues referring to a pre-existing relationship, there are still other cues that third parties could use. Friends use more discourse markers than strangers (Fox Tree, 2007; Macaulay, 2002), overlap more turns (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Dunne & Ng, 1994), and leave more information unspoken in openings and have more complex closings (Hornstein, 1985). Friends are also more likely to recognize sarcasm than strangers, though they may not use sarcasm more (Rockwell, 2003).

For all the shared history and greater shared common ground, it is unclear whether working with a friend is more effective than working with a stranger. People tend to be better at guessing the thoughts of friends (Colvin, Vogt, & Ickes, 1997) but this advantage does not necessarily lead to better performance on complex cognitive tasks. In fact, it more consistently leads to overconfidence and less accurate judgment of performance. There is evidence that friends who work together seem to be better at avoiding the costs of collaborative inhibition in retrieval of certain types of information (Andersson, 2001; Andersson & Rönnberg, 1995), though friends may also be more prone to influencing false memories in each other (Hope, Ost, Gabbert, Healey, & Lenton, 2008). Friends also overestimate their success this collaborative recall tasks, even in cases where they do not outperform strangers (Gould, Osborn, Krein, & Mortenson, 2002).

In terms of communication-specific tasks, friends are better at sending covert messages that go undetected to overhearers within larger messages (Fleming, Darley, Hilton, & Kojetin, 1990). They are also better at identifying figures that are described by friends rather than strangers, though there is no difference in mean length of the description (Fussell & Krauss, 1989). Yet, friends are only marginally more successful than strangers at correctly interpreting ambiguous statements. Echoing the collaborative recall literature, friends also more likely to overestimate their communicative success (Savitsky, Keysar, Epley, Carter, & Swanson, 2011). In certain situations, it may be that working with a friend makes one feel better about one's performance without actually benefitting performance.

Work done on the HCRC Map Task Corpus (Anderson, et al., 1991), which has equal numbers of friend and stranger dyads, has produced very few published studies on the communication between friend and stranger dyads. This may suggest a positivity bias at work (cf. Fanelli, 2010), as a prominent factor in a well-studied corpus is likely to have been examined for group differences; lack of published results in this case may indicate lack of positive results, rather than lack of study.

One of the few studies published on the differences between friends and strangers in the Map Task was on laughter: while laughter has been documented to be more common and of a different quality between friends engaged in conversations or games (Campbell, 2007; Smoski & Bachoroski, 2003). Truong and Trouvain (2012) were unable to find such differences within the Map Task corpus. However, this may be because conversations that are short, goal-oriented and constrained by the

pressures of a laboratory setting do not allow for variance in communicative or collaborative behaviors. It is possible that a less sterile setting, a more engaging task or a task that requires repeated collaboration between speakers could bring out differences that are otherwise not noticeable in most one-shot studies done in the lab. Liu, Fox Tree, Blackwell and Walker (*unpublished manuscript*) demonstrated that, when taken out of the lab and given a longer task, friends and strangers could be distinguished when extraversion was controlled for. Additionally, stimuli taken from this task showed that third-party overhearers were remarkably bad at explicitly telling apart friends and strangers but nonetheless reliably judged strangers to be more polite than friends. The difference, however, between friends and strangers narrowed significantly when comparing the beginning of the dialogue vs. 20-30 minutes into the dialogue (Liu & Fox Tree, *unpublished manuscript*).

A longitudinal corpus of online conversation conducted with participants taking part in an environment they find comfortable, using another type of referential communication task, the tangram task (Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989) was collected for this dissertation: this was done in order to elicit conversation that is still goal-oriented but is possibly more likely to result in more nuanced conversational styles that may show more variance between friend and stranger dyads. The first study of this dissertation focuses purely on how effective friend and stranger dyads are at this task. The second and third experiments focus on politeness, a social aspect of conversation. Planalp (1993) showed that people distinguished between the conversations of friends and strangers based on formality,

one aspect of politeness. There is disagreement, however, what exactly constitutes linguistic politeness, though theorists agree that what is polite depends somewhat on the relationship between speakers. I am particularly interested in two opposing theories: Brown and Levinson's (1987) Politeness Theory and Arundale's Conjoint Co-Constitution Model (1999).

Politeness

If consensus in the politeness literature is measured by the number of papers that use a certain theory, then Brown and Levinson's (1987) Politeness Theory is the best fitting theory for how politeness works. However, it does seem to be the case that it is the most widely cited because it is the most specific and testable theory of politeness. Brown and Levinson posit that people are worried about maintaining *face*: both their own face and the face of others. Speech, particularly requests, is adjusted depending on social distance, relative power, and the degree of imposition of the request. Adjustments are made primarily to the directness of the speech. According to Politeness Theory, the most impolite is the most direct type of statement ("Shut the window") while the most polite is an utterance that is indirect and off-record ("Oh, it is kind of cold in here"). However, this theory has been shown to be inaccurate for many situations and cultures (Arndt & Janney, 1985; Matsumoto, 1988; Wolfson, 1983) and only discusses politeness on the level of individual utterances, rather than looking at the larger context of the conversation (or even adjacent utterances).

Arundale (1999), on the other hand, almost takes the opposite tack: that no utterance is inherently polite or impolite. Politeness is determined jointly *after* an

utterance and what is polite for that utterance is situation- and dyad-specific: it is determined by what the conversation is on, the history of conversations between the speakers, the knowledge they have about the other, where the conversation is taking place, what it was about, and so on. Conceivably, the same utterance can be considered impolite one minute and polite the next, depending on how the conversation develops. It is also conceivable that a dyad operates by guidelines exactly as Levinson and Brown lay out because that is what is necessary to be successful in conversation with each other.

It seems reasonable to hypothesize (under both Brown/Levinson and Arundale) that the more people are comfortable and acquainted with each other, the more willing they are to say things that are considered impolite to overhearers who do not know them or their history. Providing evidence for, or against, Arundale's theory of conjointly determined politeness is then a matter of showing two things: 1. That the perceived politeness by an overhearer of a speaker's utterance can change depending on the response of the speaker's interlocutor (e.g. Haugh, 2007) and 2. that the politeness of an impolite utterance depends on whether the overhearer thinks they are listening to friends or strangers speaking.

Politeness in Text-Based Discourse Between Human Interlocutors

To test whether politeness is expressed in the form and phrasing of an utterance or mutually constructed between interlocutors, I have constrained the scope of this dissertation to text-based communication. Text-based interaction supposedly lacks the nuance of co-present face-to-face interaction (Clark & Brennan, 1991; Daft

& Lengel, 1986), but also avoids the confounds that intonation and prosody may introduce into the interpretation of emotion in speech (Ishii, Reyes, & Kitayama, 2003).

Text-based interaction is not only prevalent, it can create genuine emotional closeness. Even work done 8-10 years ago indicates that many close relationships are well-served by keeping in touch primarily via text-based methods. Cummings, Lee, and Kraut (2006) found that college students were more likely to be closer to high school friends with whom they stayed in touch via email or online chat than those they called or saw in person. Similarly, Walther and Bazarova (2008) found that non-romantic relationships between people who were accustomed to communicating online were as strong as those who used other methods of communication. Interlocutors who communicate using text-based online methods also mitigate the potentially impoverished signal by adapting their behavior to the medium (Carlson & Zmud, 1999; D'Urso & Rains, 2008) or try to make the conversation seem more like a face-to-face conversation (J. F. Anderson, Beard, & Walther, 2007).

There has been some research on politeness in different text-based modes of communication and within different contexts. This work has been mostly based on Brown and Levinson's Politeness Theory and on other easily quantified markers of politeness.

Email. Mutuality was found to be important when it came to email politeness and etiquette. Mutual expectations resulting from the roles and past history of communication influence the explicitly stated etiquette preferences for email

exchanges over time between faculty and students in a university setting (Lewin-Jones & Mason, 2014). Between strangers, Bunz and Campbell (2004) showed that emails that contained politeness markers such as "please" and used proper salutations and closing remarks were more likely to produce responses that also contained these formalities. Francis, Holmvall and O'Brien (2015) found an analogous pattern of incivility begetting incivility (where civility is defined by a subjective rating), particularly in high-workload situations.

Online Communities. Work on politeness in online communities has shown that community norms and community-accepted hierarchy can be influential in how politeness in enacted in these groups. Standards of politeness may also vary depending on the norms of the community: under the assumption that increased number of replies is a desirable outcome, Burke and Kraut (2008) found that greater politeness is more effective in technical newsgroups while rudeness is more effective in political newsgroups. A study of open-peer review systems, where peer reviews of an academic paper are signed and posted in a public forum, shows that less experienced reviewers tended to be more harsh and use fewer mitigating strategies than more experienced reviewers. Moreover, the manuscripts by less experienced researchers received more statements of positive politeness (e.g. compliments) than more experienced researchers (Nobarany & Booth, 2014).

A study that defined politeness in short messages as a combination of Brown and Levinson-like indirectness and typical politeness markers found that as people ascend in an online community's hierarchy (either becoming an admin editor at

Wikipedia, or gaining reputation at Stack Exchange), their messages become less and less polite (Danescu-Niculescu-Mizil, Sudhof, Jurafsky, Leskovec, & Potts, 2013). In contrast, requests that contained politeness markers or hedges in the Random Acts of Pizza forum on Reddit were not more likely to result in strangers sending each other pizza (Althoff, Salehi, & Nguyen, 2013).

Instant Messaging. Work that is specifically on dyadic chat suggests that instant messaging is often thought of as an informal communication channel that can be used for both simple (Nardi, Whittaker, & Bradner, 2000) and complex conversational interactions (Isaacs, Walendowski, Whittaker, Schiano, & Kamm, 2002) between coworkers as well as between friends. Linguistically, a dyadic IM interaction more closely resembles spoken communication rather than written communication (Baron, 2010) though familiarity with the medium does tend to influence the length of turns (Fox Tree, Mayer, & Betts, 2011). This perception of IM as an informal communication channel that is largely used between people who already know each other, combined with the time/effort cost of being polite (e.g. Brennan & Ohaeri, 1999), may be part of why IM conversations are perceived to be less polite than face-to-face conversations. There is some evidence that the closer people are, the less formal and polite they feel they need to be when speaking via IM (Darics, 2014; Lam & Mackiewicz, 2007).

However, when situations call for politeness, IM is as conducive to relaying politeness as speaking face-to-face, even when IM is perceived as being less formal. Politeness in online chat has been extensively studied within the very specific context

of reference librarians' interactions with college-age patrons, as it has been found to be a major determinant of patron satisfaction (Mon, 2006). Due to the roles of the players in this type of conversation (a librarian whose professional training emphasizes the need for polite, cordial behavior and a student who is "imposing" on a professional for help), polite behavior of some sort is likely to be present. However, there are subtle differences in politeness in online vs. face-to-face "reference interviews": Carlo and Yoo (2007) found that both librarians and patrons used more negative politeness strategies (e.g. apologizing for imposition, being indirect, acting deferential, asking questions) in online requests than in face-to-face requests. Both librarians and patrons assume that a reference chat is meant to be quick but often need to step away from the chat to look for something. As a result, positive politeness strategies (e.g. joking or asserting common ground) are lost and negative politeness when pointing out necessary pauses in the chat is increased. Furthermore, formality from the librarian is associated with being robotic, impersonal and condescending while informality is associated with a sense of interpersonal connection (Waugh, 2013). These findings, however, may not be generalizable for many reasons, not the least of which is that the reference interview is essentially a professional script on the librarian's side.

Current Study

This dissertation contains three sets of studies, all focusing on text-based, online chat from the perspective of interlocutors as well as over-readers. Experiment 1 is a longitudinal corpus collection that examines how efficiently friend and stranger

dyads are able to perform over time on a referential communication task: while previous studies did not find major differences between friend and stranger dyads in referential communication tasks, I hypothesize that this is because those tasks were essentially one-shot conversations (even if the task itself was repeated within that session) within a sterile lab environment that was not conducive towards bringing out the differences between friend and stranger dyads in conversation. This may change when performing these types of tasks repeatedly over a longer span of time and within a relaxed environment. I predict that friends will have an immediate advantage and will be able to complete the task with fewer turns and relevant descriptive words than strangers in the first session but that this difference will dissipate as strangers learn how to adjust their conversation to their interlocutor. (Dyads will also do a novel task each week, though that data is not explicitly analyzed for this dissertation.)

Experiment 2 uses the excerpts from the naturalistic experiment done in Study 1 to look at how *over-readers* (third parties who "overhear" these chats) perceive politeness in text-based conversation. Specifically, I examine whether perceived politeness and perceived closeness of dyads is influenced by whether a task is practiced or novel, as well as whether it matters that the excerpts were taken from the first, second or third weeks of Experiment 1. I predict that strangers will always seem more polite than friends, but that this difference will decrease over the three weeks. This trend towards sounding less polite will be particularly strong for the practiced task but not for the novel task.

Experiment 3 orthogonally manipulates some of those excerpts to examine whether over-readers' perception of politeness changes depending on how polite or impolite their interlocutors are acting. According to Brown and Levinson, the form of a single utterance determines its politeness, while Arundale posits that the response from an interlocutor is also influential towards determining an utterance's politeness. If Brown and Levinson are more accurate in their characterization of politeness, then an interlocutor's response does not matter – an utterance will be rated similarly in terms of politeness regardless of what the interlocutor says. If Arundale more accurate, then the politeness of an utterance changes depending on how it is received by the interlocutor.

Experiment 1

Experiment 1 is a longitudinal corpus collection of online chat dialogue between dyads performing two types of tasks over the course of three weeks. Over the three sessions, one task became practiced, while the other one was always novel. The practiced task was a two-round tangram task, a type of referential communication task whereby a director led a matcher to pick out a series of tangrams in a specific order. This has traditionally been used to show how interlocutors negotiate referring expressions and reuse those expressions in subsequent utterances. For Experiment 1, I look at whether friend or stranger dyads have an advantage in terms of the speed at which they are able to complete the task (as measured by number of relevant descriptive words). Participants do this task every week so that it becomes a task that

is predictable for them: they can develop strategies and learn their partner's preferences in descriptions.

The second task is a novel task that changes every week: puzzles that draw on different skill sets and is less conducive to the same give-and-take exchange that the tangram task lends itself to. This data is not analyzed for Experiment 1 but the dialogue is repurposed as stimuli for Experiments 2 and 3, which focus on whether the perception of politeness changes over time differently for friend and stranger dyads.

Participants. Stranger and friend participants were recruited using several different methods: the UCSC participant pool, acquaintances of the lab's research assistants (none of which were psychology or cognitive science majors), and recruiting students from various courses (mostly introductory statistics and clinical psychology courses). Those recruited through the participant pool were given five credits for their participation, which essentially consisted of five different sessions: an initial recruitment survey, three experimental sessions with a partner and a final session which included personality questionnaires, working memory capacity tasks and a post-experiment questionnaire or interview. Those recruited through research assistants were given a $30 Amazon gift code. Those recruited through their classes were given a choice of the gift code or participant pool credit. About two-thirds of the stranger dyads came from the participant pool, while almost all the friend dyads came from in-person recruitment efforts.

Methodology

Friend and stranger dyads perform on an online version of the tangram task and novel puzzles over the course of three weeks. Each participant was assigned to an anonymous Google account labeled either Participant1 or Participant2 (or alternative accounts, Participant3 and Participant4, when two different dyads were scheduled for the same timeslot); they used the same accounts throughout the entire duration of the experiment. Most conversation during the experiment was conducted through a Google group chat with both participants and the experimenter. However, to keep Matchers from seeing the Directors' tangram arrays, experimenters either sent links in separate emails to the anonymized accounts or through individual chat windows.

Initial recruitment survey (done online or in-lab). This survey was intended meant to solicit participants through the participant pool by informing them of the structure of the study and the time commitment required. I also used it to gather basic information about their familiarity with online and text-based communication methods and asked participants to take a one-minute typing test that measured typing speed and accuracy. This was meant to limit partners who were mismatched in their ability to keep up with the other. I also asked potential participants about their weekly availability for the next several weeks. Ultimately, the dearth of participants who wanted to participate in such a lengthy study meant that partners were primarily matched on availability, though an effort was made to ensure participants did not differ more than 20-30 wpm. Non-participant pool participants did not take this survey.

Tangram task. The tangram task was set up as a two-round task, with both rounds done in rapid succession. Directors were presented with an array of nine tangrams in a specific order that they had to lead the Matchers to replicate. Matchers saw a layout with twelve tangrams, to hinder the use of a process of elimination strategy towards the end of each round. The second round presented Director and Matchers with the same tangrams but in a different order. The first week used nine novel tangrams; subsequent weeks would introduce seven new tangrams but pulled two tangrams from the previous week into the array. Tangrams were not randomized: there were six preset arrays for Directors and six preset arrays for Matchers. Participants were not informed that the arrays were preset.

Directors were asked to go in order (right to left, top to bottom), an instruction that was usually observed but occasionally violated for different reasons (e.g. not heeding instructions, correcting errors made along the way or as a strategy adopted by the participants). Matchers then dragged and dropped tangrams into the order they thought the Director was describing. Participants were limited to 20 minutes a round, which generally provided sufficient time for completing all nine tangrams. Participants were shown their score after each round.

Participants switched off as Director and Matcher: in Week A, Participant1 was Director, in Week B, Participant2 was director and in Week C, participants would be asked to assign themselves (mutually agreed upon) roles.

Puzzle tasks. Every week had a novel 15-minute puzzle task where participants are given the choice of quitting after 10 minutes. Puzzle tasks were

deliberately chosen to be radically different week to week, so that participants could not rely on a previously utilized strategy to solve them. Actual accuracy and performance on these puzzles, while measured, was not analyzed for this dissertation. Rather, the puzzles were used to generate problem-solving dialogue, where dyads could choose how actively they coordinated their puzzle-solving efforts.

In order to minimize the ability for participants to solve the puzzles by searching out the answers the internet, three puzzles from old back issues of Games Magazine were chosen. Puzzle A (Week A, *Common Elements*) resembled a Remote Associates Test, where participants were given arrays of line drawings and asked to 1. determine what characteristic was shared in a series of line drawings, and 2. choose another line drawing from a separate lineup that also shared that characteristic. Puzzle B (Week B, *Reading Between the Lines*) was a visual puzzle where two classic book titles are superimposed: though finding the solutions is considerably faster if participants have heard of these well-known books – or movies based on the books – it is entirely possible to solve the puzzle by looking at the ways the letters are formed. Puzzle C (Week C, *Logi*-quiz) featured traditional word puzzles that scaled up in difficulty (fill in the missing letters from pairs of words, starting from non-scrambled words to anagrams). It was possible to solve the entire puzzle fairly quickly if participants figured out the metapuzzle (missing letters in a set spelled out names of American cities) but almost no participants did so. All puzzles can be found in Appendix B. The puzzle task data was not analyzed for this study as the puzzles were

purposefully different from each other. Excerpts from the dialogues were, however, used for Experiments 2 and 3.

Other Tasks. Participants were asked to do several other tasks during the course of this study. These will be used for future analyses of this corpus but will not be analyzed in this dissertation.

