Supplementary Materials

Clicking in conversation: Short gaps between turns signal connection

Emma M. Templeton, Luke J. Chang, Elizabeth A. Reynolds, Marie D. Cone LeBeaumont and Thalia P. Wheatley

Github repository for this project: https://github.com/emtempleton/GapPaper

Department of Psychological and Brain Sciences
Dartmouth College
Hanover, NH 03755
Corresponding author (emma.m.templeton.gr@dartmouth.edu)

Studies 1-2

Transcription Details

Many of our analyses relied on the timestamps in each conversation transcript. Here we detail exactly what the transcripts contained and our decisions about what constituted a speech 'turn.'

Format. Below is a screenshot of the first few turns in one transcript, to illustrate the format of the transcripts. Each turn included (i) speaker information (S1 or S2), (ii) a START timestamp, (iii) an END timestamp, and (iv) text of the words spoken.

To compute gap length, we subtracted the END timestamp of the previous turn from the START timestamp of a given turn.

```
S1: 00:00:00.000 How's it going? END 00:01.335 END
S2: 00:00:02.236 I'm okay. How are you? END 00:03.215 END
S1: 00:00:03.123 Yeah, I'm good. It's been a busy... END 00:05.599 END
S2: 00:00:05.636 It's been a week. END 00:06.043 END
S1: 00:00:06.397 Yeah, honestly. With rush. Did you rush? END 00:08.880 END
S2: 00:00:08.761 I dropped on Monday. END 00:09.686 END
S1: 00:00:09.582 You dropped? Oh, I'm sorry. END 00:10.731 END
S2: 00:00:10.388 Yeah, I didn't get callbacks. END 00:11.686 END
S1: 00:00:12.067 Oh, okay. END 00:12.432 END
```

Defining a speech 'turn'. The primary unit of analysis is a *speech turn* which we defined in the following ways. For non-overlapping speech, a turn was simply all the words one speaker said before their partner began. For overlapping speech (e.g., backchannels, interjections, interruptions, and false starts), we kept the "coherent thoughts" together (Version A below). Another way to handle overlapping speech would be to initiate a new turn every time another speaker talks (Version B below). Consider how this real moment of overlapping speech could be

transcribed in two different ways:

Version A

Turn 1, S1: It was a beautiful day. That's true. Things could be much worse, but...

Turn 2, S2: How can you be hanging in there today with weather like this?

Version B

Turn 1, S1: It was a beautiful day. That's true. Things could be much--

Turn 2, S2: How can you be--

Turn 3, S1: --worse--

Turn 4, S2: --hanging in there today--

Turn 5, S1: --but.

Turn 6, S2: --with weather like this?

During moments of overlapping speech, Version B results in many more turns with sentences “broken apart” across those turns. We felt that Version A better captured our own experiences engaging in conversation as well as listening to these recorded conversations -- where we are able to effortlessly integrate the words one speaker is saying even if another speaker is simultaneously speaking. Note that in the conversations we recorded, speech doesn’t overlap by much time (median negative latency is 297 ms). This is in line with what has been found in previous work on overlaps and so the task of assigning words to the correct speaker is easier than if we were to instruct people to talk over each other in a way that does not tend to happen naturally.

Although we removed the timestamps from this example to increase readability, note that for Version A, the START timestamp for Turn 2 would occur earlier in time than the END timestamp for Turn 1. This would allow us to compute how far into Turn 1 Speaker 2 initiated Turn 2.

It is also important to note that Version A and Version B are both correct representations of what happened in the conversation. However, the different versions emphasize different aspects of turn-taking. While we believed Version A was better suited for our particular question, there may be other projects that benefit from defining turn taking like Version B. What we want to emphasize is that it is important for researchers interested in similar questions to think carefully about how to define a "turn" and to ensure that definition is (i) disclosed and (ii) applied consistently across all transcripts in a given dataset.

Transcription Company. The transcriptions (and therefore the timestamps) that we used to run the gap length analyses presented in the paper were all done by one company -- Scribie (https://scribie.com/). More details about the transcription that was completed for each individual conversation video can be found in the Supplement folder of this project's Github repository (ConversationDatasetDetails.doc).

Robustness checks for within-conversation analysis

We examined the relationship between gap length and social connection over the course of a conversation by dividing the 10-minute conversation into twenty 30s bins. Our first robustness check was to examine this same relationship over a variety of bin sizes. We examined these bin sizes: 5, 10, 15, 25, 30, 40, 50, 60, 100, 120, 150, 200, and 300 seconds. For each bin, we computed the average gap length for turns that occurred in that bin as well as the average

connection rating for each speaker in that conversation. We ran a mixed linear effects model predicting the temporal dynamics of social connection based on fluctuations in average gap length controlling for linear effects of time. To account for variations in average gap length between dyads, we included Dyad ID as a random intercept and additionally modeled Subject ID as a random intercept because subjects participated in multiple conversations. For stranger conversations (Study 1) we observed a significant negative ($p<.05$) effect of gap length on connection for all of these bin sizes. For friend conversations (Study 2) we observed a significant negative effect of gap length on connection of all of these bin sizes except for the largest bin size of 300 seconds ($p=.06$). This is consistent with the across conversation finding that at these larger timescales friend reports connection become uniformly high and do not sufficiently vary across the conversations to assess the relationship with average gap length.

For our second robustness check, we created an empirical null distribution to ensure that the effect we were observing could not be explained by any offsets in lag between changes in gap length and connection ratings. To do this, we generated surrogate data by randomly permuting the order of gap lengths within each conversation using a circle-shifting procedure and re-fitting the model predicting social connection 100 times. We performed this procedure for each bin size listed above. As you can see in Figs S3 and S4, our results cannot be explained by any offsets in lag between changes in gap length and connection ratings.

