# Chapter 4

# Data Collection Methodology

In Chapters 5, 6, and 7 I will go into specific methodologies for each study. Before delving into the specifics of each study, though, this chapter first describes the characteristics of participants as well as the data collected, and then describes the overall methodology used for data collection.

The overall goal of my experimental setup was to emulate an online text chat environment, akin to what a person would encounter in chatting with online customer service, or engaging in a conversation on Slack. At the same time, behind the scenes, I aimed to collect not only the text appearing in the conversation, but the timing of every individual keystroke, from when a key was pressed to when it was released. This included not only visible keys such as letters and numbers, but also non-printing keys such as CONTROL or SHIFT. A screenshot of the experimental apparatus is provided in Figure 4.7. The individual components of the experimental setup will be explained in Section 4.4.3.

The experiments were run on the Prolific crowdsourcing platform, which is similar to Amazon's Mechanical Turk. However, Prolific has better quality control and is more respectful of participant privacy (Palan and Schitter, 2018). Participants were randomly paired, and pseudonyms were used to maintain anonymity. Further details are provided below in Section 4.3.

The experimental apparatus described below was used for data collection in all of the studies of my thesis. The entire experiment can be broken down into three phases:

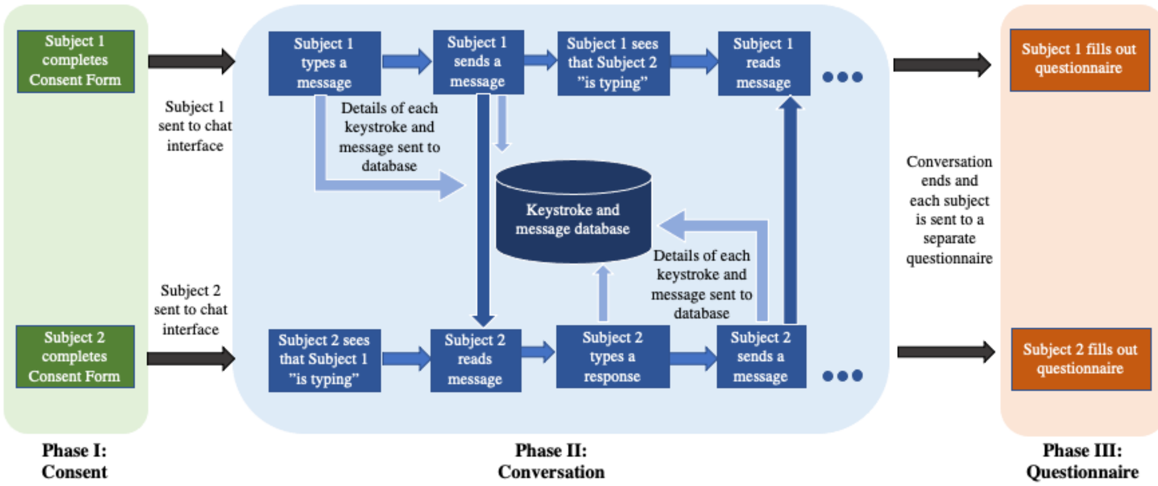


**Figure 4.1**

A map of all 3 phases of my data collection experiment. The horizontal lines show how the experiment proceeds for each participant. The lines pointing to the central database are intended to convey that all keystroke and message information is being sent simultaneously, in the background, while the conversation is taking place.

- Phase I: The initial advertisement and instructions for the experiment were sent to potential participants. Importantly, and as will be discussed in Section 4.4.1, the advertisement and initial consent did not mention keystrokes, but rather only that I was interested in collecting opinions about movies and television shows. The participants were then directed to a consent form, which followed IRB guidelines. Upon consenting, the participant was routed to the conversational apparatus, or the actual experiment.

- Phase II: This is the bulk of the experiment. The participant took part in a 16-minute conversation about movies and TV shows. In the first half, one participant recommended entertainment to the other participant; in the second half the participants switched roles. At the end of the 16 minutes, each participant was routed to an individual post-conversation questionnaire.

  Importantly, the term "message" in Figure 4.1 is not just the sum of individual keystrokes. For example, if a participant types "`A SPACE B A T DELETE DELETE E D ENTER`," then 10 individual keystrokes will be sent to the database. However, only the final message, "`A BED`," will be recorded as the transmitted message in the database (and seen by the other participant). This highlights an advantage of using keystroke data: both the production process and final product are preserved.

- Phase III: The participant rated multiple aspects of the conversation itself, as well as aspects of their partner (see Section 4.4.5 for details about the questionnaire). Upon completion of the questionnaire, the participant was routed to a final consent form that informed them as to the true nature of the experiment (i.e. studying keystroke patterns themselves), and the participant was asked to consent again, now with knowledge of the entire experiment. The participant was also provided with the option to contact the researchers if they were uncomfortable with this information being collected. None of the participants had further questions, nor did any withdraw consent after full disclosure.

## 4.1 Demographics

In total I collected 102 usable conversations, comprised of 204 participants. No participants participated multiple times in the experiment. Partners were assigned randomly: The order in which participants joined the experiment determined which room they were assigned to and they were paired with the participant who joined right before or right afterwards. All participants in my study were required to currently reside in the United States and be native English speakers.

In my final data, 119 participants identified as female, 84 identified as male, and 1 preferred not to say. Almost half of the conversations (49) involved partners who identified as different genders. 34 conversations took place between participants who both identified as female, while 17 took place between two participants who identified as male.