- *Personality questionnaires*. Two Big Five personality questionnaires were administered at the end of the study: the Ten-Item Personality Inventory (Gosling, Rentfrow, & Swann, 2003) and the Big Five Inventory (John, Donahue, & Kentle, 1991).
- *Working memory capacity tasks (WMC).* Online versions of reading span (RSPAN), operation span (OSPAN) and count span (CSPAN), as described by Just and Carpenter (1992) were provided to participants in random order. Participants were given a couple of short practice trials for each and then progressed through arrays consisting of two to seven items; each set size was repeated twice (Conway et al., 2005; Just & Carpenter, 1992).
- *Post-experiment questionnaire/in-person interview*. Participants were either asked to fill out a post-experiment questionnaire or come in for an interview where the same questions were asked (but where they could elaborate on their responses if they wished). Participants were asked about their perceptions of task difficulty for both tangrams and puzzles, comfort with their partner, their use of computer-mediated communication, and

other questions that related to their participation in the study. Those who came in for the in-person interview were also walked through all of their transcripts and asked to clarify any dialogue exchange that was unclear in intent or tone.

Corpus Coding

Descriptors and Turns within trial were coded separately for Directors and Matchers. Descriptors were defined as unique-within-round adjectives or nouns that were descriptive of the target objective, such as colors, shapes, media type (e.g., painting, sculpture, metal, stone) and patterns (e.g. striped). To avoid arbitrarily determining what counts as synonyms, similar Descriptors were counted separately. For example, the uses of turquoise, blue, and "kind of blue" were treated as unique. Each unique Descriptor was only counted once per trial.

Turns were counted for both Directors and Matchers as well. They were defined as each separate line that is used by a participant, regardless if they cede the floor immediately thereafter (i.e., every time a person hits the "return" key is considered a turn). Backchannels such as "ok" and "got it" were also counted as a turn.

Results

Unless otherwise indicated, all post-hoc pairwise comparisons were adjusted for multiple comparisons using the Sidak adjustment.

Participants. Attrition for this experiment was substantial. Due to the logistical difficulties of scheduling three people simultaneously (two participants and

an experimenter), a "week" could be anywhere from five to 10 days: research assistants were instructed to schedule participants for the same day and time every week but scheduling conflicts, missed sessions, and technical problems meant that sometimes sessions could not be equally spaced. Of the 154 dyads recruited, 73 did not complete the three central experimental sessions. An additional 20 did not have complete, or near-complete, data for the three central experimental sessions; this was either due to experimenter error, participants' technical issues or participants' being late (and thus running out the time their partners set aside for their participation).

Figure 1. Examples of Director Strategies in Experiment 1 (Tangram Task).

There are essentially three strategies for completing the tangram task for Directors after their description of any given tangram (assuming the Matcher does not have follow-up questions). A Director can wait for the Matcher to confirm they found the tangram and acknowledge the Matcher's success. She can wait for the Matcher to confirm they found the tangram and not acknowledge the Matcher's success. Or she does not wait for the Matcher at all.

Director Strategy #1 (wait and acknowledge; two contributions per trial)
Director: Tangram 1 is the guy with the bulging calf
Matcher: Got it
Director: Great
Director: Tangram 2 is the Pokémon

Director Strategy #2 (wait but don't acknowledge; one contribution per trial)
Director: Tangram 1 is the guy with the bulging calf
Matcher: Got it
Director: Tangram 2 is the Pokémon

Director Strategy #3 (don't wait; one contribution per trial)
Director: Tangram 1 is the guy with the bulging calf
Director: Tangram 2 is the Pokémon

Director Descriptors. There was no main effect of Acquaintanceship for Director Descriptors, $F(1, 51) = .138$, $p = .71$, $\eta_p^2 = .003$. There were no interactions between Acquaintanceship and either Week or Round. Friend and stranger dyads were indistinguishable in terms of how detailed they were in their description of tangrams, even after six rounds of the task over three weeks (see Figure 2).

There was a main effect of Week, $F(1, 51) = 23.41$, $p < .0001$, $\eta_p^2 = .32$. Pairwise comparisons indicate that there was a difference from Week A ($M = 6.75$, $SD = 2.08$) to Week B ($M = 5.58$, $SD = 1.65$), $t(51) = 5.05$, $p < .0001$, 95% CI [0.60, 1.75], Cohen's $d_z = 0.69$. There was no difference between Week B and Week C ($M = 5.33$, $SD = 1.61$, $t(51) = 1.12$, $p < .54$, 95% CI [-0.24, .723].

There was a main effect of Round. Directors used fewer Descriptors in Round 2 ($M = 5.16$, $SD = 1.56$) than in Round 1 ($M = 6.12$, $SD = 1.92$), $F(1, 51) = 93.49$, $p < .0001$, $\eta_p^2 = .65$. There was no interaction between Week and Round: Directors seemed to shorten their referring expressions between the two rounds by similar amounts each week.

Director Turns. The correlation between Director Descriptors and Director Turns for each round range from anywhere between $r = .11$ (weakly related and not statistically significant) to $r = .58$ (strongly related and significant), so separate analyses were carried out for these two dependent measures. Previous research indicates that the more familiar people are with instant messaging, the more turns they take (Fox Tree et al., 2011). It may also be the case that people use more (possibly shorter) turns the more comfortable they are with a specific interlocutor.

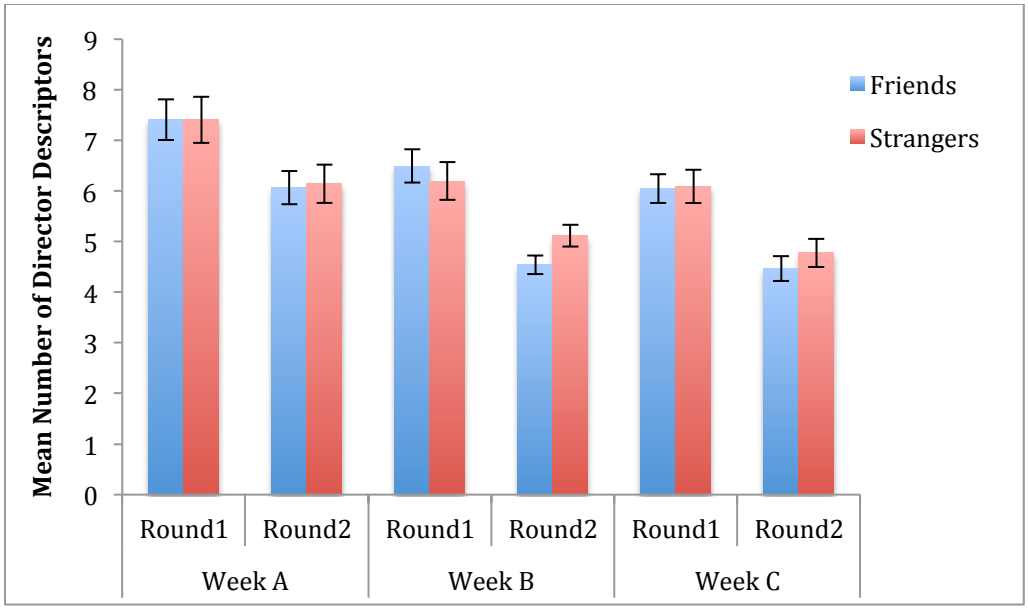


Figure 2. Mean number of Director Descriptors across all three weeks of the study.

Unlike Director Descriptors, there was a main effect for Acquaintanceship $F(1, 51) = 10.38$, $p < .002$, $\eta_p^2 = .17$, with friend dyads ($M = 2.21$, $SD = 1.20$) using more lines but shorter descriptions on each line than strangers ($M = 1.67$, $SD = 0.73$). In other words, friends seemed to perform slower than strangers (see Figure 3).

There was a main effect of Week, $F(2, 51) = 11.07$, $p < .001$, $\eta_p^2 = .18$. Pairwise comparisons indicate participants used fewer lines in Week B ($M = 1.82$, $SD = 0.84$) than in Week A ($M = 2.32$, $SD = 1.36$), $t(51) = 3.52$, $p < .003$, 95% CI [0.14, .79], Cohen's $d_z = 0.48$. There was no reduction in lines used from Week B to Week C ($M = 1.78$, $SD = 0.79$), $t(51) = 0.62$, $p = .90$. Repeating the task for the third time did not result in improved performance beyond the improvement already gained from doing the task twice.

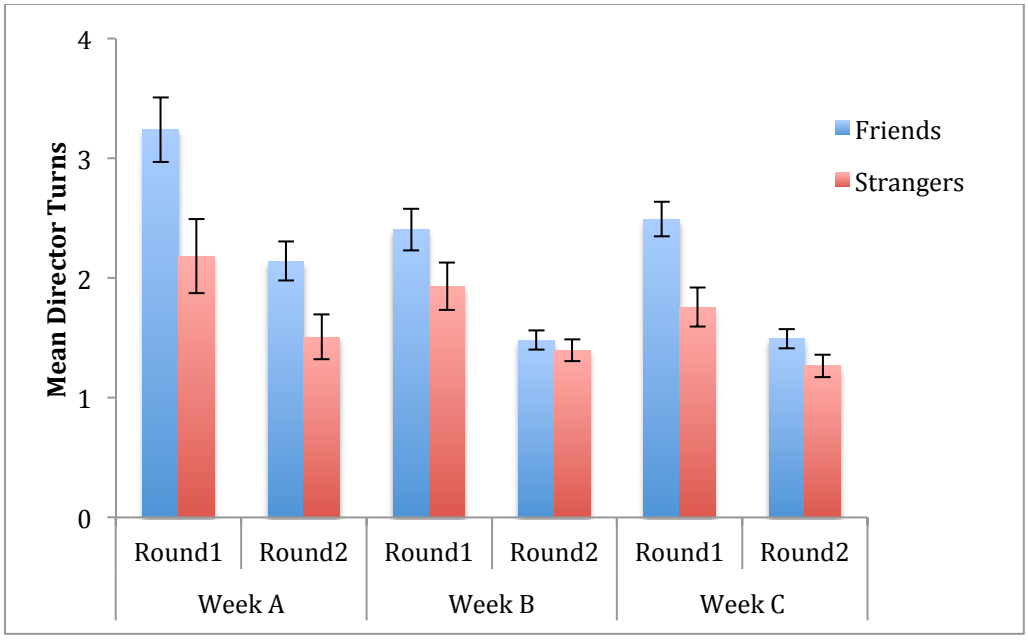


Figure 3. Mean number of Director Turns across the three weeks broken up by Round.

There was also a marginally significant interaction between Week and Acquaintanceship, $F(2, 51) = 5.51$, $p = .06$, $\eta_p^2 = .05$. Whereas friends showed improvement over the three weeks, strangers' performance stayed relatively stable. Though they should be interpreted with caution given the omnibus F test did not clear the predetermined alpha of 0.05, post-hoc tests showed that friends used fewer lines from Week A to Week C, $M_{diff} = 0.70$, $t(28) = 3.82$, $p = .001$, 95% CI [0.25, 1.15], Cohen's $d_z = 0.70$, but strangers used the same number of lines $M_{diff} = 0.33$, $t(21) = 1.59$, $p = .31$.

There was a main effect of Round, $F(1, 51) = 71.07$, $p < .0001$, $\eta_p^2 = .58$, with fewer Director Turns in Round 2 ($M = 2.38$, $SD = 1.20$) than Round 1 ($M = 1.57$, $SD = 0.69$), suggesting that it took dyads less time to identify each tangram in the second time than the first round. There was also an interaction between Round and

Acquaintanceship, $F(1, 51) = 5.05$, $p = .02$, $\eta_p^2 = .10$. There was indication that while both friends and strangers speeded up between rounds, the potential improvement in speed was greater for friends than for strangers. There was no interaction between Week and Round, $F(2, 102) = 0.20$, $p = .67$, $\eta_p^2 = .01$.

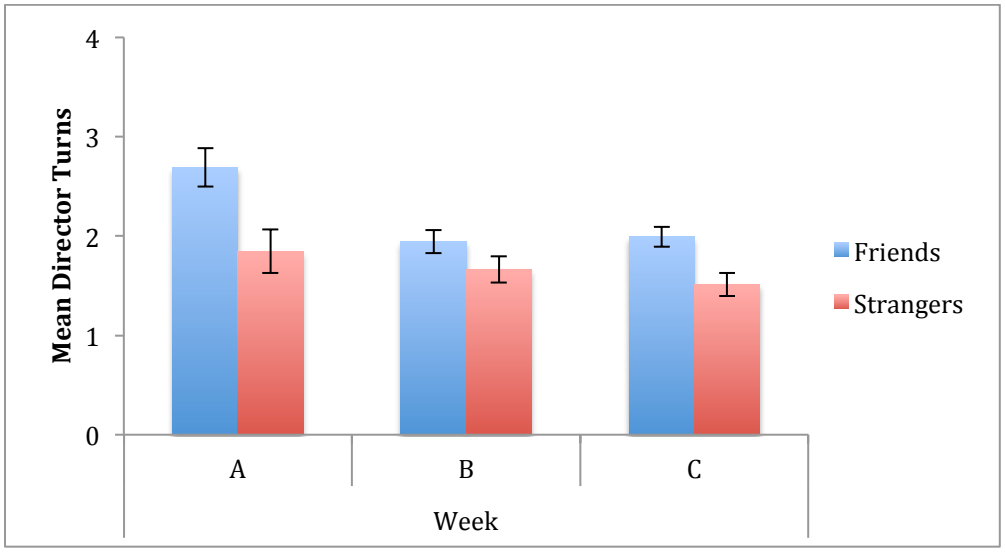


Figure 4. Mean number of Director Turns across the three weeks (Rounds collapsed).

Overall, friend dyads used more turns than stranger dyads. Performance speeded up on the second round on all weeks, which was expected as a robust finding in previous studies. For friends, task familiarity only improved performance on the second week of the task; performance plateaued for the third week. For strangers, task familiarity was less influential, with performance staying similar across the three weeks.

Matcher Descriptors. The range for Matcher Descriptors was more compressed than the Director Descriptors, as Matchers only tended describe tangrams when they were unsure about which one to select. However, there was a main effect

of Acquaintanceship for Matcher Descriptors, $F(1, 51) = 9.15$, $p = .004$, $\eta_p^2 = .15$. In general, friend matchers ($M = 0.97$, $SD = 1.03$) used more descriptors than stranger matchers ($M = 0.58$, $SD = 0.77$) (see Figure 5). Of the 138 stranger rounds included in this analysis, 33% of them included no matcher descriptors whatsoever, while only 35 of the 180 friend rounds (or 19%) included no matcher descriptors. A breakdown of the average number of Matcher Descriptors per trial shows that friends were more willing to at least either repeat or elaborate on descriptions throughout the entire study, while strangers were far more likely to only acknowledge finding the tangram (see Figures 6 and 7). There were no interactions between Acquaintanceship and either Week or Round.

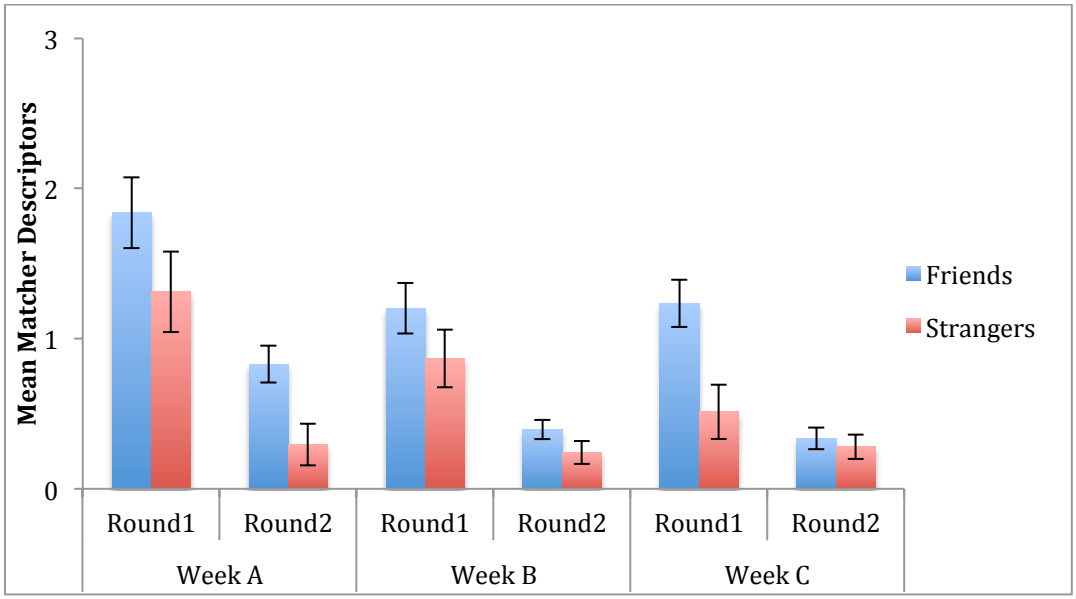


Figure 5. Number of Matcher Descriptors across the three weeks.

There was a main effect of Week, $F(2, 102) = 11.84$, $p < .0001$, $\eta_p^2 = .19$. Again, while there was a difference from Week A to Week B ($M_{diff} = 0.39$, $t(51) =$

3.35, $p = .005$, 95% CI [0.10, 0.68], Cohen's $d_z = 0.47$. There was no difference between Week B and Week C, ($M_{diff} = 0.09$, $t(51) = 0.88$, $p = .77$, 95% CI [-0.16, 0.33].

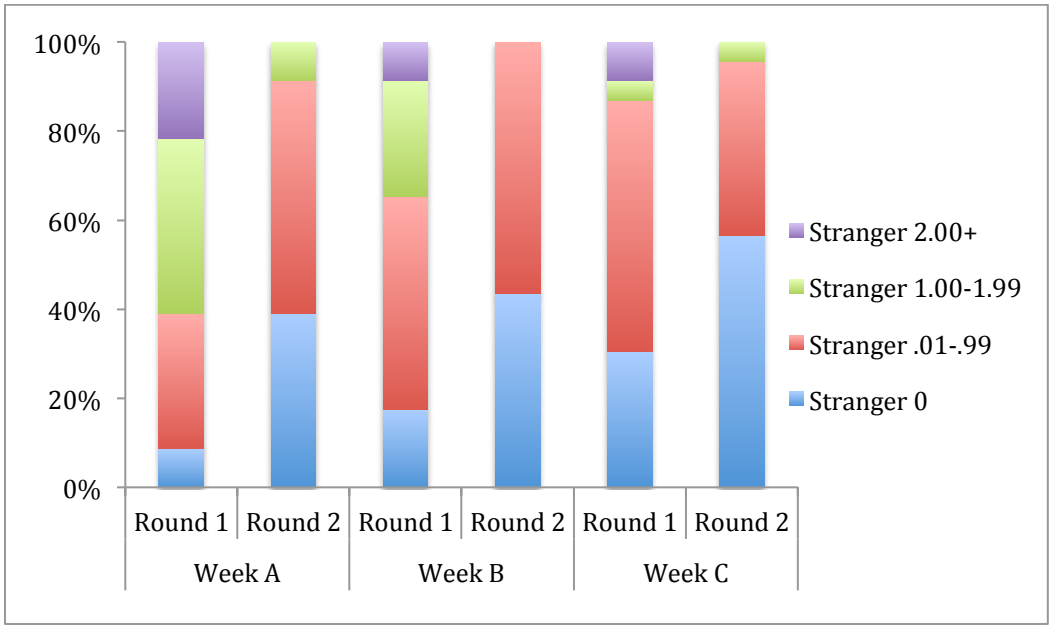


Figure 6. Breakdown of the number of Matcher Descriptors for all Stranger dyads.