Study 3

There are two ways to manipulate the size of gaps in a recorded conversation: *proportional* (changing each original gap to be longer or shorter by a specified proportion) and *distributed* (replacing each original gap with a gap pulled from a specified distribution: short or

long). For the pre-registered Study 3 in the main text, we used proportional manipulation which had the benefit of changing the size of each gap while maintaining the natural variance of gap lengths across the conversation. However, we also ran a version of this study that used the alternative method in which the original gaps were replaced with gaps taken from one of two *distributions* (short or long). We are including the methods and results of this additional manipulation study here for two reasons. First, the results demonstrate how our effect replicated in a different set of participants, using a different method of manipulating gap length. Second, the methods provide helpful context to explain the changes we made before running the pre-registered version of this study.

Study 3 Replication

Methods. For six of the conversations from Study 1 (3 male and 3 female), we selected a short segment (mean length = 23.33 seconds) to use as stimuli. We picked conversation segments that had minimal overlapping speech, where both participants had signed a video release. These were the same 6 conversation segments that we used in the preregistered version as well.

We used these 6 stimuli to create two different conditions: Long gap and Short gap. Specifically, for each segment, we produced two versions by inserting gaps with the length drawn from two different distributions: Long gaps (mean gap = 500 ms, std = 10 ms) and Short gaps (mean gap = 50 ms, std = 10 ms).

300 participants on Amazon's Mechanical Turk listened to each of 6 conversation segments, presented in a random order. All participants heard each conversational segment only once and the version (Long, Short) of each conversation they were presented was randomly assigned. This random assignment was blocked such that, over all participants, each conversation segment was presented an equal number of times in both conditions.

After listening to each conversation segment, participants responded to two questions -- *1) How much do you think these people enjoyed their conversation?* and *2) How connected do you think these people felt toward each other?*. Participants responded using a slider bar anchored by "Not at all" (0) and "Very much" (100).

To access the study, participants were first required to successfully complete a task delivered via audio instructions. This ensured that we only included those participants who were able to listen and respond to audio instructions, a requirement for the study.

To check for participant compliance, we also collected timing information on each page of the survey. This allowed us to determine whether or not participants submitted their response before the audio file stopped playing. When we filtered the dataset based on this inclusion criteria, our number of observations dropped from 1,800 (300 participants x 6 items) to 1,584 (88% retention). We ran the same set of analyses on the full dataset as well as on this subset and the pattern of results do not change between the two. We present results on the full dataset here.

Results. We ran two mixed linear effects models with condition (Short gap, Long gap) predicting each of our two DVs: perceived enjoyment and perceived connection. We included subject and item (e.g. which of the 6 conversations was being judged) as random intercepts. Results showed a significant effect of condition such that the same conversation with short gaps was perceived as more enjoyable ($b=6.50$, $SE=0.77$, $p<.001$) and connected ($b=7.52$, $SE=0.88$, $p<.001$; Fig S5) than the version with long gaps.

Additional methods for reported Study 3

Preregistration and materials. We preregistered Study 3 before collecting data. Our full preregistration plan is here: osf.io/u2brn. We've highlighted details from the preregistration plan in this section.

Improvements over the original study. We were encouraged by the results from the first manipulation study, but wanted to be sure our result couldn't be explained by the decision to manipulate the gaps by a distributed vs proportional method or acoustic properties in the stimuli. We made three changes. First, we moved from a distributed to a proportional method of manipulating the gap lengths. Second, we recorded the natural background noise of the testing room and used that (rather than pure silence) as the audio for the gaps. Third, we included a baseline condition where the gap lengths were not manipulated.

Audio processing. Each conversation segment consisted of audio clips for every speech turn and for every gap in between each speech turn (e.g. silences). The speech turns were spliced out of the original conversation. The silences were taken from a recording of the empty testing room (to re-create the ambient noise) and were trimmed to be the length of the silences in the original conversation (or trimmed to be the length of the new, manipulated gap).

For the Control condition, these audio clips were stitched together, in the order that they appeared in the original conversation segment. We persisted in stitching together the clips in the Control condition (rather than simply playing the entire segment) to maintain continuity across the conditions. The mean gap length in the Control condition was 278.88ms.

For the Long Gap condition, we manipulated the length of the audio clips that consist of silences to be twice as long as they are in the Control condition. The mean gap length in the Long condition was 557ms (278 * 2).

For the Short Gap condition, we manipulated the length of the audio clips that consist of silences to be a fifth as short as they are in the Control condition. The mean gap length in the Short condition was 55.7ms (278 / 5).

We picked these multipliers to best match the distributions that we used in our original

manipulation study (where the Short condition had a mean gap length of 50ms and the Long condition had a mean gap length of 500ms).

We took two additional steps to improve the quality of the final conversation segment. For all conditions, the speech turn audio clips have a fade in of length 100ms and a fade out of 300ms. The silence audio clips have a fade in and fade out of length 20ms. This is done to make the audio file sound "smoother" and more natural. For consistency, we decided to keep these lengths constant across all conditions and all files rather than 'tailoring' them to each clip. Finally, because participants in the original study complained that the audio clips were hard to hear, we increased the volume of all audio clips by 6 dB.

Figure S1: Factor loadings for Study 1. Factor loadings for questions asked across all round robins. Factor 1 was our measure of "conversation enjoyment". The full questionnaire is in Appendix A.

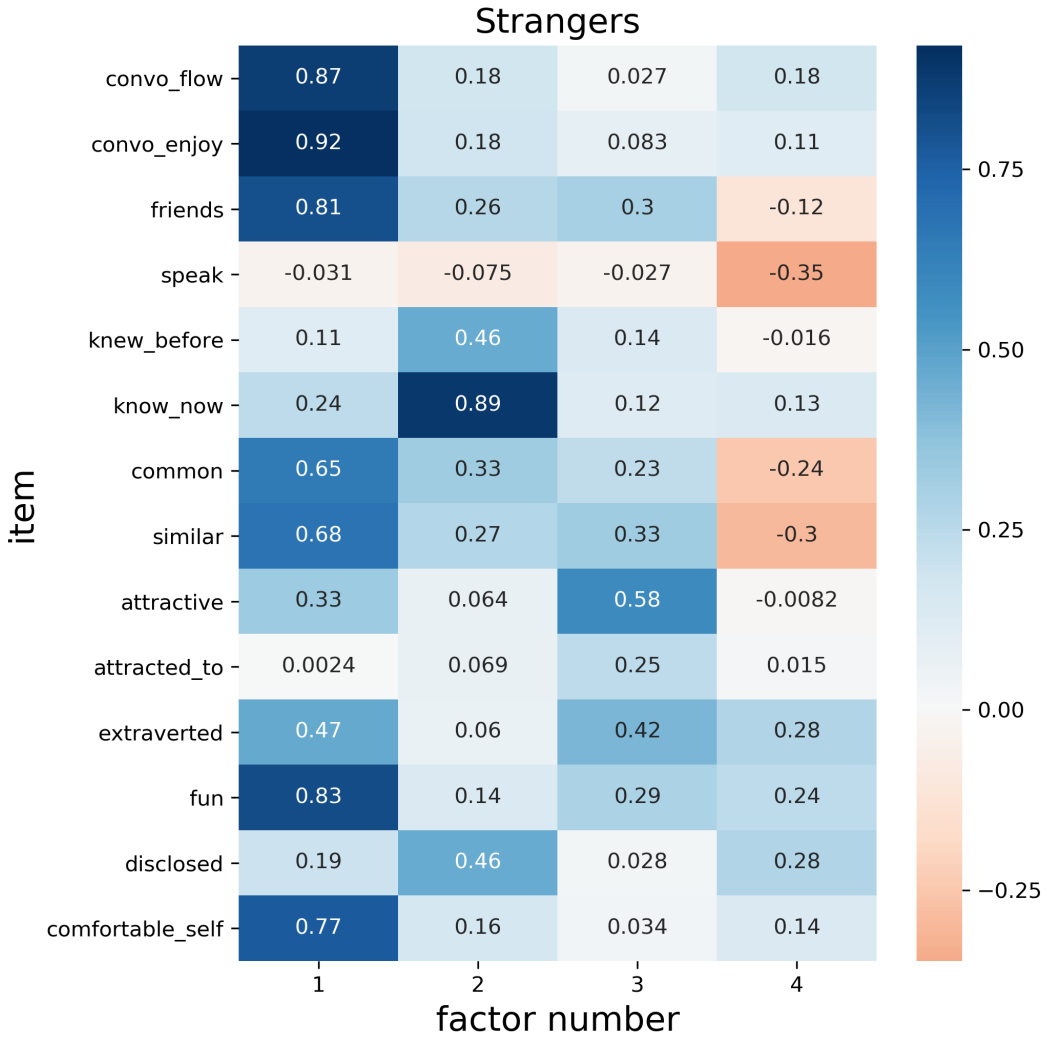

Figure S2: Factor loadings for Study 2. Factor loadings for questions answered after friend conversations. As in Study 1, the first factor maps onto "conversation enjoyment". The full questionnaire is in Appendix B.