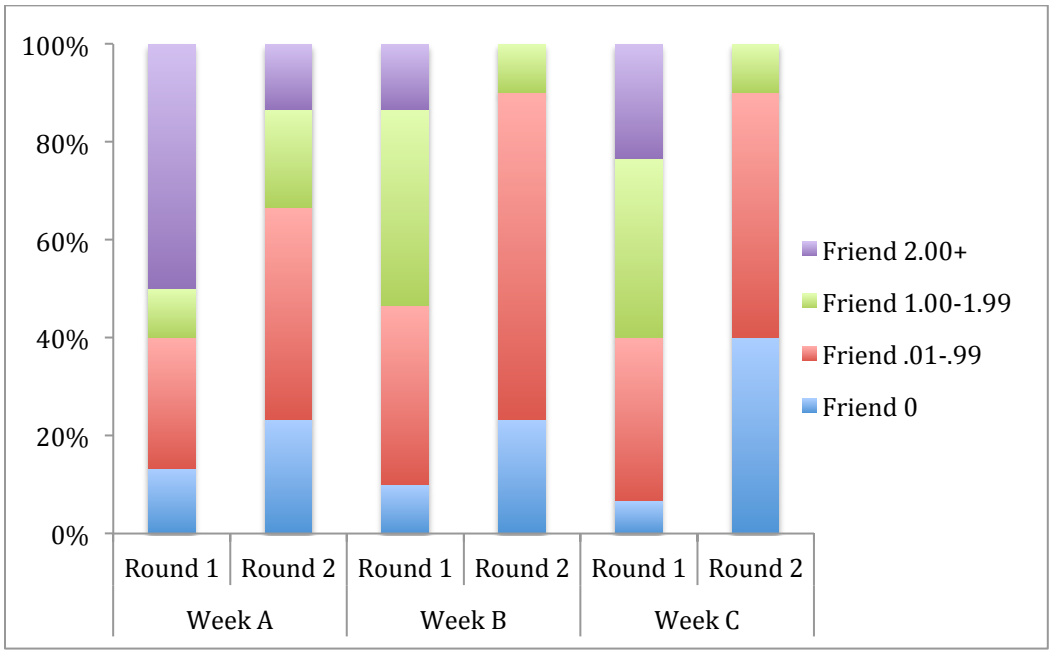


Figure 7. Breakdown number of Matcher Descriptors for all Friend dyads.

There was a main effect of Round. Directors used fewer Descriptors in Round 2 (M = 5.16, SD = 1.56) than in Round 1 (M = 6.12, SD = 1.92), $F(1, 51) = 56.90$, $p < .0001$, $\eta_p^2 = .53$. There was an interaction between Week and Round, $F(2,102) = 3.55$, $p = .03$, $\eta_p^2 = .07$, which suggests that the reduction from Round 1 to Round 2 was not the same across the three weeks. Post-hoc pairwise comparisons indicate that there was significant reduction from Round 1 to Round 2 across all three weeks, but that, numerically, the reduction in Week A (M_{diff} = 1.01, SD = 1.31) was greater than the reduction in Weeks B (M_{diff} = 0.73, SD = 0.83) or Week B (M_{diff} = 0.61, SD = 0.88).

Matcher Turns. Unlike inconsistent relationship between Director Descriptors and Director Turns, Matcher Descriptors and Matcher Turns are very strongly correlated (between r = .50 and .83). Statistical analyses for this dependent variable are likely to be highly redundant. Therefore, they were not carried out for this dependent variable. Figures illustrating the breakdown of this dependent variable can be found in Appendix A (Figures A1 and A2).

Experiment 1 Discussion

Overall, Directors and Matchers showed very similar patterns of behavior in terms of the number of descriptors used. Matchers were less verbose than Directors, choosing only to describe tangrams when they were unsure. Similar to the Directors, Matchers decreased in the number of descriptors used from Week A to Week B, but not Week B to Week C. And, as expected, Round 2 was also faster than Round 1.

Acquaintanceship had a varying effect: friend Matchers used more descriptors than stranger Matchers, possibly indicating greater comfort or willingness to question or contradict Directors when they were unsure. Directors in friend dyads, however, made more contributions in the form of turns than those in stranger dyads. Those who are more experienced with IM in general tend to use shorter, more frequent turns as a way of maintaining more effective grounding (Fox Tree et al., 2011); this may be a similar effect where friends worry more about appearing responsive to their interlocutor and so take more turns that have fewer descriptors on each line than strangers. There were also interaction effects with Acquaintanceship for Round and for Week, indicating that friends got faster from Round 1 to Round 2 and across the three weeks, while strangers remained stable. In general, friends seemed to be more verbose at first, though by the end, the routine of doing the task repeatedly made them appear more like the stranger pairs.

The corpus collected for Experiment 1 provides naturalistic stimuli for Experiments 2 and 3, which focus on the perception of politeness by third-party over-readers. Since the corpus features dialogues between friend and stranger dyads over the course of three weeks performing one task that becomes practiced (tangram) and one novel task (a puzzle), it provides a way to examine whether the perception of politeness differs between friends and strangers and whether any difference persists over time. It also allows a comparison between a task that lends itself to nearly scripted dialogue and a task that needs to be sorted out anew each time.

Experiment 2a: Unedited Tangram Excerpts

Introduction

Given the structure, demands and time-pressured nature of corpus collected in Experiment 1, relatively few typical politeness markers were used throughout the corpus: for example, there were relatively few typical politeness markers (such as "please" and "thank you", or even many examples of indirect speech) actually exist in the tangram portion of the task. In absence of the most straightforward examples of politeness, Experiment 2 uses subjective politeness judgments from third-party over-readers.

This experiment addresses two questions using excerpts from the Experiment 1 corpus. First, can third-party over-readers distinguish between the conversations between friends and strangers in online, text-based conversations? Second, does politeness differ between friends and strangers, or over the three weeks?

Methodology

Stimuli. Dialogue excerpts were taken from twelve corpus dyads – six friend and six stranger dyads – who had relatively straightforward back-and-forth exchanges during the first round of Weeks A, B and C (i.e., did not backtrack or experience inordinate difficulty in identifying tangrams). All twelve dyads contributed three excerpts each. In order to minimize the inherent differences found in naturalistically-produced speech, dialogues from the same tangram trials were taken from every week (e.g., every Week A tested the tangrams #6, #7, and #8 of that week's array). Due to the structure of the task that produced the corpus, tangrams changed between Weeks

A, B and C. To preserve as much as the dialogue as possible, the text of the stimuli were only minimally edited for clarity, including spelling error correction. None of the original excerpts included direct or indirect reference to whether the interlocutors knew each other. Because it would substantially change the dialogue, naturally occurring variation such as number of turns and descriptors were kept.

The dialogue was reformatted to be easier for people to read. In all excerpts, Directors were labeled Participant 1 and their dialogue was presented in dark green font; Matchers were labeled Participant 2 and their dialogue was presented in blue font. The words *Participant 1* and *Participant 2* were presented in bold while the timestamp was italicized. (To minimize confusion, they will henceforth be referred to as *Speaker 1 and Speaker 2*; calling them "participants" within the experiment was to minimize MTurk participants assuming that they were reading transcripts of verbal conversation). Dialogue itself was in regular, non-bolded, non-italicized font. Timestamps showing hour, minute and second were retained from the original transcripts. Dates were not shown in order to not cue participants into the longitudinal structure of the original corpus collection. It should also be noted that all the friend dyads used in Experiments 2 and 3 reported having known each other for at least a year and talking at least several times a week.

Procedure. The experiment was put on PsychSurveys.org. Participants were recruited via Mechanical Turk for $0.95 each. MTurk participants were shown all 36 excerpts (3 excerpts from 12 dyads) in random order. They were told that the stimuli were taken from real dialogues from participants of a different experiment; no

mention was made of the fact that the dialogues came from three different weeks nor that the participants did the task more than once.

Four dependent variables were measured: Speaker 1 Politeness (S1Polite), Speaker 2 Politeness (S2Politeness), Relationship Intensity (Intensity) and Offline/Online Relationship. Participants were shown an excerpt on the same page that was used to ask questions about that particular excerpt. They were asked to rate S1Polite and S2Polite separately on a 7-point scale (1=*Extremely Impolite*, 4=*Neither Polite nor Impolite*, 7=*Extremely Polite*). Participants were also asked to guess how often the interlocutors in the dialogue spoke and how close they are (Intensity) on a 5-point ordinal scale: 1 = *Speakers have NEVER talked before*, 2 = *Speakers have spoken ONCE OR TWICE*, 3 = *Speakers have talked MORE THAN A COUPLE OF TIMES before this dialogue but do not talk regularly,* 4 = *They have talked REGULARLY but are not close friends*, and 5 = *Speakers talk REGULARLY and are close friends.*

Participants were asked whether the speakers primarily communicated in person or online, though this variable was not analyzed in this dissertation).

Results

Sixty-three MTurk participants rated at least 35 of the 36 excerpts; three participants accidentally skipped one trial (though not the same trial). These 63 participants were included in the following analyses.

Can third-party over-readers distinguish between friends and strangers over time in online, text-based conversations? Previous research indicates that third-party

overhearers are ineffective at telling the difference between the conversations friends and strangers in overheard verbal conversations when the conversation is devoid of references to a pre-existing relationship or mutually-shared mutual plans (Planalp, 1993; Planalp & Benson, 1992). Figure 8 shows the raw Relationship Intensity ratings given for friend and stranger dyads over the three weeks: as weeks went along, participants gave more ratings of *Never Talked* and fewer ratings of *Regularly Talked (Close Friends)* and *Regularly Talked (Not Close Friends)*, regardless of whether they were listening to friends or strangers.

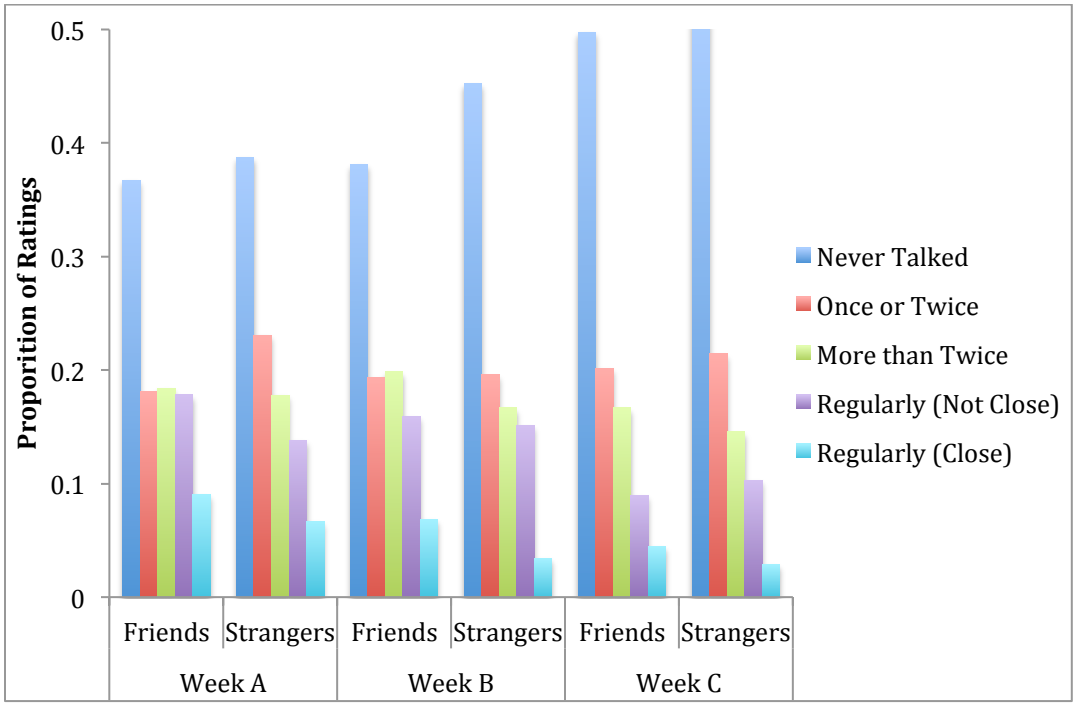


Figure 8. Proportion of Relationship Intensity ratings given to all excerpts for Experiment 2a (Tangram), broken down by actual Acquaintanceship and Week.

A binary logistic regression was run in order to determine the effects of (actual) Acquaintanceship and Week could have on naïve over-readers' ability to correctly classify dyads as strangers, or, more specifically, people who have never

spoken to each other outside of the experiment. A binary logistic regression was used because the ordinal dependent variable, Relationship Intensity, was dichotomized into "strangers" and "friends." This was done because the correct identification for dyads changed from Week A to Week B. Even though speaking once or twice is not what anyone would count as evidence of "friendship", it is technically incorrect to say that the stranger pairs in Week A have spoken "once or twice" before. Stranger dyads had technically spoken to each other once or twice but only in the context of the experiment in Week B and Week C, but not in Week A. So, for simplicity's sake, pairs were either dichotomously labeled as "Strangers" or "Friends" and this variable is referred to as Guessed Acquaintanceship. See Table 1 for how responses were categorized for Week A vs. Week B/C.

	Week A	**Week B**	**Week C**
Never	Strangers	Strangers	Strangers
Once or twice	Friends	Strangers	Strangers
More than a couple	Friends	Friends	Friends
Regularly/not close friends	Friends	Friends	Friends
Regularly/close friends	Friends	Friends	Friends

Table 1. Dichotomization of the Relationship Intensity variable into the Guessed Acquaintanceship variable during each Week.

The results of the binary logistic regression indicated that a model that included both actual Acquaintanceship and Week was predictive of naïve participants Guessed Acquaintanceship, $\chi^2(3) = 187.59$, $p < .001$. This model could account for about 10.2% (Nagelkerke R^2) of the variance in Guessed Acquaintanceship. With these two variables, the model could classify participant guesses 77.9% for actual stranger dyads but only 47.9% of actual friend dyads. The model that contained

Acquaintanceship and Week as predictors worked better for identifying strangers than identifying friends.

Actual Acquaintanceship alone had a small but significant effect on naïve participants' Guessed Acquaintanceship, *Wald* = 3.93, $p = .05$, *Exp(B)* = 1.19, *95% CI* [1.00, 1.42]. This indicates that the odds of a pair of strangers being accurately guessed as being Strangers were 1.2x greater than being guessed as Friends, which is only a small advantage.

Week, however, was more influential in naïve participants' Guessed Acquaintanceship. Comparing Week A to Week B, naïve participants were less than half as likely to correctly identify a friend or stranger dyad in Week B than in Week A, *Wald* = 82.38, $p < .001$, *Exp(B)* = 0.38, *95% CI* [0.31, 0.47]. Comparing Week A to Week C, participants were less than a third as likely to correctly identify dyads, *Wald* = 163.79, $p < .001$, *Exp(B)* = 0.25, *95% CI* [0.20, 0.30]. In other words, the longer the dyad had been participating, the less easily identifiable they were as friends or strangers.

In summary, the increasingly goal-oriented, increasingly formulaic exchanges that developed over the three weeks (Director asks for tangram, Matcher either acknowledges or asks for clarification) made identification of friend and stranger dyads difficult for third-party over-readers. Actual Acquaintanceship had relatively little influence on these ratings, suggesting that over-readers cannot actually reliably distinguish between friend and stranger conversations.

Does politeness differ between friends and strangers, or over time? A preliminary examination of the data by dyad showed that, of the dyads used in the stimuli, strangers followed a fairly straightforward downward trend in average Speaker 1 Politeness rating (see Figure 9). There was some deviation but all stranger dyads decreased in politeness from Week A to Week B; politeness for Week C varied, with some groups becoming more polite and others becoming less polite. Friends, on the other hand, started with a slightly wider range in Week A and remained unpredictable in Weeks B and C. This suggests that friend and stranger dyads may differ from each other across time and that individual dyad may account for some random variance. Thus, a repeated measures mixed effects model was used to analyze the politeness ratings.

The mixed effects model analyzed the two major politeness outcome variables (S1Polite, S2Polite) separately to assess the effects of actual acquaintanceship, perceived relationship intensity and MTurk participant (Participant) and corpus dyad (Dyad) were entered as random effects. Study week (Week), perceived intensity of acquaintanceship (Intensity) and actual acquaintanceship (Acquaint, a binary categorical variable) were entered as fixed effects. An interaction between Acquaintanceship and Week was also entered. Trial (i.e., excerpt) was entered as a repeated measure. Restricted Maximum Likelihood was the estimation method used.

Speaker 1 Politeness (S1Polite). Using chi-square likelihood ratio test (Locker, et al, 2007), it was shown that a model for S1Polite that used random effects was a significant improvement upon an empty model, $X^2(2) = 1642.21$, $p < .001$. The

final model with fixed effects included was an improvement on the random effects-only model, $X^2(3) = 155.30$, $p < .001$.

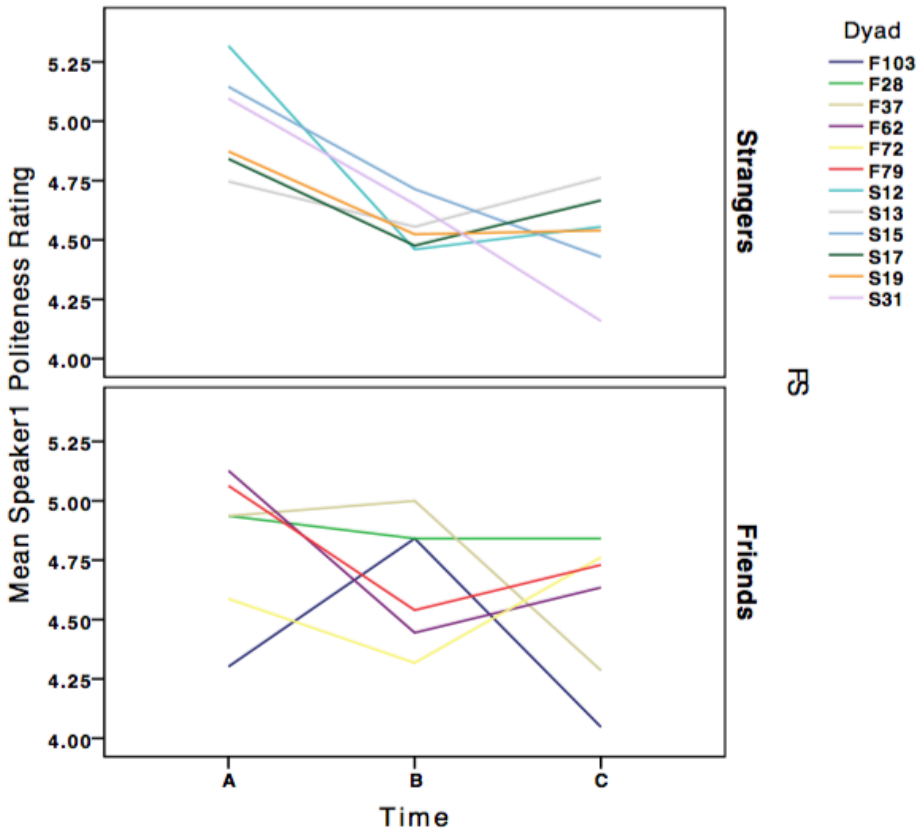


Figure 9. Mean politeness rating for Speaker 1 in each dyad across time broken up by Acquaintanceship.

The mixed effects analysis shows an effect of Week ($F(1, 1541) = 55.11$, $p < .0001$), Intensity ($F(1, 2241) = 92.95$, $p < .0001$) and Acquaintanceship ($F(1, 2114) = 7.88$, $p = .005$), as well as the interaction between Week and Acquaintanceship ($F(1, 1510) = 7.40$, $p = .007$), were all significant predictors of Speaker1 Politeness. The longer participants were in the study (Week), the less polite Speaker1 was ($\beta = -0.10$, $SE = 03$, $p = .001$, 95% CI [-0.15, -0.04]).