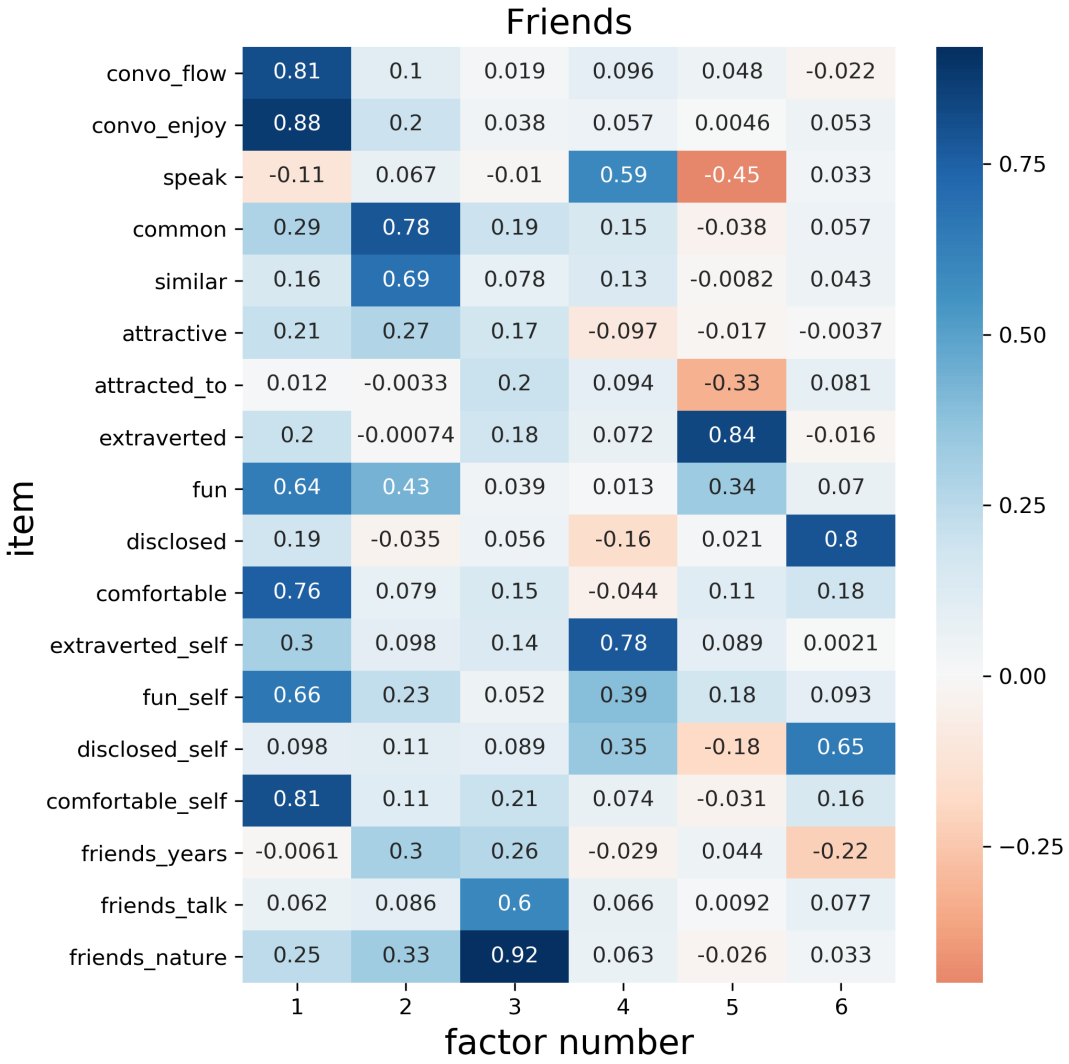

Figure S3: Robustness checks for within-conversation analysis (Study 1). First, how does the relationship between gap length and connection ratings within a conversation change by how time is binned? We plot the beta coefficients for this effect across a range of different bin sizes ("real data"). Every bin size yielded a significant ($p < .05$) effect of gap length on connection. Note that the larger the bin size, the greater the magnitude of the effect. However, as the bin size decreases the likelihood of missing data also increases. This is because we can only compute an average gap length for a given bin if a turn occurred in that bin. Second, do these estimates outperform an empirical null distribution? For each bin size, we generated surrogate data by randomly permuting the order of gap lengths within each conversation using a circle-shifting procedure and re-fitting the model predicting social connection 100 times. We plot the beta coefficients for these effects at each bin size, for each permutation ("circle-shifted data").

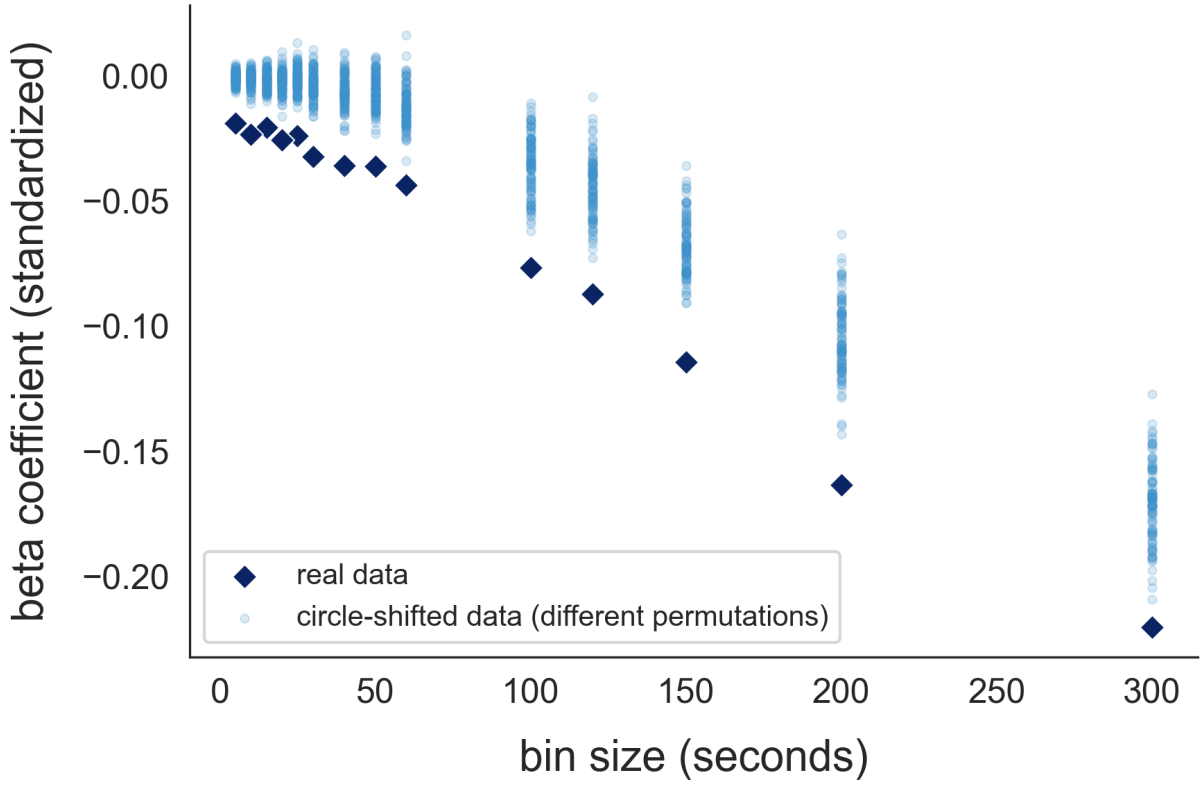

Figure S4: Self / partner effects across different bin sizes (Study 1). How does the relationship between self and partner gap length and connection ratings within a conversation change by how time is binned? We plot the beta coefficients for both self and partner effects across a range of different bin sizes. For every bin size, the partner effect is stronger than the self effect. The partner effect yielded a significant ($p < .05$) effect of on connection for every bin size.