Those in dyads that had higher Intensity ratings were rated as more polite than those with lower Intensity ratings ($\beta = 0.17$, $SE = 0.02$, $p < .001$, 95% CI [0.14, 0.21]). There was also a small but significant difference between friend and stranger dyads that also contributed to the variance accounted for by this overall model, $t(2114.30) = 2.81$, $\beta = 0.15$, $SE = 0.05$, $p = .005$, 95% CI [0.14, 0.21], with stranger dyads rated as being more polite than friends. This however was not a difference that held across the three weeks, as shown by a significant interaction between Acquaintanceship and Week, $\beta = -0.11$, $SE = 0.04$, $p = .007$, 95% CI [-0.19, -0.03]. Speaker 1 in stranger dyads was only more polite than friends during Week A (see Figure 10). By Week B, Speaker 1 Politeness ratings for the stranger dyads were indistinguishable from the friend dyads.

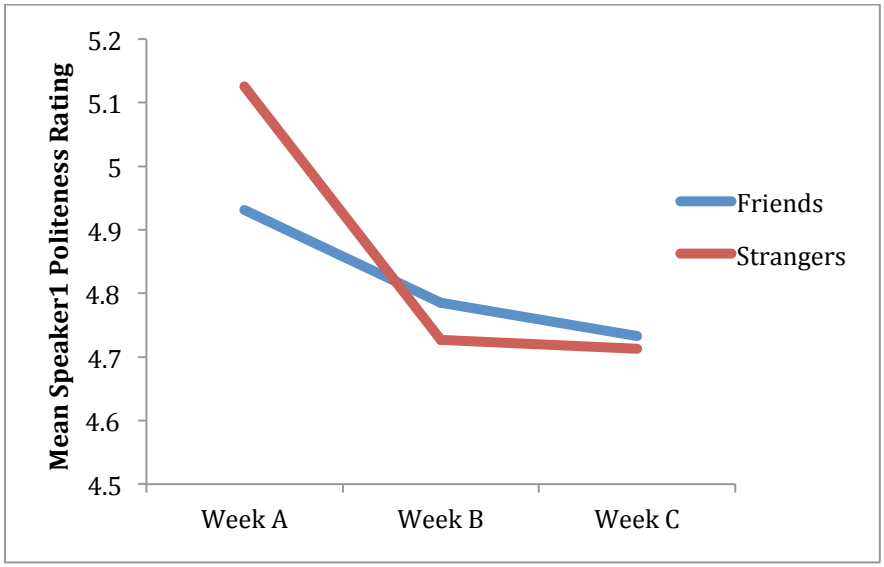


Figure 10. Mean politeness rating for Speaker 1 for friend and stranger dyads across the three weeks.

The finding that strangers are more polite than friends, even if it is a short-lived difference, is somewhat paradoxical considering that participants

simultaneously rated dyads that they guessed were closer (higher Intensity ratings) as being more polite (see Figure 11) but rated actual friend dyads as being less polite than strangers. Strangers become nearly indistinguishable from Friends in terms of S1Polite by Week B. It is possible that this is because people may generally believe that they have reason to be polite to people they know but not people they do not know: a continuing positive relationship requires the maintenance of civility. However, increased familiarity with someone may broaden the definition of what is polite or impolite behavior: calling what a friend is wearing "ugly", for instance, can be taken as a joke (or at least a good-natured jab), whereas strangers would need multiple cues (such as tacking on a "lol") to convey that they do not mean offense.

Finally, the random variance in S1Polite that can be attributed to MTurk Participant is 0.82 (*SE* = 0.15), while almost no variance can be attributed to the effect of Dyad (0.007, *SE* = 0.01); this suggests that, for tangram task excerpts, there are no systematic difference in Speaker 1 Politeness that can be attributed to differences between dyads (though there is variance attributable to individual participants).

As an estimate of effect size, the proportion of variance in S1Polite explained by the final model (over the empty model) is approximately 60.20% (Carson & Beeson, 2013): the proportion of variance explained by the predictors over and above the random effects is approximately 8.5%.

Speaker 2 Politeness (S2Polite). The same analyses used for S1Polite was used on S2Polite to examine the politeness rating data for Speaker 2. Using chi-

square likelihood ratio test (Locker, et al, 2007), it was shown that a model for S1Polite that used random effects was a significant improvement upon an empty model, $X^2(2) = 1708.13$, $p < .001$. The final model with fixed effects included was an improvement on the random effects-only model, $X^2(3) = 195.54$, $p < .001$.

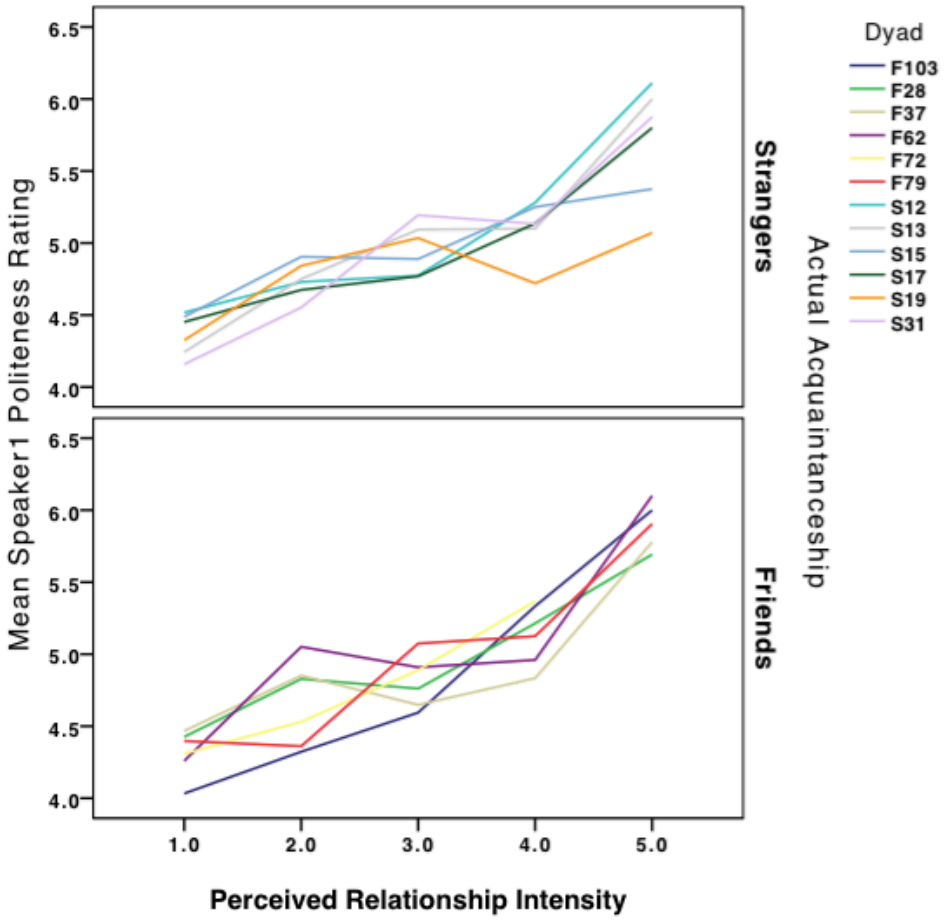


Figure 11. Relationship between Speaker 1 Politeness Rating and perceived Relationship Intensity in dyads, split up by actual Acquaintanceship status.

Week ($F(1, 1537) = 40.65$, $p < .0001$), Intensity ($F(1, 2249) = 136.50$, $p < .0001$) and Acquaintanceship ($F(1, 2004) = 19.86$, $p < .0001$), as well as the interaction between Week and Acquaintanceship ($F(1, 1508) = 19.47$, $p < .0001$),

were all significant predictors of Speaker 2 Politeness. The longer participants were in the study (Week), the less polite Speaker 2 was (β= -0.04, SE = 0.03, p < .001, 95% CI [-0. 09, 0.01]). Those in dyads that had higher Intensity ratings were rated as more polite than those with lower Intensity ratings (β = 0.21, SE = 0.02, p < .001, 95% CI [0.17, 0.24]). There was also a significant difference between friend and stranger dyads, with stranger dyads rated as being more polite than strangers, t(2003.85) = 4.46, β = 0.24, SE = 0.05, p < .001, 95% CI [0.13, 0.34]. Again, the major difference between friends and strangers was seen in the first week, where friends were significantly less polite than strangers (p = .0001) but friend and stranger dyads were indistinguishable by the second week (see Figure 12).

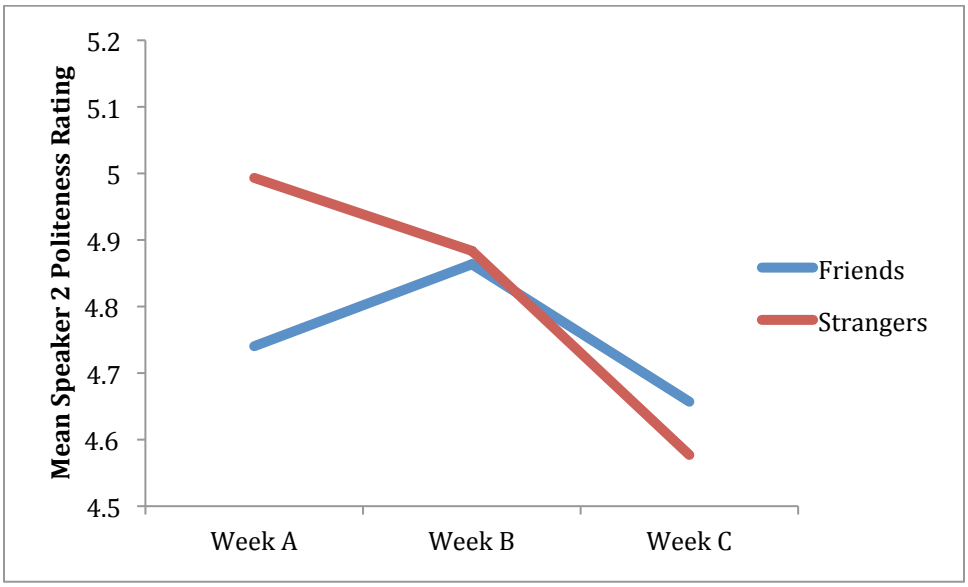


Figure 12. Speaker 2 Politeness for Experiment 2a over the three weeks.

Again, there was random variance that is attributable to MTurk Participant (β = 0.83, SE = 0.15) while almost no variance can be attributed to the effect of Dyad (β = 0.05, SE = 0.01). For these excerpts, there are no systematic difference in Speaker 2

Politeness that can be attributed to differences between dyads (though there is some attributable to the rating habits of individual participants).

As an estimate of effect size, the proportion of variance in S1Polite explained by the final model (over the empty model) is approximately 64.44% (Carson & Beeson, 2013); the proportion of variance explained by the predictors over and above the random effects is approximately 10.07%.

The decrease in politeness seen over the three weeks may be attributable to increased familiarity with the task, which translates into speaking in an abrupt shorthand that may seem less polite by people outside of the conversation. The next experiment aims to examine what happens to politeness ratings when tasks are always novel.

Experiment 2b: Unedited Puzzle Excerpts

Introduction

Except for the stimuli, this experiment was identical to Experiment 2a. Instead of using tangram task stimuli, which was the same task repeated over the three weeks, I used dialogue taken from the puzzle tasks. Whereas Experiment 2a's stimuli from Weeks B and C showed friends and strangers communicating when they had become acclimated to the task, Experiment 2b examines whether perceived politeness from the exact same dyads holds when confronted with totally new tasks for all three weeks. The dialogue for the tangram task becomes very routinized across the three weeks: as seen in Experiment 1, length and number of turns shortens with time, which may make later excerpts sound more curt and less polite than earlier excerpts.

However, this may not be the case with the puzzle tasks, as each weekly puzzle has different task demands and does not assign roles to participants.[1] Experiment 2b uses this less routinized (but still task-oriented) dialogue to ask the same questions that were asked in Experiment 2a: Can third-party over-readers distinguish between the conversations between friends and strangers? And does politeness differ between friends and strangers, or over the three weeks?

Methodology

Stimuli. Dialogue excerpts were taken from the same twelve corpus dyads used in Experiment 2a. Again, all twelve dyads contributed three excerpts, one from each week. Because participants did not solve the puzzle items in a consistent order (unlike tangrams where the vast majority of participants went from left to right, top to bottom), three minutes of worth of dialogue were excerpted from the middle of each puzzle task. Dialogue was taken between minutes 3 and 10: in other words, the dialogue was taken from the conversation after dyads had time to sit with the task and possibly figure out a few answers but before the experimenters gave pairs the option of quitting (at 10 minutes).

Again, the text of the stimuli was only minimally edited for clarity, including spelling error correction. None of the original excerpts included direct or indirect reference to whether the interlocutors knew each other. Because it would substantially change the dialogue, naturally occurring variation such as number of turns were kept.

[1] It should be noted that, on the whole, puzzles elicited fewer words from Experiment 1 dyad participants than tangram tasks. The average number of words elicited in the three weeks of the tangram task was 4142.86 (*SD* = 1323.32) while the average number of words for the puzzle task was 2931.86 (*SD* = 904.54). Table A1 (Appendix A) shows the averages for number of words used for each session's tangram and puzzle tasks.

Participants were labeled “Participant 1” (in green font) and “Participant 2” (in blue font). The words *Participant 1* and *Participant 2* were presented in bold while the timestamp was italicized. Dialogue itself was in regular, non-bolded, non-italicized font. Timestamps showing hour, minute and second were retained from the original transcripts. Dates were not shown in order to not cue participants into the longitudinal structure of the original corpus collection.

Procedure. The Experiment 2b procedure was identical to Experiment 2a.

Results

Sixty MTurk participants rated at least 35 of the 36 excerpts. Three participants skipped one trial.

Can third-party over-readers distinguish between friends and strangers in online, text-based conversations over time? Similar to Experiment 2a, a binary logistic regression was run in order to determine the effects of actual Acquaintanceship and Week could have on naïve over-hearers’ ability to correctly classify dyads as strangers. Experiment 2a featured almost formulaic conversations that were generated using the easily routinized tangram task. Dialogues for Experiment 2b, however, were generated using the puzzle tasks, which allowed for less predictable dialogue.

A quick glance at just the frequencies of Relationship Intensity ratings indicate that there were some differences in the way third-party over-readers rated these more free-form dialogues from the more formulaic ones (see Figure 13). While the increasingly formulaic dialogues of the tangram task were also increasingly rated

as taking place between people who did not know each other, there is a slightly more even distribution of ratings for these less formulaic puzzle dialogues. There was roughly the same number of *Never Talked* ratings across the three weeks and across friends and strangers. Except for a spike to 42% of participants choosing *Never Talked* for the actual stranger dyads during Week C, *Never Talked* ratings stayed around 37-38% for both friends and strangers. The *Never Talked* rating for the tangram task, by comparison, steadily increased from 37% to 50%.

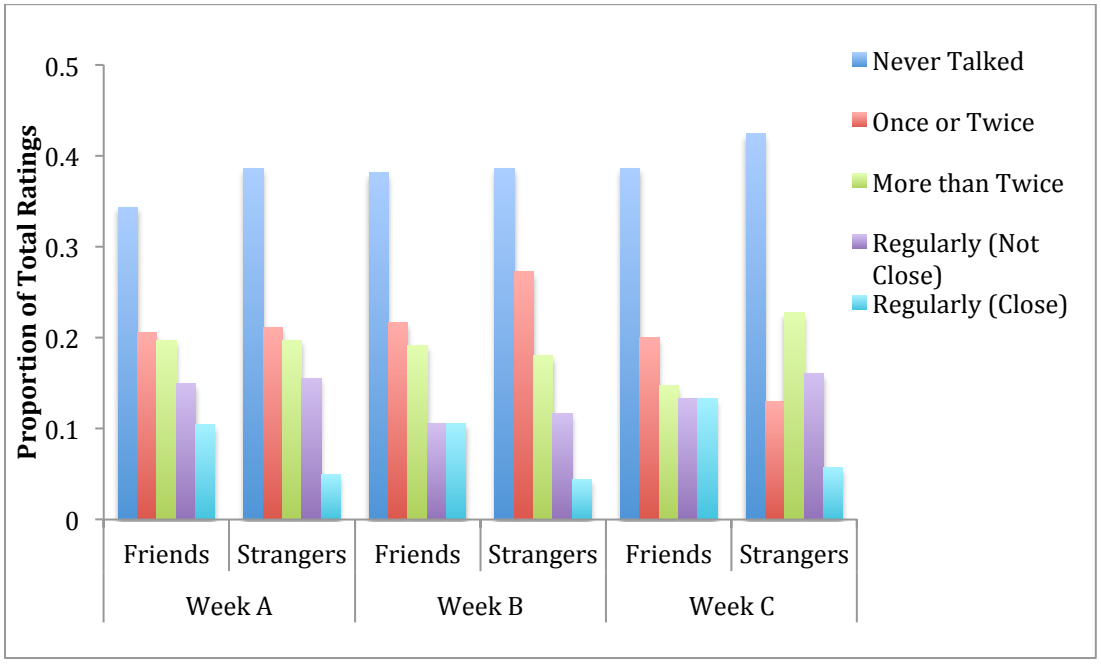


Figure 13. Proportion of Relationship Intensity ratings given to all excerpts for Experiment 2b (Puzzle), broken down by actual Acquaintanceship and Week.

Additionally, the number of *Talks Regularly (Close Friends)* ratings for the puzzle task differed in all three weeks as well, with more *Close Friends* ratings for actual friend dyads and fewer for stranger dyads. In comparison, the number of

participants who chose *Talks Regularly (Close Friends)* in the tangram task decreases steadily for both friend and stranger dyads in the tangram task (Experiment 2a).

Again, a binary logistic regression was run, with the Relationship Intensity variable dichotomized into "strangers" and "friends" depending on whether participants were rating Week A or Week B/C (see Table 2 for dichotomization schema); this dichotomized variable is referred to as Guessed Acquaintanceship.

The results of the binary logistic regression indicated that a model that included both actual Acquaintanceship and Week was predictive of naïve participants Guessed Acquaintanceship, $\chi^2(3) = 127.13$, $p < .001$. This model could account for about 7.6% (Nagelkerke R^2) of the variance in Guessed Acquaintanceship. Similar to the results of Experiment 2a, the model could accurately classify participant guesses 77.2% for actual stranger dyads but only 45.1% of actual friend dyads. This suggests that Acquaintanceship and Week worked better for identifying strangers than identifying friends.

However, even though the initial glance at the rating frequencies indicated that Acquaintanceship made more of a difference in Experiment 2b than in Experiment 2a, once these ratings were dichotomized into Guessed Acquaintanceship, actual Acquaintanceship in this model was not shown to be significantly predictive, *Wald* = 2.41, $p = .12$, *Exp(B)* = 1.15, *95% CI* [0.97, 1.37].

Week, again, was more influential in naïve participants' Guessed Acquaintanceship. Comparing Week A to Week B, naïve participants were a third as likely to correctly identify a friend or stranger dyad in Week B than in Week A, *Wald*

= 105.12, $p < .001$, *Exp(B)* = 0.33, *95% CI* [0.26, 0.40]. Comparing Week A to Week C, participants were a little more than a third as likely to correctly identify dyads, *Wald* = 72.74, $p < .001$, *Exp(B)* = 0.40, *95% CI* [0.32, 0.49]. Again, it was easiest to correctly identify dyads as friends or strangers in Week A than in Week B or Week C. Dyads became harder to distinguish over time, even though the puzzle dialogues initially seemed more likely to produce more idiosyncratic conversations.

Does politeness differ between friends and strangers, or over time? Repeated-measures mixed effects modeling was used to analyze the politeness ratings. A repeated measures mixed effects model was run to analyze the two major politeness outcome variables (S1Polite, S2Polite) separately to assess the effects of actual acquaintanceship, perceived relationship intensity and MTurk participant (Participant) and corpus dyad (Dyad) were entered as random effects. Study week (Week), perceived intensity of acquaintanceship (Intensity) and actual acquaintanceship (Acquaintanceship, a binary categorical variable) were entered as fixed effects. An interaction between Acquaintanceship and Week was also entered. Trial (i.e., excerpt) was entered as a repeated measure. Restricted Maximum Likelihood was the estimation method used.

Speaker 1 Politeness (S1Polite). Using chi-square likelihood ratio test (Locker, Hoffman, & Bovaird, 2007), it was shown that a model for S1Polite that used random effects was a significant improvement upon an empty model, $X^2(2) = 1080.66$, $p < .001$. The final model with fixed effects included was an improvement on the random effects-only model, $X^2(3) = 87.24$, $p < .001$.

The mixed effects analysis shows an effect of Week ($F(1, 1432) = 57.424$, $p < .0001$), Intensity ($F(1, 2142) = 17.82$, $p < .0001$) and Acquaintanceship ($F(1, 1941) = 30.37$, $p < .0001$) were all significant predictors of Speaker1 Politeness. There was also an interaction between Week and Acquaintanceship ($F(1, 1432) = 4.65$, $p = .03$). The longer participants were in the study (Week), the less polite Speaker1 was ($\beta = -0.11$, $SE = 0.03$, $p < .001$, 95% CI [-0.17, -0.05]). Those in dyads that had received higher Intensity ratings were rated as more polite than those with lower Intensity ratings ($\beta = 0.07$, $SE = 0.02$, $p < .001$, 95% CI [0.04, 0.10]).

There was also a significant difference between friend and stranger dyads, with stranger dyads rated as being more polite than strangers $t(1941.26) = 5.51$, $\beta = 0.30$, $SE = 0.05$, $p < .001$, 95% CI [0.19, 0.40]. Additionally, though friend and stranger dyads were indistinguishable in terms of Speaker 1 Politeness during Week B, friends were rated as less polite than strangers in Weeks A ($p < .0001$) and C ($p < .0001$) (see Figure 14).

Finally, there was random variance in S1Polite that was attributable to individual MTurk Participant ($\beta = 0.53$, $SE = 0.02$) while only a small amount of variance can be attributed to the effect of Dyad ($\beta = 0.04$, $SE = 0.01$). Similar to tangram task excerpts, differences in dyads do not account for much variance in Speaker 1 Politeness while there is an effect of Participant.

As an estimate of effect size, the proportion of variance in S1Polite explained by the final model (over the empty model) is approximately 50.50% (Carson &

Beeson, 2013); the proportion of variance explained by the predictors over and above the random effects is approximately 6.92%.

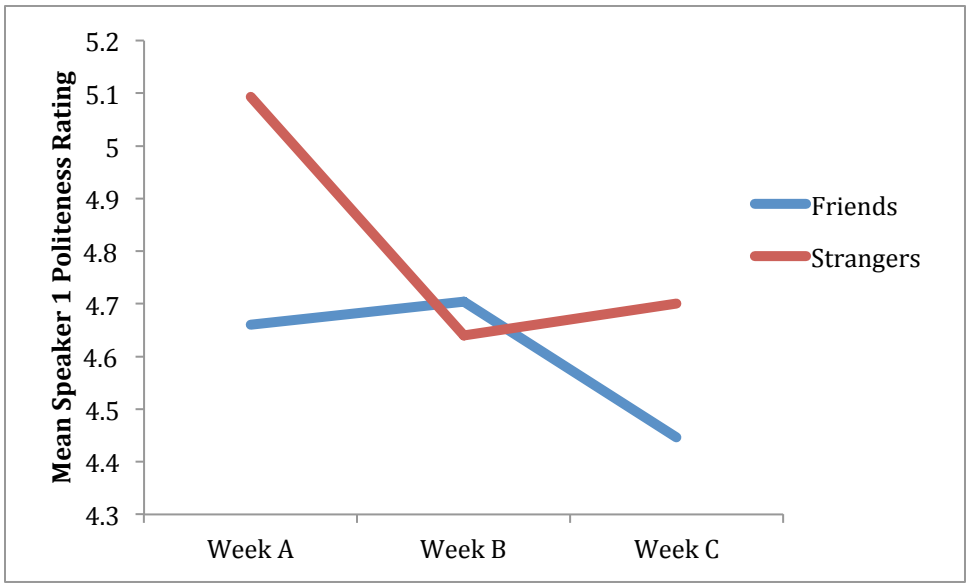


Figure 14. Mean Speaker 1 Politeness rating across the three weeks for the novel puzzle tasks.

Speaker 2 Politeness (S2Polite). The same analyses used for S1Polite was used on S2Polite to examine the politeness rating data for Speaker 2. Using chi-square likelihood ratio test (Locker, et al, 2007), it was shown that a model for S1Polite that used random effects was a significant improvement upon an empty model, $X^2(2) = 1096.00$, $p < .001$. The final model with fixed effects included was an improvement on the random effects-only model, $X^2(3) = 87.24$, $p < .001$.

Week ($F(1, 1435) = 56.13$, $p < .0001$), Intensity ($F(1, 2140) = 8.10$, $p = .004$) and Acquaintanceship ($F(1, 1961) = 24.94$, $p < .0001$) were significant predictors of Speaker 2 Politeness. There was also an interaction between Week and Acquaintanceship ($F(1, 1435) = 7.56$, $p = .006$), were all significant predictors of

S2Polite. The longer participants were in the study (Week), the less polite Speaker 2 was (β = -0.09, *SE* = 0.03, 95% CI [-0.15, -0.04]). Those in dyads that had higher Intensity ratings were rated as more polite than those with lower Intensity ratings (β = 0.04, *SE* = 0.02, 95% CI [0.01, 0.07]).

There was also a significant difference between friend and stranger dyads, with stranger dyads rated as being more polite than strangers, $t(1961.03) = 4.99$, β = 0.26, *SE* = 0.05, $p < .001$, 95% CI [0.16, 0.36]. Post-hoc tests indicated that Speaker 2 Politeness differed in both Week A ($p < .0001$) and Week C ($p < .03$), though the difference was more pronounced in Week A (see Figure 15).

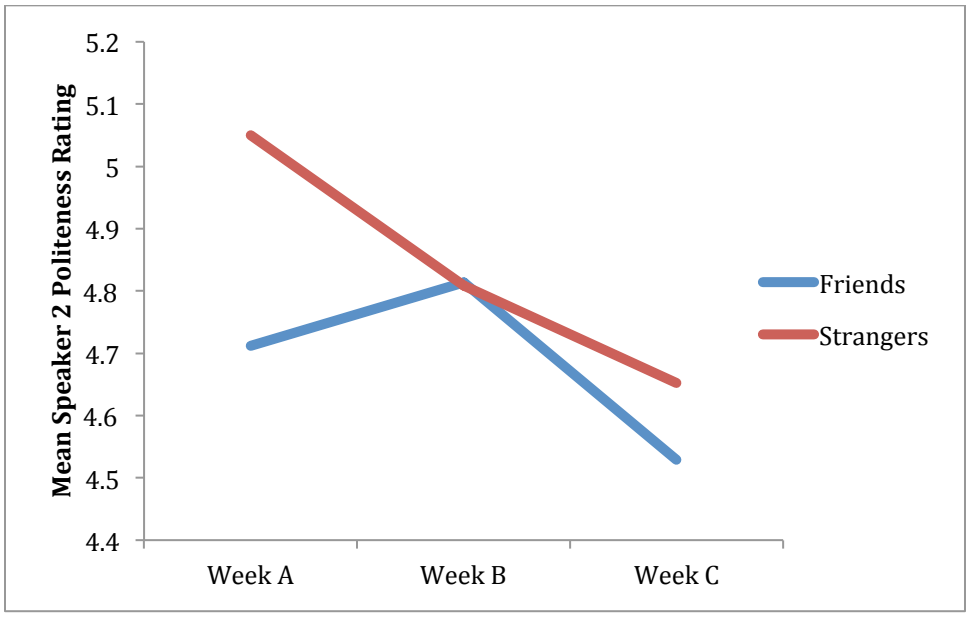


Figure 15. Speaker 2 Politeness during non-repeated puzzle tasks throughout the three weeks.

Finally, some random variance in S2Polite was attributable to Participant (β = 0.49, *SE* = 0.05, while almost no variance can be attributed to the effect of Dyad (β = 0.03, *SE* = 0.01). Again, while variance due to Participant seemed to account for

some variance, the variance due to systematic differences between Dyads was negligible.

As an estimate of effect size, the proportion of variance in S2Polite explained by the final model (over the empty model) is approximately 49.77% (Carson & Beeson, 2013); the proportion of variance explained by the predictors over and above the random effects is approximately 2.49%.

Experiments 2a and 2b Discussion

Friend dyads were either less polite or similarly polite as compared to stranger dyads in both tangram (Exp 2a) and puzzle (Exp 2b) tasks. The tangram task (2a) allowed to participants to develop a rhythm and strategy to their performance, while the novel puzzle tasks (2b) were less likely to generate a routinized script that dyads would follow. In the repeated tangram task, Friend Directors were judged as being slightly less polite (though not outright impolite) when compared to Stranger Directors during the first week. The difference evaporated by the second week. Friend Matchers were also considered less polite than Stranger Matchers the first week but were comparable from the second week onwards. Stranger Matchers likely decreased in politeness (as opposed to plateaued like the Stranger Directors) because Matchers became fairly efficient by the end of the experiment, often only giving very brief responses that merely let the Directors know that they found the target tangram.

Despite the fact that dialogue for the puzzle tasks was not as constrained as the tangram task, the perception of politeness by third-party over-readers was similar across the tangram and the puzzle tasks: friends were perceived as less polite than

strangers at first but then generally became less polite over time, with strangers experiencing a relatively steeper drop than friends. The only major difference is that in the third week of the puzzle task, Speaker 1 in stranger dyads were more polite those in the friend dyads. Nevertheless, this suggests that even 1 to 2 hours of non-personal conversation can be enough to make strangers largely indistinguishable from friends in terms of subjective politeness.

Third-party over-readers were, generally speaking, not good at explicitly identifying whether speakers were friends or strangers, in either tangram or puzzle tasks, regardless of week. This was surprising, given that the friend dyads that participated in the corpus collection were close friends who knew each other for at least a year and talked multiple times a week. Even in a goal-oriented task done through text, where dialogue is naturally constrained and multimodal cues are absent, it still seems remarkable that friends could so easily be mistaken for strangers who have spoken for less than three hours (and vice versa). Participants' guesses as to whether the dyads were friends or strangers were more likely to be influenced by what week (A, B or C) the dialogue took place than whether the dyad knew each other prior to the study.

Experiment 3a: Manipulating the Perception of Politeness in Tangram Task Dialogue

Introduction

Experiment 3a looks at whether the perception of politeness can be manipulated via interlocutor response to a potentially impolite statement by a speaker.

Specifically, one line of director dialogue from an original, unedited excerpt was changed to say something that would be considered colloquially impolite. Brown and Levinson (1987) claimed that politeness depended on the directness of a request: for instance, "Could you find the ice skater tangram?" would be considered less polite than "Find the ice skater tangram." In addition to lacking most of the usual politeness markers, Directors also mostly lacked this specific request format when asking for tangrams. The prescribed demands of the task essentially excused Directors from having to mitigate their requests, or even having to make requests explicitly. The role of the Director was obviously to ask the Matcher to do things and both Director and Matcher understood this. For Experiments 3a and 3b, instead of manipulating the dialogues using the more subtle direct/indirect distinction, Director dialogue was manipulated to be either obviously impolite or neutral in nature and Matcher response was manipulated to be offended or unoffended.

As Brown and Levinson (1987) focus on the politeness of specific, single utterances, their theory would predict that blatantly impolite utterances (such as insults) would always be considered less polite by third party eavesdroppers (or over-readers) than the neutral utterances. Arundale (1999), however, claims that politeness is co-constructed: utterances are not considered impolite until both speakers accept it as not-impolite after it is spoken. This predicts that a Matcher's response to a Director's impolite utterance could mitigate how polite the Director's utterance is perceived to an over-reader. If a mitigating response is inserted after an impolite utterance, then there should be an increase in perceived politeness compared to when

there isn't a mitigating-interlocutor response. Experiment 3 tests Brown and Levinson against Arundale on this point: if interlocutor response can mitigate perceived politeness, it provides evidence for Arundale's theory. If it doesn't, then it provides evidence for Brown and Levinson.

Brown and Levinson allow that those who are closer can get away with being less polite. Arundale does not offer up a neat prediction, choosing instead to simply say that social relationship, history and conversational context also play into the perceived politeness of an utterance. Perceived (by the over-reader) relationship intensity will be examined but no specific predictions are made about this.

Methodology

Stimuli. After examining the results from Experiment 2a, the dialogues were narrowed down to a subset of the eight dialogues: four friend dyads and four stranger dyads from Experiment 2a were used. Only Week B excerpts were used, as they were most likely to be rated as being neither polite nor impolite (all with an average of politeness rating of 4 out of 7, with standard deviation of less than 1). That is to say, they were neutral in tone.

The dialogue for the first two tangrams discussed in each dialogue was left unedited from Experiment 2a (where most dialogue was unedited except for small edits to increase clarity for over-readers). Only the third tangram discussed in each dialogue was edited. Most of the dialogue lines in the "neutral" condition are the original dialogue, though a couple were changed if they seemed as if they could be misconstrued as anything besides neutral.

Each excerpt had four versions created by crossing Speaker 1's contribution (S1Utterance, either Impolite or Neutral) and Speaker 2's (S2Response, either Offended or Not Offended): Impolite-Offended (IO), Impolite-Not Offended (IN), Neutral-Offended (NO), Neutral-Not Offended (NN). Speaker1 was always the one either saying something impolite or neutral (Speaker1 Utterance), while Speaker2 was always the one reacting in an offended or not offended manner (Speaker2 Utterance).

Figure 16. Example of Experiment 3 stimuli manipulation (using a stimulus set used in Experiment 3b)

Base Dialogue
1:13:11 PM Participant1: 12 is moby dick
1:13:36 PM Participant2: and jane eyre
1:13:45 PM Participant1: i know
1:13:51 PM Participant2: 15 is great expectations
1:15:49 PM Participant1: is 2 also Lord Breckienridges?
1:15:59 PM Participant1: *Lord Breckenridges
1:16:03 PM Participant2: i haven't heard of that book tho

Speaker1 Impolite – Speaker2 Offended
1:16:24 PM Participant1: well, great. it's your fault if we fail then.
1:16:53 PM Participant2: wtf it's your fault for starting this late

Speaker1 Impolite – Speaker2 Not Offended
1:16:24 PM Participant1: well, great. it's your fault if we fail then.
1:16:53 PM Participant2: at least we'll be in it together haha XD

Speaker1 Neutral – Speaker2 Offended
1:16:24 PM Participant1: Man I hate that we're going to do so badly! :(
1:16:53 PM Participant2: wtf it's your fault for starting this late

Speaker1 Neutral – Speaker2 Not Offended
1:16:24 PM Participant1: Man I hate that we're going to do so badly! :(
1:16:53 PM Participant2: at least we'll be in together haha XD

The portion of the manipulated dialogue was not pointed out in any way. Though the different dialogue conditions were tested within-subjects, the different version of each excerpt was presented between-subjects so that any given MTurk participant would only see one version of the dialogue. See Figure 16 for an example of how the stimuli were constructed (n.b. this example uses dialogue from Experiment 3b, but the way the stimuli constructed is similar for both tangram and puzzle task components).

Procedure. The experiment was put on PsychSurveys and participant recruitment was done through MTurk. Participants were paid $0.85 for their participation.

A total of 95 participants were run in Experiment 3a. Every participant saw and rated eight dialogues from the eight chosen dyads using the same dependent variables used in Experiments 2a and 2b: Speaker1 Politeness (S1Polite), Speaker2 Politeness (S2Polite) and Relationship Intensity (non-dichotomized). Politeness was rated on a scale of 1 (*Extremely Impolite*) through 7 (*Extremely Polite*), while Relationship Intensity was on a 5 point scale: 1 = *Speakers have NEVER talked before*, 2 = *Speakers have spoken ONCE OR TWICE*, 3 = *Speakers have talked MORE THAN A COUPLE OF TIMES before this dialogue but do not talk regularly,* 4 = *They have talked REGULARLY but are not close friends*, and 5 = *Speakers talk REGULARLY and are close friends.*

Participants were also asked whether speakers knew each other in real life or only online, though this data was not analyzed in this study. Every participant saw one excerpt each IO, IN, NO, or NN from both friend and stranger dyads. Partial

Latin square counterbalancing was used to determine which version (IO, IN, NO, or NN) of each of the eight excerpts the participants would see. The order of presentation was randomized by participant.

The data were analyzed to examine the following questions:

1. As a basic manipulation check, is Speaker 2 rated differently than Speaker 1 (i.e., rated as a separate person)?
2. Is the perceived politeness of Speaker 1's utterance mediated by Speaker 2's reaction and vice versa?
3. Does perceived politeness change depending on how close over-readers think they are?

Results

Are Speaker 1 and Speaker 2 rated as separate entities? To ensure that Speaker 1 and Speaker 2 were rated as separate entities and not as a single unit, dependent samples t-tests were run on all the ratings for the situation-incongruent dialogue conditions, IN (Impolite-Not Offended) and NO (Neutral-Offended). In dialogue condition IN, Speaker 1 ($M = 3.62$, $SD = 1.50$) was rated as significantly less polite than Speaker 2 ($M = 5.29$, $SD = 1.10$), $t(94) = -10.45$, $p < .001$, $d = 1.07$. In NO, Speaker 2 ($M = 3.20$, $SD = 1.56$) was rated as being significantly less polite than Speaker 1 ($M = 4.75$, $SD = 1.15$), $t(94) = -10.37$, $p < .001$, $d = 1.06$.

Does interlocutor influence Speaker 1 Politeness (S1Polite)? To analyze whether manipulating interlocutor response to impolite utterances, a 2x2 (Speaker 1 Utterance x Speaker 2 Response) repeated measures ANOVA was run. This is unlike

the previous experiment, which used mixed effects modeling: Experiment 2 was looking at multiple factors that could be involved in influencing third-party perception of politeness in stimuli that were not systematically manipulated; that dataset also had some missing data, which are better addressed with a mixed effects model than a repeated measures ANOVA. In comparison, Experiment 3 asks a straightforward question about whether a specific experimental manipulation has made a difference to two ratings. To address concerns about random effects due to item or due to subject, both by-subject (F1) and by-item (F2) analyses, as well as minF' (Clark, 1973; Raaijmakers, Schrijnemakers, & Gremmen, 1999), were run.

Overall, there was a significant main effect for S1Utterance on S1Polite in the by-subject analysis, $F_1(1, 94) = 25.87$, $p < .001$, $\eta p^2 = 0.22$, as well as in the by-items analysis, $F_2(1, 28) = 156.65$, $p = .001$, $\eta p^2 = 0.85$; $MinF'(1, 86) = 78.51$, $p < .001$. The experimental manipulation to alter the politeness of Speaker 1's utterance worked: the utterances that were meant to be impolite were perceived as impolite ($M = 3.44$, $SD = 1.21$), while the neutral ones were perceived to be more polite ($M = 4.99$, $SD = 0.74$).