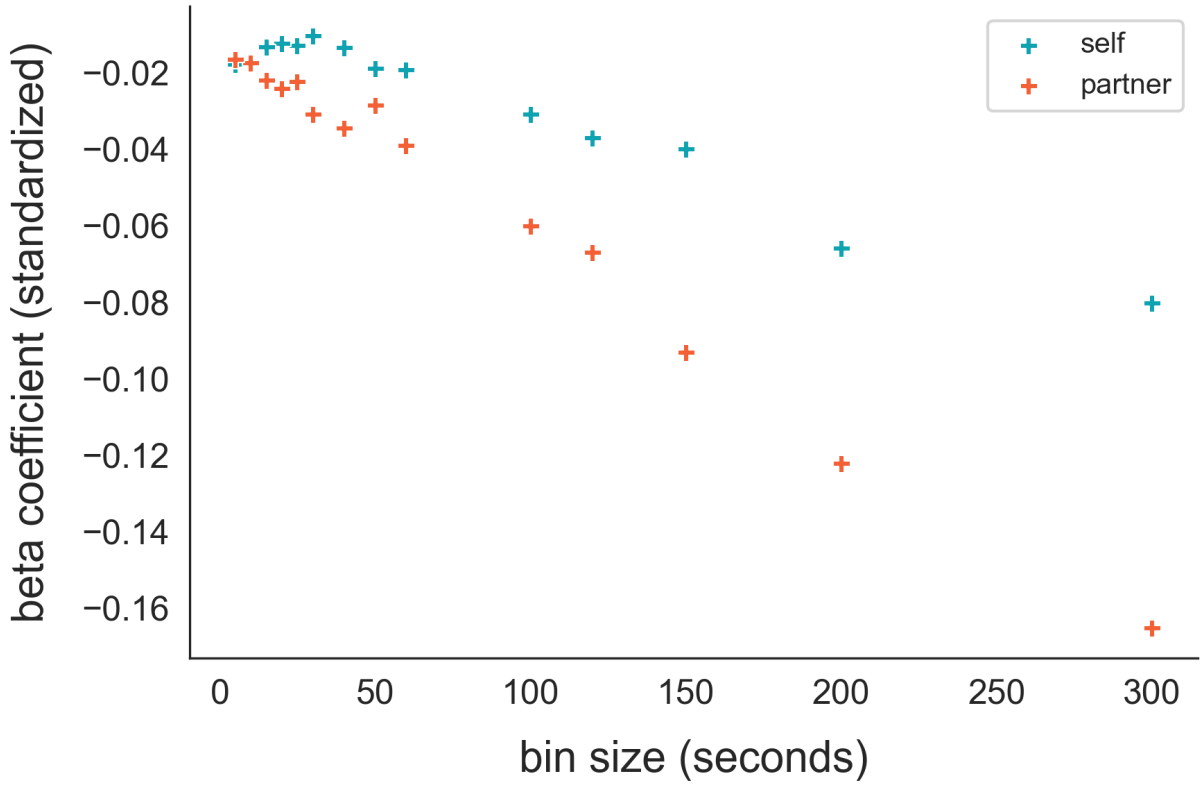

Figure S5: Robustness checks for within-conversation analysis (Study 2). First, how does the relationship between gap length and connection ratings within a conversation change by how time is binned? We plot the beta coefficients for this effect across a range of different bin sizes ("real data"). Every bin size except the largest one (300 seconds) yielded a significant ($p < .05$) effect of gap length on connection. In general, the larger the bin size the greater the magnitude of the effect. However, as the bin size decreases the likelihood of missing data also increases. This is because we can only compute an average gap length for a given bin if a turn occurred in that bin. Second, do these estimates outperform an empirical null distribution? For each bin size, we generated surrogate data by randomly permuting the order of gap lengths within each conversation using a circle-shifting procedure and re-fitting the model predicting social connection 100 times. We plot the beta coefficients for these effects at each bin size, for each permutation ("circle-shifted data").

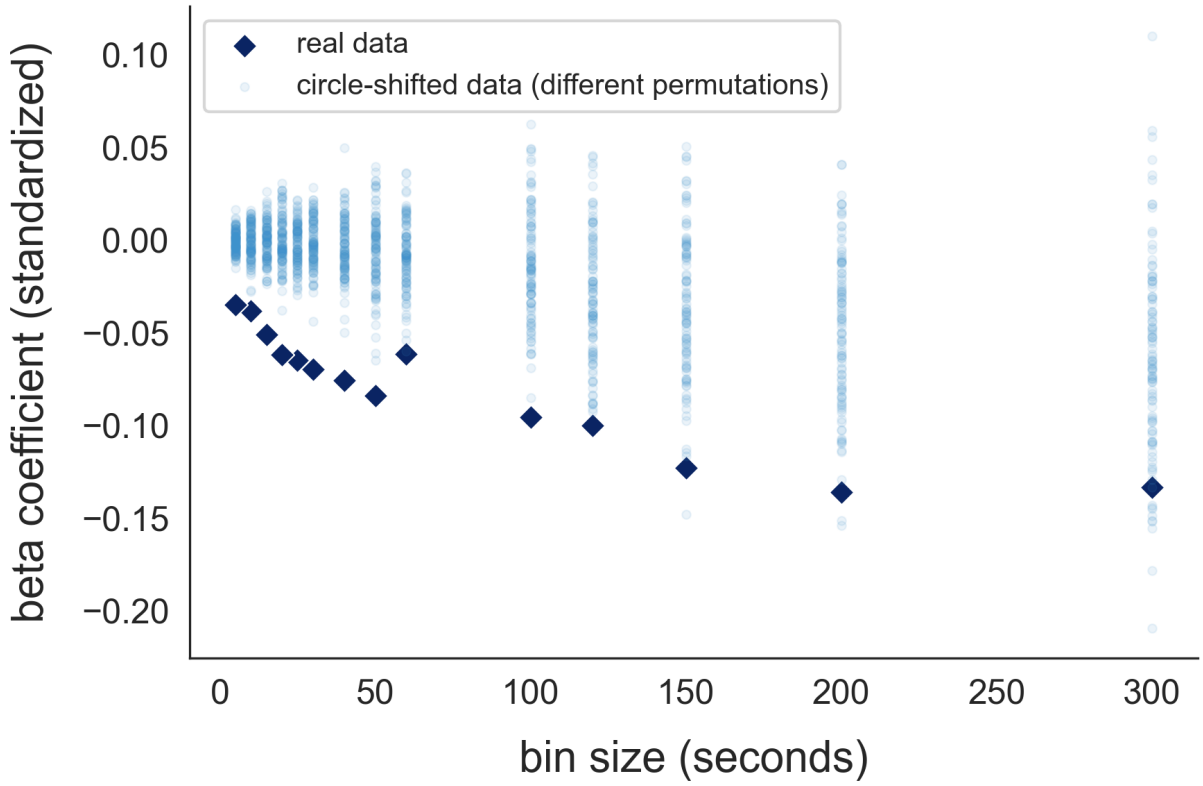

Figure S6: Self / partner effects across different bin sizes (Study 2). How does the relationship between self and partner gap length and connection ratings within a conversation change by how time is binned? We plot the beta coefficients for both self and partner effects across a range of different bin sizes. The partner effect is consistently stronger than the self effect for all bin sizes 60 seconds and less. The partner effect consistently yielded a significant ($p < .05$) effect of on connection for bin sizes 60 and less. The self effect consistently yielded a significant ($p < .05$) effect of on connection for bin sizes 40 and less.