Most importantly, there was a main effect of S2Response on Speaker 1's perceived politeness, $F_1(1, 94) = 157.38$, $p < .001$, $\eta p^2 = 0.63$, as well as in the by-items analysis, $F_2(1, 28) = 13.12$, $p = .001$, $\eta p^2 = 0.32$; $MinF'(1, 59) = 8.71$, $p = .004$. Speaker 2 Response had an effect on the perception of Speaker 1's politeness. Regardless of what Speaker 1 said, the perception of how polite Speaker 1 had been was influenced by Speaker 2 sounding offended ($M = 4.42$, $SD = 0.89$) or not

offended ($M = 4.00$, $SD = 0.91$). Planned comparisons indicate that for impolite utterances, Speaker 1 was considered more polite if Speaker 2 responded in a non-offended manner ($M = 3.62$, $SD = 1.34$) than if Speaker 2 had responded in an offended manner ($M = 3.25$, $SD = 1.31$), $t(94) = 3.24$, $p = .002$, M_{diff} *95% CI* [0.14, 0.58], $d = 0.33$ (see Figure 17).

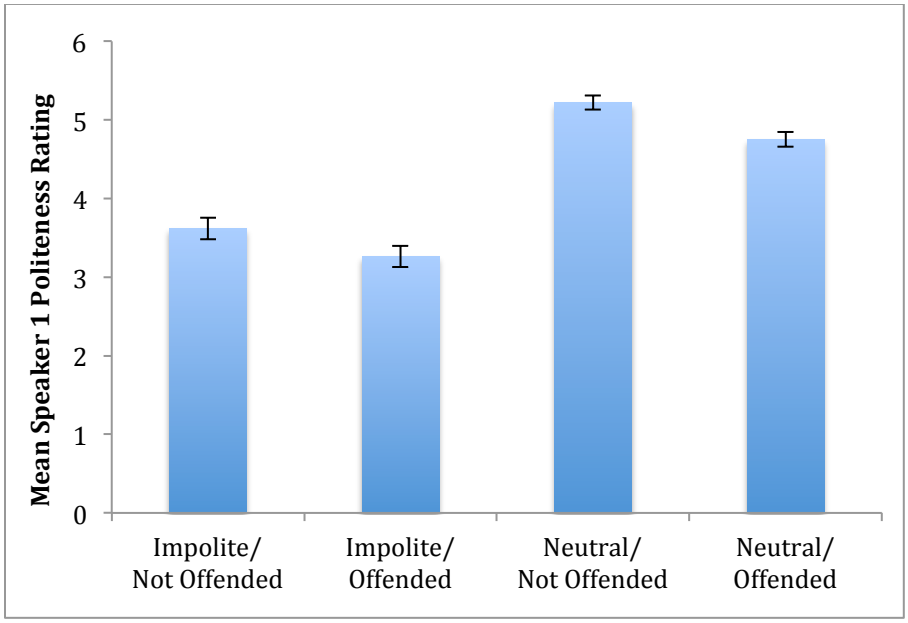


Figure 17. Mean politeness rating for Speaker 1 across the four dialogue conditions in Experiment

The influence of S2Response even extended to neutral (polite) utterances: though Speaker 1 was not, on average, considered impolite for saying something neutral, Speaker 1's politeness rating for neutral utterances was significantly higher when Speaker 2 gave non-offended responses ($M = 5.23$, $SD = 0.85$) than when Speaker 2 gave offended responses ($M = 4.75$, $SD = 0.92$), $t(94) = 4.87$, $p < .001$, M_{diff} *95% CI* [0.28, 0.67], $d = 0.50$.

The interaction between Speaker 1 Utterance and Speaker 2 Response was not significant, $F_1(1, 94) = 0.88$, $p = .35$, $F_2(1, 28) = 0.19$, $p = .66$, $MinF'(1, 41) = 0.16$, $p = .69$.

Does interlocutor response influence Speaker2 Politeness (S2Polite). The same analysis (2x2 repeated measures ANOVA) used for S1Polite was used on S2Polite to examine the politeness rating data for Speaker 2.

There was a main effect of S2Response on Speaker 2's perceived politeness, $F_1(1, 94) = 255.70$, $p < .001$, $\eta p^2 = 0.73$, as well as in the by-items analysis, $F_2(1, 28) = 86.40$, $p = .001$, $\eta p^2 = 0.87$; $MinF'(1, 72) = 107.81$, $p < .001$. The manipulation for Speaker 2 Response resulted in the offended responses sounding less polite ($M = 3.38$, $SD = 1.19$) than the non-offended responses ($M = 5.40$, $SD = 0.83$). It was not explicitly anticipated that sounding offended would be considered less polite than not sounding offended but it is not surprising that it is more polite to not show offense.

There was no main effect for S1Utterance on S2Polite in the by-subject analysis, $F_1(1, 94) = 0.49$, $p = .49$, or in the by-items analysis, $F_2(1, 28) = 0.28$, $p = .63$; $MinF'(1, 63) = 0.18$, $p = .68$. There was, however, a significant interaction between Speaker 1 Utterance and Speaker 2 Response in the by-subjects analysis, $F_1(1, 94) = 17.01$, $p < .001$, $\eta p^2 = 0.15$ but is only marginally significant in the by-items analysis, $F_2(1, 28) = 3.69$, $p = .07$, $\eta p^2 = 0.12$. MinF' was also not significant, $MinF'(1, 41) = 3.03$, $p = .09$. The results trend towards the suggestion that Speaker 2 was considered slightly more polite when their offended response is "justified" by

Speaker 1 being impolite than when they were responding in an offended manner to a neutral comment.

Planned comparisons done for the by-subjects analysis indicate that Speaker 1 did have an influence on Speaker 2's perceived politeness. Somewhat unsurprisingly, Speaker 2 acting offended by a neutral Speaker 1 Utterance was considered significantly less polite (M = 3.22, SD = 1.35) than Speaker 2 acting offended by an impolite utterance (M = 3.54, SD = 1.28), $t(94) = 2.84$, $p = .002$, M_{diff} *95% CI* [0.10, 0.55], $d = 0.29$. Though both non-offended response ratings were quite high, Speaker 2's was also rated as less polite if they responded to an impolite Speaker 1 Utterance with a non-offended response (M = 5.29, SD = 0.94) than if they had responded to a neutral utterance (M = 5.51, SD = 0.83), $t(94) = -2.83$, $p = .006$, M_{diff} *95% CI* [-0.38, -0.07], $d = 0.29$ (see Figure 18).

Does estimated Relationship Intensity influence the perception of politeness of Speaker 1's Utterance or Speaker 2's Response? A series of separate regressions for each condition (IN, IO, NN, NO) were run regressing respective Politeness ratings on estimated Relationship Intensity.

Relationship Intensity only significantly predicted Politeness in incongruent conditions: 20.7% of the variance in Speaker 1 Politeness could be explained by Relationship Intensity in the Impolite-Not Offended condition, $F(1, 93) = 25.47$, $p < .001$, *adjusted-*$R^2 = .21$, while 9.4% variance in Speaker 2 Politeness was explained by the same predictor, $F(1, 93) = 10.76$, $p = .001$, *adjusted-*$R^2 = .09$.

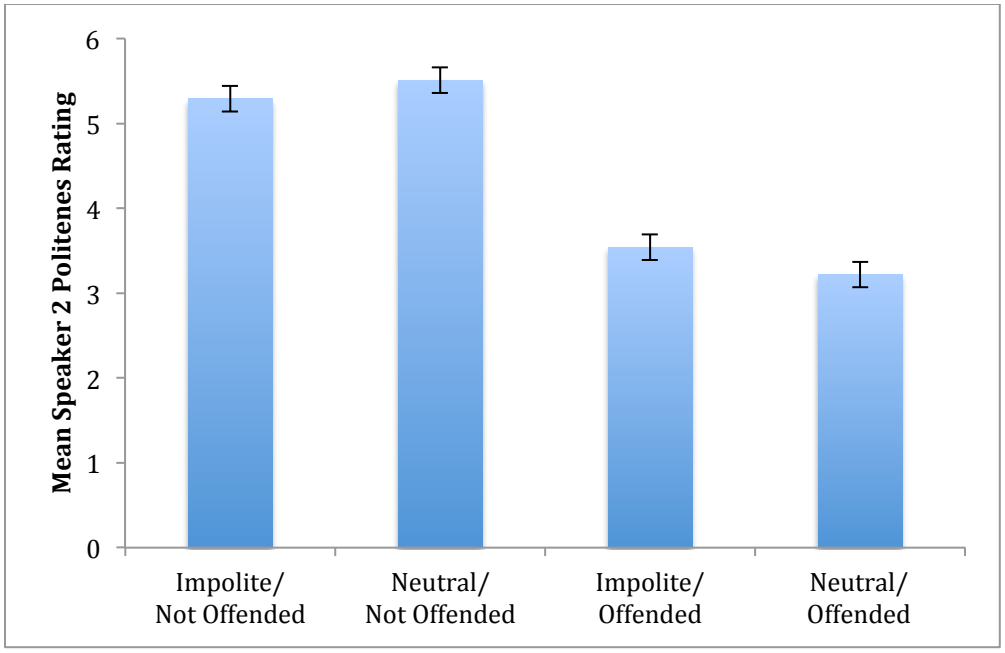


Figure 18. Mean politeness ratings for Speaker 2 across the four dialogue conditions in Experiment 2a.

This is interesting because it suggests that the over-reader perception of politeness is somewhat dependent on the relationship that the over-reader thinks conversational partners have, but only when the reaction seems unusual. When Speaker 2 acts unconcerned that Speaker 1 has been impolite, the same impolite utterance is far more polite-sounding to the over-reader when he thinks that the two speakers are friends and not strangers. The closer they are thought to be, the more polite an impolite utterance is rated. For Speaker 1 in the Impolite/Not Offended condition, every point increase in the Relationship Intensity rating corresponds to nearly half a point increase in their politeness rating ($\beta = 0.46$, $t(93) = 5.05$, $p < .001$) (see Figure 19). For Speaker 2 in the Neutral/Offended category, a point increase in Relationship Intensity was associated with about an increase in politeness rating of about a third of a point, ($\beta = 0.32$, $t(93) = 3.28$, $p = .001$).

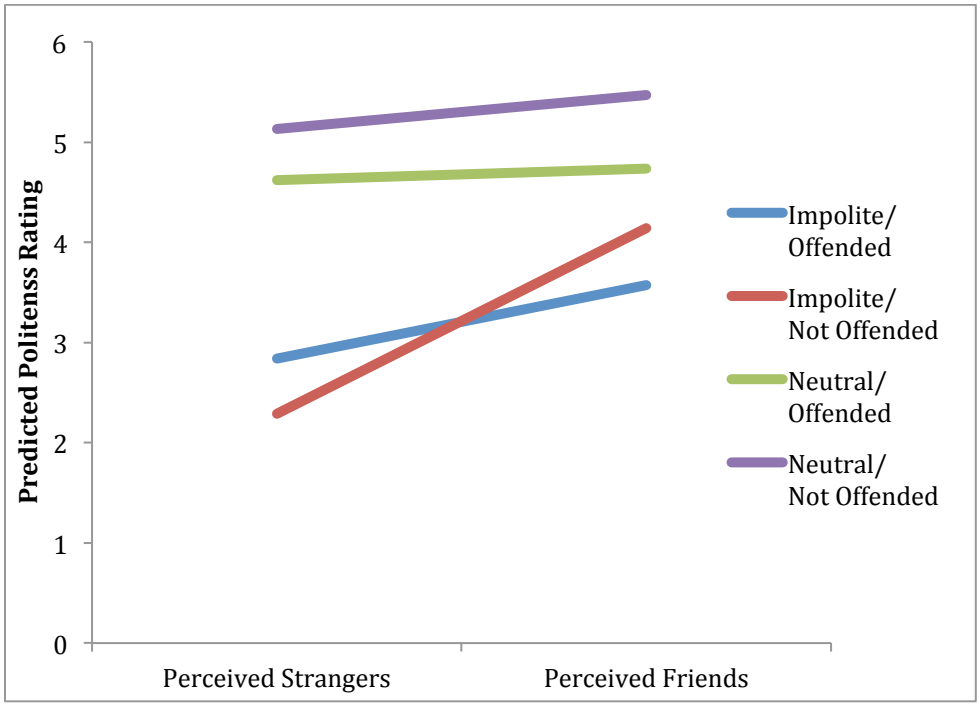


Figure 19. Separate regression lines for predicted Speaker 1 politeness rating for the different dialogue conditions when dyads are perceived to be friends vs. dyads perceived to be strangers.

Discussion for Experiment 3a

In the highly routinized tangram task, Speaker 1, the director, was in a role where she had to tell Speaker 2, the matcher, what to do. This led to Speaker 1 utterances generally being unapologetically direct. While this manner of speaking could possibly sound impolite in other situations, it was an appropriate tone for the role they were asked to play. Without any extra manipulation, Speaker 1 in these dialogues were rated by third-party over-readers as being neutral in politeness. When a blatantly impolite utterance was added to the end of dialogues, over-readers rated Speaker 1 as being much less polite, regardless of what Speaker 2 responded with. However, Speaker 2 response (either expressing offense or not) mitigated the

impoliteness of Speaker 1, suggesting that the external appearance of politeness is not merely dependent on the form or content of an utterance; it is modulated by how the utterance is received by the speaker's interlocutor. This modulation extended to even when Speaker 1 was being neutral: if Speaker 2 acted offended, Speaker 1's politeness rating would be lower than if Speaker 2 did not act offended.

Speaker 1 influenced Speaker 2 ratings in a similar fashion. Speaker 2 was docked when they responded in a polite (non-offended) fashion to Speaker 1 impoliteness, which either suggests that the third-party rater felt that Speaker 2 must have done something to deserve the impoliteness or that the non-offended line of dialogue came off as more sarcastic than intended. Speaker 2 was also not let off the hook for acting offended to an impolite comment. However, Speaker 2's politeness rating was the lowest when he acted offended at a neutral Speaker 1 comment (an intuitively sensible outcome).

Interestingly, in conditions where the reactions seem mismatched (acting offended at a neutral comment or acting not offended at an impolite comment), politeness rating depended somewhat on how close the over-reader thought the speakers were. If the over-reader thought they were reading complete strangers where the second speaker acted offended at an innocuous comment, Speaker 2 would be considered less polite (predicted rating = 2.55) than if the over-reader thought they were reading close friends (predicted rating = 3.84); in other words, a stranger would be impolite while a friend would be almost completely neutral. Similarly, when paired with a forgiving Speaker 2 who brushes off Speaker 1's otherwise impolite

provocation, Speaker 1 comes off as being neutral when over-readers think this is a pair of friends (predicted rating = 4.12) than if they thought it was a pair of strangers (predicted rating = 2.14).

Experiment 3b: Manipulating the Perception of Politeness in Puzzle Task Dialogue

Introduction

Using the same basic methodology, Experiment 3b used a different set of stimuli to examine the same basic question asked in Experiment 3a: does interlocutor response influence perception of speaker politeness. The excerpts used in Experiment 3b differ in two ways from the ones used in Experiment 3a: first, Exp 3a used dialogue taken from what is essentially the third round (Week B, Round 1) of the tangram task: the dyads were familiar with the task, its demands and were likely to have already developed strategies for doing the task. Second, the tangram task dialogue also tended to be fairly formulaic with a predictable pattern to the exchange. In contrast, Experiment 3b used Week B puzzle task dialogue (Read Between the Lines). It was unlike anything they had done so far, which meant that dyads could not rely on strategies developed in Week A. There were also no prescribed roles, so the form of the conversation could be somewhat unpredictable.

Methodology

Stimuli. Stimuli were chosen and constructed in a manner that was similar to Experiment 3a. The eight excerpts used (four friend dyads, four stranger dyads) were chosen from the original twelve used in Experiment 2b. Only Week B excerpts were

used, as they received the most neutral ratings, being neither polite nor impolite. Four versions of each excerpt were created (Condition): Impolite-Offended (IO), Impolite-Not Offended (IN), Neutral-Offended (NO), Neutral-Not Offended (NN). Speaker1 was always the one either saying something impolite or neutral, while Speaker2 was always the one reacting in an offended or not offended manner. As with the Experiment 3a excerpts, the relevant experimental manipulation was put at the end of the dialogue. Unlike the tangram dialogues (Exp 3a) where a lot of the Neutral Speaker1 lines were simply left unedited, the Neutral Speaker1 lines in Exp 3b had to be constructed wholesale.

Procedure. The experiment was put on PsychSurveys and participant recruitment was done through MTurk. Participants were paid $0.85 for their participation.

A total of 108 participants participated in this study. Every participant saw and rated eight dialogues from the eight chosen dyads using the same dependent variables used in Experiments 2a and 2b: Speaker1 Politeness (S1Polite), Speaker2 Politeness (S2Polite), and an estimate of relationship intensity (Intensity). Again, though I collected the data for whether the MTurk participants believed they were reading dyads who knew each other in real life or only online, that data was not analyzed for this study.

Every participant saw one excerpt each IO, IN, NO, or NN from both friend and stranger dyads. Partial Latin square counterbalancing was used to determine

which version (IO, IN, NO, or NN) of each of the eight excerpts the participants would see. The order of presentation was randomized by participant.

Results

Before answering the question of whether the perceived politeness of Speaker 1's utterance mediated by Speaker 2's response in novel task dialogues, a basic manipulation check was run to demonstrate that Speaker 2 was indeed rated as a separate entity from Speaker 1.

The influence of S2Response even extended to neutral (polite) utterances: though Speaker 1 was not, on average, considered impolite for saying something neutral, Speaker 1's politeness rating for neutral utterances was significantly higher when Speaker 2 gave non-offended responses ($M = 5.36$, $SD = 0.85$) than when Speaker 2 gave offended responses ($M = 5.02$, $SD = 0.91$), $t(107) = 3.73$, $p < .001$, M_{diff} *95% CI* [0.16, 0.51], $d = 0.35$.

Are Speaker 1 and Speaker 2 rated as separate entities? To ensure that Speaker 1 and Speaker 2 were rated as separate entities and not as a single unit, dependent samples t-tests were run on all the ratings for the situation-incongruent Dialogue Conditions, IN (Impolite-Not Offended) and NO (Neutral-Offended). In Dialogue Condition IN, Speaker 1 ($M = 3.62$, $SD = 1.13$) was rated as significantly less polite than Speaker 2 ($M = 4.99$, $SD = 0.93$), $t(107) = -12.00$, $p < .001$, $d = 1.16$. In NO, Speaker 2 ($M = 2.92$, $SD = 1.37$) was rated as being significantly less polite than Speaker 1 ($M = 5.03$, $SD = 1.18$), $t(107) = 14.59$, $p < .001$, $d = 1.41$

Does the interlocutor influence Speaker 1 Politeness (S1Polite)? Again, to analyze whether manipulating interlocutor response to impolite utterances, a 2x2 (Speaker 1 Utterance x Speaker 2 Response) repeated measures ANOVA was run. MinF' was used to assess generalizability across both subjects and items.

Overall, there was a significant main effect for S1Utterance on S1Polite in the by-subject analysis, $F_1(1, 107) = 260.91$, $p < .001$, $\eta p^2 = 0.71$, as well as in the by-items analysis, $F_2(1, 28) = 161.78$, $p < .001$, $\eta p^2 = 0.85$; $MinF'(1, 67) = 99.86$, $p < .001$. The experimental manipulation to alter the politeness of Speaker 1's utterance worked: the utterances that were meant to be impolite were perceived as impolite ($M = 3.25$, $SD = 1.05$), while the neutral ones were perceived to be more polite ($M = 5.19$, $SD = 0.74$).

Most importantly, there was a main effect of S2Response on Speaker 1's perceived politeness, $F_1(1, 107) = 38.06$, $p < .001$, $\eta p^2 = 0.26$, as well as in the by-items analysis, $F_2(1, 28) = 9.41$, $p = .005$, $\eta p^2 = 0.25$; $MinF'(1, 59) = 7.54$, $p = .009$. Speaker 2 Response had an effect on the perception of Speaker 1's politeness. Regardless of what Speaker 1 said, the perception of how polite Speaker 1 had been was influenced by Speaker 2 sounding offended ($M = 3.98$, $SD = 0.77$) or not offended ($M = 4.46$, $SD = 0.77$), $t(107) = 5.74$, $p < .001$, M_{diff} *95% CI* [0.40, 0.83].

The interaction between Speaker 1 Utterance and Speaker 2 Response was significant for the by-subjects analysis, $F_1(1, 107) = 5.20$, $p = .03$, $\eta p^2 = 0.05$, but not for the by-items analysis, $F_2(1, 28) = 0.90$, $p = .35$; $MinF'(1, 38) = 0.77$, $p = .39$. Planned comparisons done for the by-subjects analysis indicate that Speaker 1 did

have an influence on Speaker 2's perceived politeness. When saying something impolite, Speaker 1 was considered less polite when they received an offended Speaker 2 response ($M = 2.94$, $SD = 1.26$) than when Speaker 2 gave a non-offended response ($M = 3.56$, $SD = 1.13$), $t(107) = 5.74$, $p < .001$, M_{diff} *95% CI* [0.40, 0.83], $d = 0.55$. Even neutral statements were considered less polite if they were followed by an offended reaction ($M = 5.02$, $SD = 0.91$) than a non-offended reaction ($M = 5.36$, $SD = 0.85$), $t(107) = 3.73$, $p < .001$, M_{diff} *95% CI* [0.16, 0.51], $d = 0.35$. The difference between the two Speaker 2 responses was larger for the Impolite condition ($M_{diff} = 0.62$, $SD = 1.11$) than the Polite condition ($M_{diff} = 0.33$, $SD = 0.93$), $t(107) = 2.28$, $p = .03$, M_{diff} *95% CI* [0.04, 0.53], $d = 0.22$ (see Figure 20). The perceived extent of Speaker 1's impoliteness was more affected by Speaker 2's offense when Speaker 1 was being impolite to begin with.

Does interlocutor influence Speaker 2 Politeness (S2Polite)? There was a main effect of S2Response on Speaker 2's perceived politeness, $F_1(1, 107) = 309.57$, $p < .001$, $\eta p^2 = 0.73$, as well as in the by-items analysis, $F_2(1, 28) = 131.18$, $p < .001$, $\eta p^2 = 0.82$; *MinF'*$(1, 54) = 7.54$, $p < .001$. Speaker 2 was thought of as far less polite when they responded in an offended ($M = 3.09$, $SD = 0.80$) vs. non-offended tone ($M = 5.22$, $SD = 1.07$), $t(107) = 17.56$, $p < .001$, M_{diff} *95% CI* [1.89, 2.36].

There was no main effect for S1Utterance on S2Polite in the by-subject analysis, $F_1(1, 107) = 0.51$, $p = .48$, or in the by-items analysis, $F_2(1, 28) = 0.11$, $p = .48$, $\eta p^2 = 0.74$; *MinF'*$(1, 67) = 0.77$, $p = .76$.

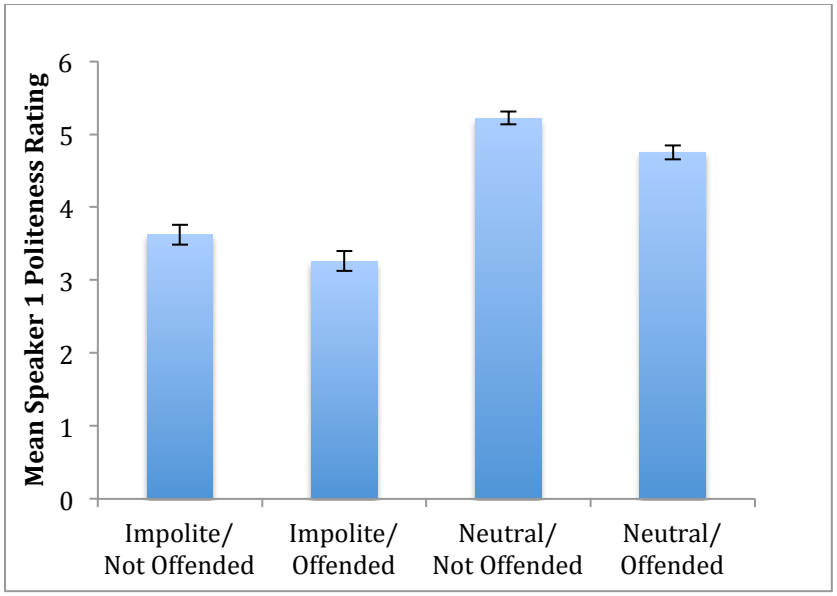


Figure 20. Speaker 1 Politeness Ratings by dialogue condition for Experiment 3b (Puzzle).

There was, however, an interaction between S2Response and S1Utterance, $F_1(1, 107) = 33.08$, $p < .001$, $\eta p^2 = 0.24$, as well as in the by-items analysis, $F_2(1, 28) = 4.49$, $p = .43$, $\eta p^2 = 0.14$; *MinF'*$(1, 36) = 3.96$, $p = .05$. Planned comparisons show a similar pattern of results to Speaker 2 in Experiment 3a: a non-offended response is judged as less polite when in reaction to an impolite utterance ($M = 4.99$, $SD = 0.93$) than to a neutral one by Speaker 1 ($M = 5.44$, $SD = 0.89$), $t(107) = -5.57$, $p < .001$, M_{diff} *95% CI* [-0.62, -0.29], $d = 0.53$. Offended responses are judged as less polite when given in reply to a neutral utterance by Speaker 1 ($M = 2.92$, $SD = 1.21$) than an impolite one ($M = 3.26$, $SD = 1.29$), $t(107) = 2.77$, $p = .007$, M_{diff} *95% CI* [0.10, 0.59], $d = 0.27$ (see Figure 21).

Does estimated Relationship Intensity influence the perception of politeness of Speaker 1's Utterance or Speaker 2's Response? A series of separate regressions for

each condition (IN, IO, NN, NO) were run regressing respective Politeness ratings on estimated Relationship Intensity.

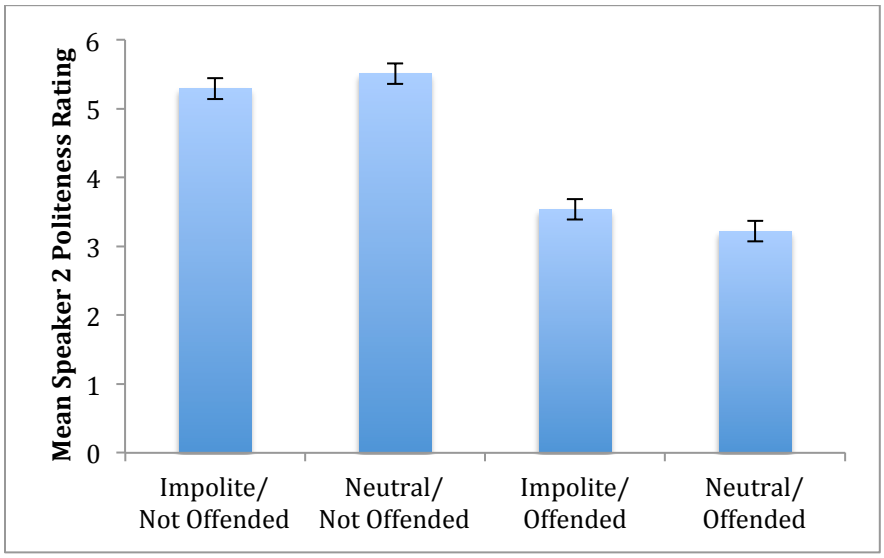


Figure 21. Speaker 2 Politeness Ratings by dialogue condition for Experiment 3b (Puzzle).

Unlike in Experiment 3a where the only Relationship Intensity only significantly predicted Politeness in incongruent conditions (Impolite/Not Offended and Neutral/Offended), Relationship Intensity significantly predicted Politeness in both Speaker 1 ratings for Impolite/Offended, $F(1, 106) = 6.11$, $p = .02$, *adjusted*-$R^2 = .05$, and $F(1, 106) = 15.95$, $p < .001$, *adjusted*-$R^2 = .12$. For Speaker 2, results were similar to Experiment 3a, where Relationship Intensity was a significantly predictor in only the incongruent condition, Neutral/Offended, $F(1, 106) = 6.11$, $p = .02$, *adjusted*-$R^2 = .05$.

The perception of Speaker 2's politeness only changes with the over-reader's perception of the closeness of the dyad in cases where Speaker 2 acts out of line with expectations ($\beta = 0.32$, $t(107) = 3.49$, $p = .001$). Over-reader participants who thought

they were listening to complete strangers rated Speaker 2 as being impolite (predicted rating = 2.46), while those who thought they were listening to close friends excused Speaker 2 sounding offended (predicted rating = 3.74).

For Speaker 1, perceived politeness changed in both expected (Impolite-Offended; $\beta = 0.23$, $t(107) = 2.47$, $p = .02$) and unexpected (Impolite-Not Offended; $\beta = 0.36$, $t(107) = 3.99$, $p < .001$) conditions. When Speaker 2 gave a non-offended response, Speaker 1 was considered neutral in politeness when third-party over-readers thought they were reading a pair of close friends (predicted rating = 4.09). When over-readers thought they were reading strangers, even a non-offended response did make Speaker 1 seem impolite (predicted rating = 3.15). This advantage of being close friends also protects Speaker 1 when Speaker 2 is obviously offended (predicted rating = 3.84). However, Speaker 1 becomes unequivocally rude when Speaker 2 is offended and when over-readers think they are reading complete strangers (predicted rating = 2.40) (see Figure 22).

Discussion for Experiment 3a and 3b

Overall, it did not appear that the two different types of tasks (novel and less routine vs. practiced and more routine) produced dialogues that impacted the perception of politeness differently. The results for Experiment 3a and 3b largely mirror each other. Impolite utterances from Speaker 1 or offended responses from Speaker 2 are always going to be less polite than polite utterances and non-offended responses. Nevertheless, conversational context can shift the perception of the politeness.

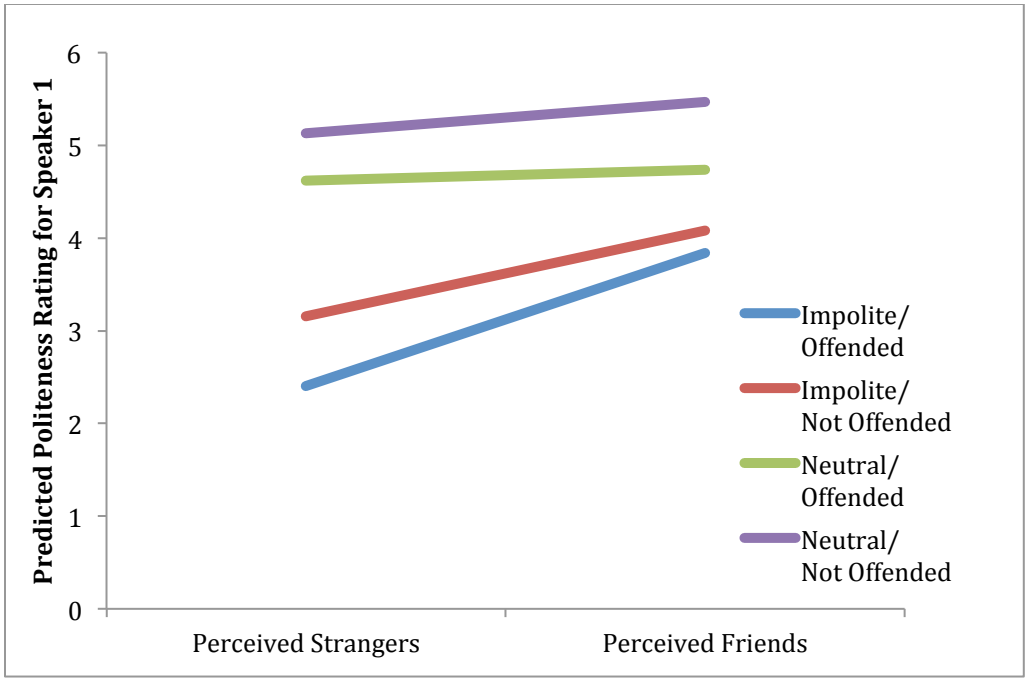


Figure 22. Separate regression lines for predicted Speaker 1 politeness rating in Experiment 3b (Puzzle) for the different dialogue conditions when dyads are perceived to be friends vs. dyads perceived to be strangers.

A forgiving Speaker 2 can shift an impolite Speaker 1's average politeness rating up to the neutral mid-point of the scale. An insulting Speaker 1 somewhat justifies the rudeness of Speaker 2 sounding offended. Conversely, Speaker 1 is penalized for completely innocuous comments that offend Speaker 2. And a gracious Speaker 2 is somehow tainted for simply responding to a rude Speaker 1.

Across the two experiments, perception of politeness depended on whether the over-reader thought they were reading a friend or stranger dyad. Most of the conditions that triggered this distinction contained some socially curious behavior: acting impolite in general or acting offended at nothing. Friends were mostly given passes while strangers were thought of as impolite.

General Discussion

The results of this dissertation suggest that there are initial differences between friends and strangers when they first start working together on concrete, goal-oriented tasks using instant messaging as their only source of communication. Some differences, however, fade as dyads become accustomed to this mode of working. This is one of the few longitudinal studies that examine pairs of friends and strangers engaging in relatively lengthy text-based conversations on specific experimental tasks. It is also one of the few that uses largely naturalistic stimuli to examine how third-party over-readers perceive politeness in these types of conversations. Finally, it is one of the first, if not the first, empirical study to test politeness-related hypotheses derived from Arundale's Conjoint Co-Constitution Model of discourse.

Experiment 1, which compared the efficiency with which friends and strangers were able to complete a standard tangram task using real-time online chat over time, indicated that there were largely no differences between the number of unique descriptive words that friend and stranger dyads used. However, it was shown that friend directors took more frequent turns in order to achieve the same goals as the stranger directors, particularly for the first round, suggesting that their turns are shorter. This may be an extension of previous research that has shown that increased familiarity with this medium of communication leads to the use of a greater number of shorter turns. The more familiar one is with communicating with a specific person using this particular medium, the quicker turns will become.

There was more variance in how friend and stranger matchers used descriptive words. Friend matchers were almost always more voluble in their descriptions than strangers. While stranger matchers quickly defaulted to merely acknowledging that they had found a target or uttering a single word that described the target at all (a perfectly sufficient strategy), some friend matchers used two or more words through the entire three weeks.

Finally, the results of Experiment 1, which showed a substantial difference between Week 1 and Week 2 performance, and no substantial differences between Week 2 and Week 3 performance, indicates that task performance was not necessarily improved with an extra week.

Experiment 2 used stimuli taken from all three weeks of the tangram task (Round 1 from Weeks A, B and C) as well as the novel puzzle tasks to examine whether third-parties could tell the difference between the dialogues of friends and strangers and to assess whether friends and strangers differed in perceived politeness. Despite the intuitive sense that friends and strangers must sound different (and the evidence from Experiment 1 that there are differences between friends and strangers), third-party over-readers were remarkably bad at telling apart friend and stranger dyads. This was particularly true for the later weeks than the earliest week. Third-party over-readers were a little less predictable in their guesses for puzzle tasks than tangram tasks, perhaps suggesting that the format of the task did differ in its ability to conceal or convey whether a given pair had been friends who spoke regularly for over

a year or whether they were strangers who had just met during the course of the study.

Even though participants seemed to be unable to explicitly tell apart who were friends and who were strangers in chat, strangers were nonetheless more polite than friends in most of the tangram task. In the puzzle task, they only differed in the first week. Politeness also declined across time for both puzzle and tangram tasks. (It should be noted that, on a 7-point scale, speakers went from an average rating of about 5.1 to 4.6 across the three weeks: speakers were quite polite to begin with and became neither polite nor impolite over time, rather than lapsing into outright impoliteness.)

In all cases (for Speakers 1 and 2, as well as in puzzle and tangram tasks), stranger dyads experienced a steeper drop in politeness between Weeks 1 and 3 than friends. With the exception of Speaker 1 during the tangram task, strangers and friends became (and stayed) indistinguishable by the second week.

Experiment 3 compared the Brown and Levinson (1987) Politeness Theory to Arundale's (1999) Conjoint Co-Constitution theory. Brown and Levinson posit that politeness is expressed through the form of an utterance; the intent of the speaker is what determines the politeness of an utterance. Arundale posits that because meaning is conjointly determined, the speaker's intent only partially determines the politeness of an interaction; the interlocutor must buy into whether an utterance is polite or impolite. The results of Experiment 3 indicate that obviously impolite comments from speakers are always going to be interpreted as being less polite than utterances that

are meant to be neutral. They also indicate that an interlocutor acting offended is always going to be considered less polite than not acting offended. However, that does not mean that the interlocutor influence does not shift the perception of politeness. On the contrary, Speaker 1 is considered more impolite when they do offend Speaker 2 than when they don't. Speaker 2 is even (somewhat unfairly) penalized for responding in a non-offended fashion to impoliteness from Speaker 1 compared to responding in the same fashion to a neutral Speaker 1 utterance.

Furthermore, the perception of politeness shifts depending on the beliefs of the third-party: if an outsider thinks that he is listening to a pair of friends, Speaker 1 is not judged as impolite even when the utterance would normally be considered very direct and face-threatening. If an outsider thinks he is listening to strangers, Speaker 1 is considered impolite for the very same comment she was previously excused for. For Speaker 2, opinion only changes in the most jarring and incongruent situation: acting offended to a remark that is, on its face, harmless. If Speaker 2 was thought to be close friends with Speaker 1, he is largely excused for the outburst; he is still considered somewhat less polite but its well within range of being "neither polite nor impolite." However, strangers are not granted this leeway: if somebody you do not know says something you know will not be seen as an offensive comment, you will likely be thought of as very impolite if you treat it as an offensive comment. The results of Experiment 3, as a whole, suggest that while Brown and Levinson were accurate in stipulating that some utterances will be considered more polite than

others. Yet Arundale was also accurate in positing that interlocutor reaction and conversational context influences perceived politeness.

Since there is some evidence that perceived closeness between speakers influences the perceived politeness of a pair of speakers, future studies that use the procedure in Experiment 3 can add in a condition where participants are explicitly biased to think they are reading a pair of friends or a pair of strangers (instead of letting participants guess the closeness of any given dyad).

The reviewable nature of text-based communication may also play a role in how the perception of politeness by a third-party is affected by interlocutor response. A chat transcript is fully reviewable by third-parties, while memory may be fallible. Would Experiment 3 have shown the same influence of response or provocation on the perception of politeness if third-parties were only given one chance to read each line? The ability to go back and re-read dialogue may have had an ameliorating effect that is even stronger in unreviewable or hard-to-review conversation (such as a small chat window), where third-parties may retrospectively alter what they think they remember hearing or reading in order to bring the rest of an exchange into tonal accordance with utterances and responses that may be jarring otherwise.