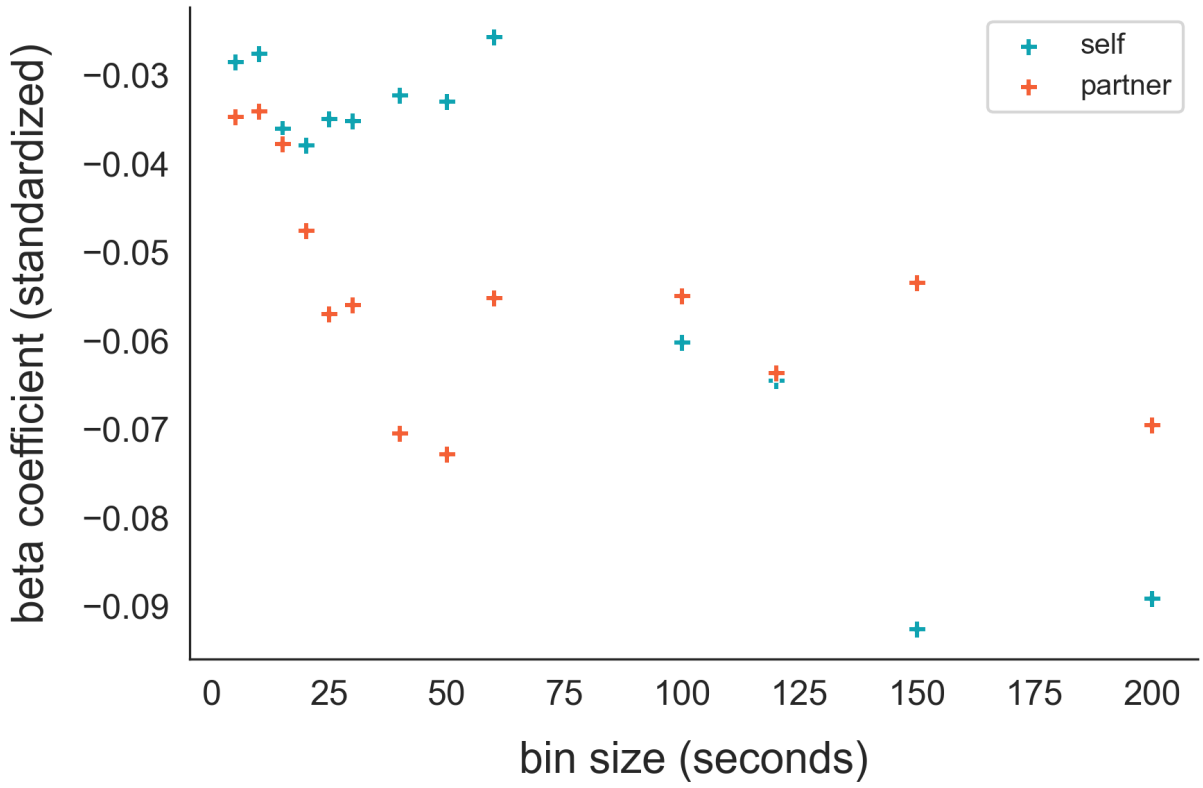

Figure S7: Replication of Study 3 results in a different sample. Main effect of condition (Short, Long) on (A) perceived conversation enjoyment and (B) connection. Effect of gap length condition on perceived (C) enjoyment and (D) connection broken down by conversation audio file. All values are centered within-subject to reflect the random effect structure used in the mixed-effects model. Error bars depict 95% confidence intervals.