The question remains as to why participants in this dissertation showed some initial (Week A) differences between friends and strangers while studies done on the HCRC Map Task have not. One possibility is that the Map Task corpus was too straightforward and short, with the experimental environment being too conducive to formal verbal behavior, which erases differences between friend dyads and stranger

dyads. Since friends and strangers can be somewhat reliably be distinguished based on politeness, it may be that any environment that discourages informality also discourages friends from acting like friends.

Furthermore, there could have been carryover effects due to the construction of the task: the HCRC Map Task used a fully within-subjects design, with every participant doing the task participating with a friend and a stranger in back-to-back sessions. As seen in this dissertation, friend and stranger dyads quickly became indistinguishable, often by the second week. Increased familiarity with the task's format – even when the specific task itself changed and even when a full week separated experimental sessions – decreased any noticeable differences between friends and strangers. It is thus unsurprising that few studies have found quantifiable differences between friends and strangers in the HCRC Map Task Corpus: repeating the same task multiple times in a single session seems likely to have flattened differences as participants took what they learned in their session with a friend partner and applied it to a subsequent iteration of the task with a stranger partner (or what they learned with a stranger and applied it to their turn with a friend).

This dissertation has provided evidence that perceived politeness is not merely determined by the form of a request or utterance by one speaker. It is also partially determined by how the interlocutor responds to the speaker. However, in future work, more attention should be given to how different forms may or may not be influenced by interlocutor response. Due to the format and time pressure of the tasks administered in the corpus collection, there were few "typical" markers of politeness:

very few say “please”, no one needs to indirectly impose on their partner to fulfill a request because this imposition is assumed to be necessary. Yet some manifestation of politeness must be lingering in the linguistic form of the dialogue, as third-party over-readers did see differences in the excerpts of dialogue that were not manipulated. A finer-grained content or discourse analysis may help shed light on what third-party over-readers were picking up on. Re-running a similar study with more complex and contentious tasks that are less likely to lead to routinized and predictable verbal behavior may also yield interesting results that further shed light on how politeness is manifested in online, text-based conversation.

Appendix A: Additional Graphs and Tables

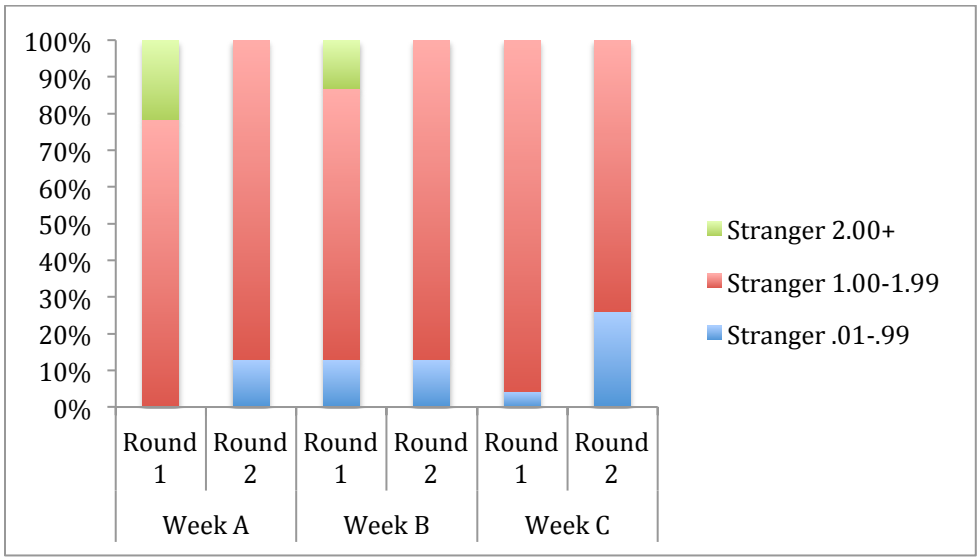


Figure A1. Breakdown of number of turns taken by Strangers matchers in Experiment 1.

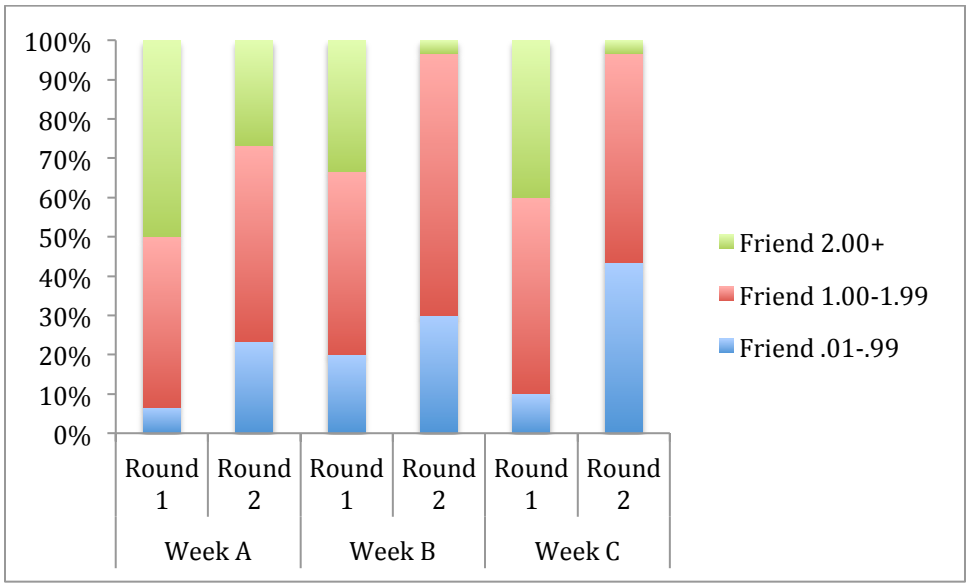


Figure A2. Breakdown of number of turns taken by Friend matchers in Experiment 1.

	Tangram	Puzzle
Week A	1754.12	1096.32
	(707.37)	(363.67)
Week B	1210.4	858.33
	(433.26)	(300.92)
Week C	1178.33	976.2
	(405.66)	(360.53)
TOTAL	4142.86	2930.86
	(1323.32)	(904.54)

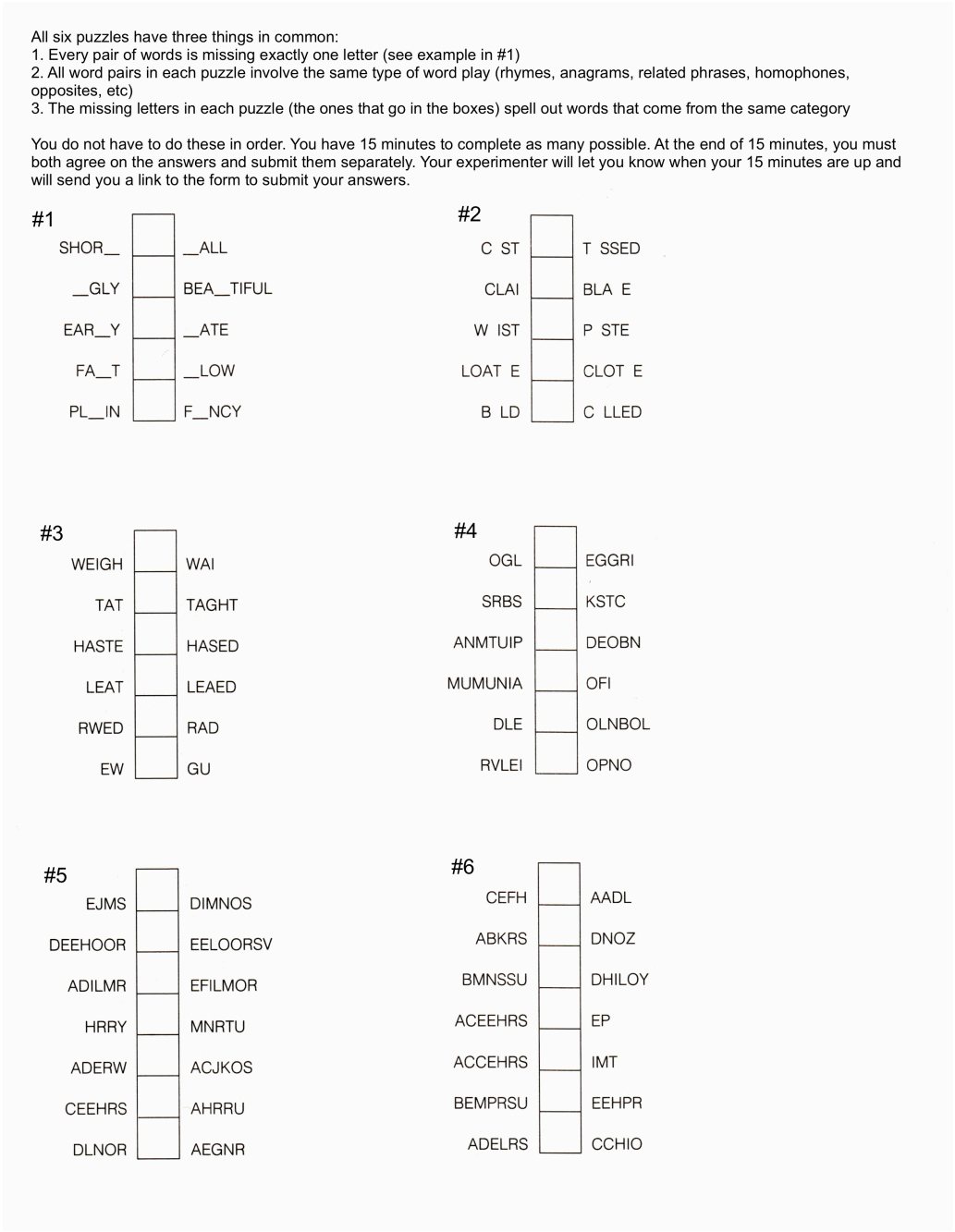
Table A1. Total number of words used in Experiment 1 by task.

Appendix B: Novel Puzzle Tasks

Common Elements (Figure B1). The first puzzle resembles a visual version of a remote associates task, where participants are given a series of diagrams of objects and asked to figure out what they have in common and then to pick a final object that belongs to the group.

Read Between the Lines (Figure B2). The second puzzle gives participants superimposed images of classic book titles and asks that the participants to visually disentangle them and to name both titles; those who read or watch movie adaptations of books will have a distinct advantage though it is solvable by someone who is good at mentally manipulating visual stimuli.

Logi-Quiz (Figure B3). The final puzzle is made up of anagram-like sub-puzzles of increasing difficulty: related word pairs are given with the same letter missing from both. The missing letters in a series of word pairs spells out a major American city (i.e. Miami, Dallas, Seattle, etc); conceivably, if participants figure that out, they need only solve a couple of pairs per set to deduce the correct answer for each series.



References

Althoff, T., Salehi, N., & Nguyen, T. (2013). Random Acts of Pizza: Success Factors of Online Requests.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., . . . Weinart, R. (1991). The HCRC Map Task Corpus. *Language and Speech, 34*, 351-366.

Anderson, J. F., Beard, F. K., & Walther, J. B. (2007). *Turn-taking and the local management of conversation in a highly simultaneous computer-mediated communication system*: DigitalCommons@ Kennesaw State University.

Andersson, J. (2001). Net effect of memory collaboration: How is collaboration affected by factors such as friendship, gender and age? *Scandinavian Journal of Psychology, 42*(4), 367-375.

Andersson, J., & Rönnberg, J. (1995). Recall suffers from collaboration: Joint recall effects of friendship and task complexity. *Applied Cognitive Psychology, 9*(3), 199-211.

Arndt, H., & Janney, R. W. (1985). Politeness revisited: cross-modal supportive strategies. *International Review of Applied Linguistics, 23*(4), 281-300.

Arundale, R. B. (1999). An alternative model and ideology of communication for an alternative to politeness theory. *Pragmatics, 9*(1), 119-154.

Baron, N. S. (2010). Discourse structures in instant messaging: The case of utterance breaks. *Language@ Internet, 7*(4), 1-32.

Bortfeld, H., Leon, S. E., Bloom, J. E., Schober, M. F., & Brennan, S. (2001). Disfluency rates in spontaneous speech: Effects of age, relationship, topic, role, and gender. *Language and Speech, 44*, 123-149.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Bunz, U., & Campbell, S. W. (2004). Politeness accommodation in electronic mail. *Communication Research Reports, 21*(1), 11-25.

Burke, M., & Kraut, R. (2008). *Mind your Ps and Qs: the impact of politeness and rudeness in online communities.* Paper presented at the Proceedings of the 2008 ACM conference on Computer supported cooperative work.

Campbell, N. (2007). *Whom we laugh with affects how we laugh.* Paper presented at the Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter.

Carlo, J. L., & Yoo, Y. (2007). "How may I help you?" Politeness in computer-mediated and face-to-face library reference transactions. *Information and Organization, 17*(4), 193-231.

Carlson, J. R., & Zmud, R. W. (1999). Channel expansion theory and the experiential nature of media richness perceptions. *Academy of management journal, 42*(2), 153-170.

Carson, R. J., & Beeson, C. M. (2013). Crossing language barriers: Using crossed random effects modelling in psycholinguistics research. *Tutorials in Quantitative Methods for Psychology, 9*(1), 25-41.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 335-359.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In J. M. L. L.B. Resnick, & S.D. Teasley (Ed.), *Perspectives on socially shared cognition*. Washington, DC: APA Books.

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

Colvin, C. R., Vogt, D., & Ickes, W. (1997). Why do friends understand each other better than strangers do?

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5), 769-786.

Cummings, J., Lee, J., & Kraut, R. (2006). Communication technology and friendship during the transition from high school to college. *Computers, phones, and the Internet: Domesticating information technology*, 265-278.

D'Urso, S. C., & Rains, S. A. (2008). Examining the Scope of Channel Expansion A Test of Channel Expansion Theory With New and Traditional Communication Media. *Management Communication Quarterly, 21*(4), 486-507.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management science, 32*(5), 554-571.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Darics, E. (2014). The Blurring Boundaries Between Synchronicity and Asynchronicity New Communicative Situations in Work-Related Instant Messaging. *International Journal of Business Communication, 51*(4), 337-358.

De Choudhury, M., & De, S. (2014). *Mental health discourse on reddit: Self-disclosure, social support, and anonymity.* Paper presented at the Eighth International AAAI Conference on Weblogs and Social Media.

Dunne, M., & Ng, S. H. (1994). Simultaneous speech in small group conversation: all-together-now and one-at-a-time? *Journal of Language and Social Psychology, 13*, 45-71.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS One, 5*(4), e10068.

Fleming, J. H., Darley, J. M., Hilton, J. L., & Kojetin, B. A. (1990). Multiple audience problem: a strategic communication perspective on social perception. *Journal of Personality and Social Psychology, 58*(4), 593.

Fox Tree, J. E. (2007). Folk notions of um and uh, you know and like. *Text & Talk, 27*(3), 297-314.

Fox Tree, J. E., Mayer, S. A., & Betts, T. E. (2011). Grounding in instant messaging. *Journal of Educational Computing Research, 45*(4), 455-475.

Francis, L., Holmvall, C. M., & O'Brien, L. E. (2015). The influence of workload and civility of treatment on the perpetration of email incivility. *Computers in Human Behavior, 46*, 191-201.

Fussell, S. R., & Krauss, R. M. (1989). Understanding friends and strangers: The effects of audience design on message comprehension. *European Journal of Social Psychology, 19*(6), 509-525.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in personality, 37*(6), 504-528.

Gould, O. N., Osborn, C., Krein, H., & Mortenson, M. (2002). Collaborative recall in married and unacquainted dyads. *International Journal of Behavioral Development, 26*(1), 36-44.

Haugh, M. (2007). The co-constitution of politeness implicature in conversation. *Journal of Pragmatics, 39*, 84-110.

Hope, L., Ost, J., Gabbert, F., Healey, S., & Lenton, E. (2008). "With a little help from my friends…": The role of co-witness relationship in susceptibility to misinformation. *Acta Psychologica, 127*(2), 476-484.

Hornstein, G. A. (1985). Intimacy in conversational style as a function of the degree of closeness between members of a dyad. *Journal of Personality and Social Psychology, 49*, 671–681.

Isaacs, E., Walendowski, A., Whittaker, S., Schiano, D. J., & Kamm, C. (2002). *The character, functions, and styles of instant messaging in the workplace.* Paper presented at the Proceedings of the 2002 ACM conference on Computer supported cooperative work.

Ishii, K., Reyes, J. A., & Kitayama, S. (2003). Spontaneous attention to word content versus emotional tone differences among three cultures. *Psychological Science, 14*(1), 39-46.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). The big five inventory—versions 4a and 54. *Berkeley: University of California, Berkeley, Institute of Personality and Social Research*.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122-149.

Lam, C., & Mackiewicz, J. (2007). *A case study of coherence in workplace instant messaging.* Paper presented at the Professional Communication Conference, 2007. IPCC 2007. IEEE International.

Lewin-Jones, J., & Mason, V. (2014). Understanding style, language and etiquette in email communication in higher education: a survey. *Research in Post-Compulsory Education, 19*(1), 75-90.

Liu, K., Fox Tree, J. E., Blackwell, N., & Walker, M. (unpublished manuscript). Entrainment Between Friends and Strangers.

Locker, L., Hoffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior research methods, 39*(4), 723-730.

Macaulay, R. (2002). You know, it depends. *Journal of Pragmatics, 34*(6).

Matsumoto, Y. (1988). Reexamination of the universality of face: Politeness phenomena in Japanese. *Journal of Pragmatics, 12*(4), 403-426.

Nardi, B. A., Whittaker, S., & Bradner, E. (2000). *Interaction and outeraction: instant messaging in action.* Paper presented at the Proceedings of the 2000 ACM conference on Computer supported cooperative work.

Nobarany, S., & Booth, K. S. (2014). Use of politeness strategies in signed open peer review. *Journal of the Association for Information Science and Technology*.

Planalp, S. (1993). Friends' and Acquaintances' Conversations II: Coded Differences. *Journal of Social and Personal Relationships*(10), 339-354.

Planalp, S., & Benson, A. (1992). Friends' and Acquaintances' Conversations I: Perceived Differences. *Journal of Social and Personal Relationships, 9*, 483-506.

Raaijmakers, J. G., Schrijnemakers, J. M., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41*(3), 416-426.

Rockwell, P. (2003). Empathy and the expression and recognition of sarcasm by close relations or strangers. *Perceptual and Motor Skills, 97*(1), 251-256.

Savitsky, K., Keysar, B., Epley, N., Carter, T., & Swanson, A. (2011). The closeness-communication bias: Increased egocentrism among friends versus strangers. *Journal of Experimental Social Psychology, 47*, 269-273.

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

Smoski, M., & Bachoroski, J.-A. (2003). Antiphonal laughter between friends and strangers. *Cognition & Emotion, 17*, 327-340.

Truong, K. P., & Trouvain, J. (2012). Laughter annotations in conversational speech corpora-possibilities and limitations for phonetic analysis. *Proceedings of the 4th International Worskhop on Corpora for Research on Emotion Sentiment and Social Signals*, 20-24.

Walther, J. B., & Bazarova, N. N. (2008). Validation and application of electronic propinquity theory to computer-mediated communication in groups. *Communication Research, 35*(5), 622-645.

Waugh, J. (2013). Formality in Chat Reference: Perceptions of 17-to 25-Year-Old University Students. *Evidence Based Library and Information Practice, 8*(1), 19-34.

Wolfson, N. (1983). An empirically based analysis of complimenting in American English. *Sociolinguistics and language acquisition*, 82-95.