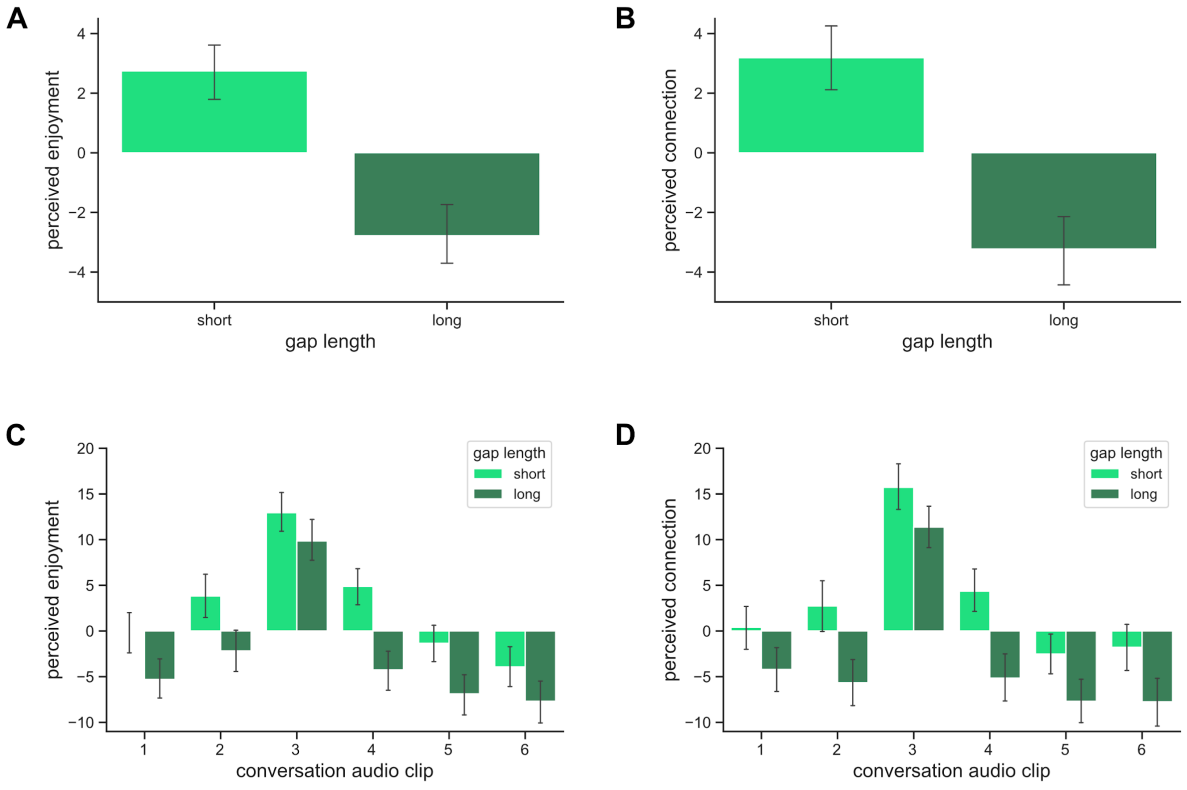

Table S1: Participants rated conversations with friends more positively than conversations with strangers. Differences in conversation ratings between stranger dyads and friend dyads. As expected, participants rate conversations with their friends more positively than conversations with strangers.

Variable	Mean Friends	Mean Stranger	Welch's t test
convo_flow	88.02	73.25	t(278.12)=10.22, p<.001 ***
convo_enjoy	87.95	72.55	t(251.28)=10.15, p<.001 ***
speak	54.22	53.51	t(168.03)=0.45, p=.66
common	78.64	53.18	t(202.48)=13.33, p<.001 ***
similar	66.65	53.12	t(169.83)=5.73, p<.001 ***
attractive	78.57	55.24	t(220.96)=13.09, p<.001 ***
attracted_to	24.02	8.62	t(149.10)=6.09, p<.001 ***
extraverted	70.25	59.45	t(190.47)=5.02, p<.001 ***
fun	82.57	68.20	t(245.52)=9.35, p<.001 ***
disclosed	55.78	44.16	t(162.19)=4.62, p<.001 ***
comfortable_self	88.86	73.86	t(232.85)=9.87, p<.001 ***

Appendix A

Post-conversation survey items for Study 1 (Strangers)

In this short survey, you will make ratings about the conversation you just had. Please answer the following questions about your experience as honestly and completely as possible. Your responses to these questions will be kept confidential and only identified by a numeric identifier, not your name.

1. How well did this conversation "flow"? *(0=Not at all, 100=Very)* **[**variable name = **convo_flow]**
2. How much did you enjoy the conversation you had with your study partner? *(0=Not at all, 100=Very much)* **[convo_enjoy]**
3. How much would you like to be friends with your study partner? *(0=Not at all, 100=Very much)* **[friends]**
4. Think about how much you and your study partner each talked during your conversation and indicate your relative contributions on the scale below *(0=My partner spoke much more than I did, 50=My study partner and I spoke the same amount, 100=I spoke much more than my study partner did)* **[speak]**

5. How well did you know your study partner before today? *(0=Not well at all, 50=Moderately well, 100=Extremely well)* **[knew_before]**
6. If you knew your study partner before today, in what capacity did you know them? *(free response)* **[knew_before_text]**
7. How well did you think you know your study partner now? *(0=Not well at all, 50=Moderately well, 100=Extremely well)* **[know_now]**

8. My study partner and I seemed to have a lot in common. *(0=Strongly disagree, 100=Strongly agree)* **[common]**
9. My study partner and I seemed to have similar personalities. *(0=Strongly disagree, 100=Strongly agree)* **[similar]**
10. My study partner is an attractive person. *(0=Strongly disagree, 100=Strongly agree)* **[attractive]**
11. I am physically attracted to my study partner. *(0=Strongly disagree, 100=Strongly agree)* **[attracted_to]**

12. My study partner seemed to be an extroverted person. *(0=Strongly disagree, 100=Strongly agree)* **[extraverted]**
13. My study partner was a fun person to talk to. *(0=Strongly disagree, 100=Strongly agree)* **[fun]**
14. My study partner disclosed a lot of personal information during our interaction. *(0=Strongly disagree, 100=Strongly agree)* **[disclosed]**
15. My study partner felt comfortable having a conversation with me. *(0=Strongly disagree, 100=Strongly agree)* **[comfortable]**

Please rate your agreement with the following statements, as they relate to the conversation you JUST HAD.

16. I was extroverted in that conversation. *(0=Strongly disagree, 100=Strongly agree)* **[extraverted_self]**
17. I was a fun person to talk to in that conversation. *(0=Strongly disagree, 100=Strongly agree)* **[fun_self]**
18. I disclosed a lot of personal information during that conversation. *(0=Strongly disagree, 100=Strongly agree)* **[disclosed_self]**
19. I felt comfortable having a conversation with my study partner. *(0=Strongly disagree, 100=Strongly agree)* **[comfortable_self]**

The conversation you just had was about 10 minutes long. Sometimes people feel ready for a conversation to end before it actually ends. Sometimes people don't feel that way.

Think back to your conversation. Was there a point in the conversation when you felt ready for it to end? Or do you wish it had gone on longer?

20. How do you feel about the length of the conversation you just had? *(0=I wish it had been much shorter, 50=It was exactly the right length, 100=I wish it had been much longer)* **[length_self]**
21. How do you think YOUR PARTNER felt about the length of the conversation you just had? *(0=They wish it had been much shorter, 50=They thought it was exactly the right length, 100=They wish it had been much longer)* **[length_partner]**

Notes about these survey items.

- Questions **1-14** and **19** were asked across *all* round robins and were therefore the questions that we entered into the factor analysis
- Questions **20-21** were included for a collaborator and were not analyzed by us
- Round Robin **1** answered questions: **1-14**, **19**
- Round Robins **2** & **3** answered questions: **1-19**
- Round Robins **4**, **5**, & **6** answered questions: **1-21**

Appendix B

Post-conversation survey items for Study 2 (Friends)

In this short survey, you will make ratings about the conversation you just had. Please answer the following questions about your experience as honestly and completely as possible. Your responses to these questions will be kept confidential and only identified by a numeric identifier, not your name.

1. How well did this conversation "flow"? *(0=Not at all, 100=Very)* **[**variable name = **convo_flow]**
2. How much did you enjoy the conversation you had with your friend? *(0=Not at all, 100=Very much)* **[convo_enjoy]**
3. Think about how much you and your friend each talked during your conversation and indicate your relative contributions on the scale below *(0=My partner spoke much more than I did, 50=My study partner and I spoke the same amount, 100=I spoke much more than my study partner did)* **[speak]**

Please rate your agreement with the following statements:

4. My friend and I have a lot in common. *(0=Strongly disagree, 100=Strongly agree)* **[common]**
5. My friend and I have similar personalities. *(0=Strongly disagree, 100=Strongly agree)* **[similar]**
6. My friend is an attractive person. *(0=Strongly disagree, 100=Strongly agree)* **[attractive]**
7. I am physically attracted to my friend. *(0=Strongly disagree, 100=Strongly agree)* **[attracted_to]**

Please rate your agreement with the following statements:

8. My friend seemed extroverted in that conversation. *(0=Strongly disagree, 100=Strongly agree)* **[extraverted]**
9. My friend was a fun person to talk to in that conversation. *(0=Strongly disagree, 100=Strongly agree)* **[fun]**
10. My friend disclosed a lot of personal information during our interaction. *(0=Strongly disagree, 100=Strongly agree)* **[disclosed]**
11. My friend felt comfortable having a conversation with me. *(0=Strongly disagree, 100=Strongly agree)* **[comfortable]**

Please rate your agreement with the following statements, as they relate to the conversation you JUST HAD:

12. I was extroverted in that conversation. *(0=Strongly disagree, 100=Strongly agree)* **[extraverted_self]**
13. I was a fun person to talk to in that conversation. *(0=Strongly disagree, 100=Strongly agree)* **[fun_self]**

14. I disclosed a lot of personal information during that conversation. *(0=Strongly disagree, 100=Strongly agree)* **[disclosed_self]**
15. I felt comfortable having a conversation with my friend. *(0=Strongly disagree, 100=Strongly agree)* **[comfortable_self]**

Please answer the following questions about the friend you just talked to.

16. How long have you been friends with them? *(0, 1, 2, 3, 4, 5yrs)* **[friends_years]**
    a. You indicated that you've known your friend for at least 5 years. If you've known them for LONGER than 5 years, please indicate that here: *(open response)* **[friends_years_extended]**
17. How frequently do you talk to this friend? *(0=Monthly, 50=Weekly, 100=Daily)* **[friends_talk]**
18. How would you characterize the nature of your friendship with this person? *(0=acquaintances, 25=friend, 75=close friend, 100=best friend)* **[friends_nature]**
19. Pick the gender that you most identify with: *(Female, Male, Other, Prefer not to answer)* **[gender]**

Appendix C

"Common Ground" as a mechanism?

We preregistered and ran another online study exploring a possible explanation for when short gap lengths occur: having common ground. We hypothesized that transcripts from conversations that had shorter gap lengths would be rated as having more common ground compared to transcripts from conversations that had longer speech gaps. We also included transcripts from friend conversations as (i) a sanity check (e.g., if participants don't rate conversations between friends as having common ground than strangers, perhaps the way we are asking our question isn't eliciting the responses we hope) and (ii) a benchmark (e.g., how "friend-like" can a stranger conversation be rated?).

All of the details of the study design can be found in the preregistration here: osf.io/4c96b.

We found no evidence that stranger conversations with short gaps are rated as having more common ground than stranger conversations with longer gaps. We did find evidence that participants were able to recognize common ground in transcripts of friend conversations. We did not pursue this question further. However, because we preregistered it as part of this larger project (examining the social role of gap lengths in conversation) we include it here.

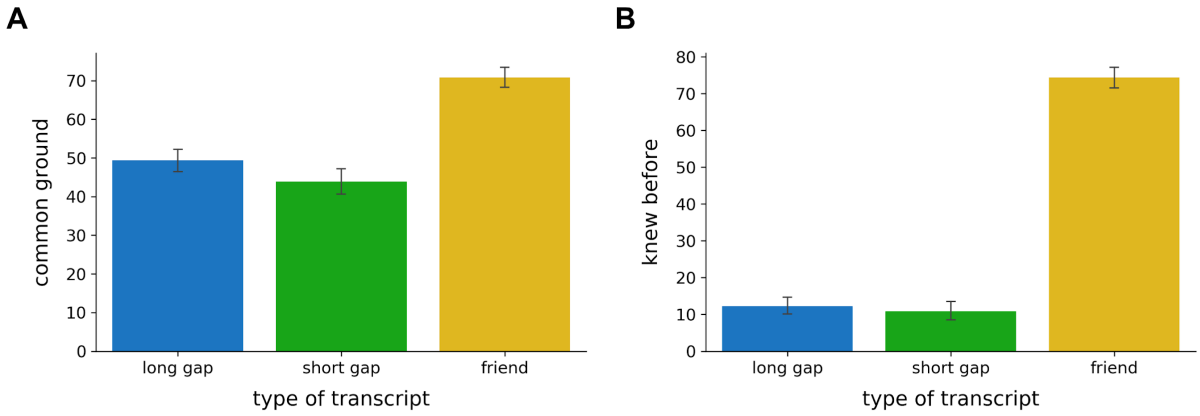

Figure S4: (A) Average responses to the question, *"To what extent did these two people find 'common ground' (e.g., talked about similar backgrounds, shared interests, mutual friends -- even if you couldn't exactly tell what they were talking about)"* rated on a sliding scale anchored by "Not at all" (0) and "A great deal" (100). The difference between long and short gaps is not significant. (B) Average responses to the question, *"How well do you think these two people knew each other before their conversation?"* rated on a sliding scale anchored by "Not at all" (0) and "Extremely well" (100). The difference between long and short gaps is not significant. Error bars depict 95% confidence intervals.