The average age was 34 years old, while the median age was 32. The youngest participant was 18, while the oldest was 72. As can be seen in Figure 4.2, even though age constraints were removed, the vast majority of participants still fell between 20-40.

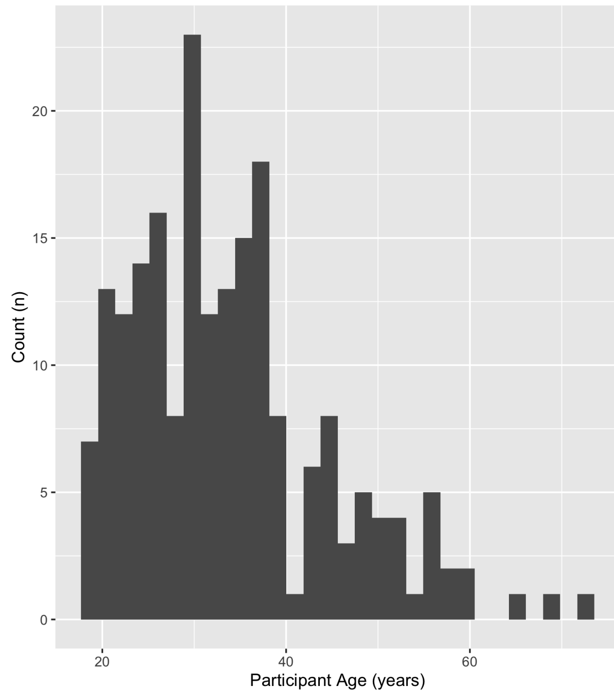


**Figure 4.2**
The overall age distribution of participants

The average age difference between partners in conversations was 12 years apart, with a median difference of 10 years apart. Figure 4.3 illustrates why removing age constraints did not materially affect the nature of the conversations.[1] The flat trendline strongly supports the notion that an increase in age difference did not impact the enjoyment ratings assigned by partners. In other words, a conversation between two 25-year-old partners was, on average, roughly as enjoyable as a conversation between a 25-year-old participant and their 40-year-old partner.

The majority of participants, 117 or 57%, have obtained an undergraduate degree or are in the process of obtaining an undergraduate degree. 58 participants have (only) a high school education, and 29 participants have or are completing graduate school (including 2 PhDs). Of those participants with current data as to their student status, 130 (64%) are no longer students while 41 are currently students.

[1] A smaller version of this analysis was run after the pilot study, which led to the decision to remove age constraints.

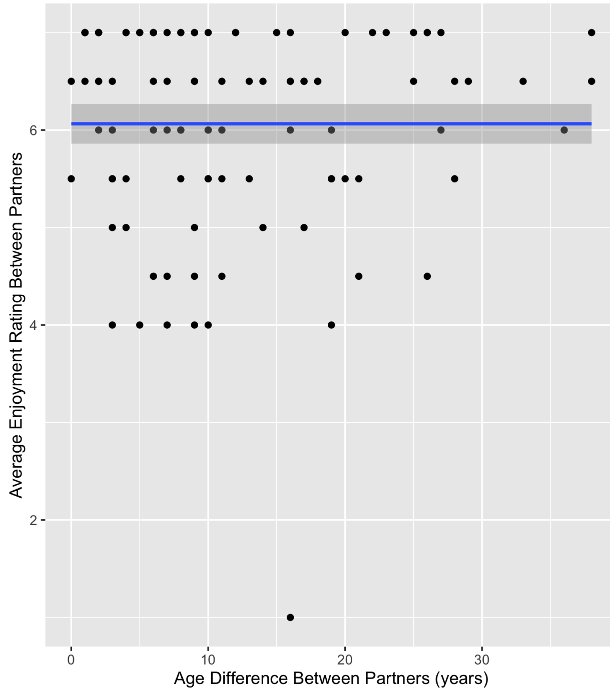


**Figure 4.3**
The average enjoyment rating between partners, as a function of the age difference between them.

Table 4.1 provides a detailed breakdown of employment status. Since the financial motivations underlying crowdworkers are important, it also seems important to highlight that the motivations for the crowdwork in my experiments were likely heterogeneous and of varying neediness.

Finally, the participants had familiarity with Prolific, and had also successfully completed a large amount of previous experiments. In this way, I ensured that typing patterns were not due to lack of familiarity with the platform, or due to a participant simply trying to rush through the experiment.

| Employment Status | *n* |
|---|---|
| Full-Time | 78 |
| Part-Time | 35 |
| Unemployed (and job seeking) | 22 |
| Not in paid work (e.g. homemaker, retired or disabled) | 19 |
| Due to start a new job within the next month | 4 |
| Expired data or Other | 46 |

**Table 4.1**
Employment status of participants

the average platform approval rating for participants was 99%. The lowest score was 96%, and 168 (82%) participants had a 100% approval rating. The average participant had completed 617 studies, with a median of 511 studies. Only 14 participants had completed less than 100 studies, while 27 had completed more than 1,000 studies.

As a final note, the demographics listed above are the only meaningful statistics released by Prolific. Prolific takes privacy very seriously, and does not release demographic information that could allow for a participant to be identified. Moreover, none of my experiments use or control for any demographics, as this was outside the scope of my studies. However, prior studies have shown that keystrokes can predict demographics such as gender or education level (Cascone et al., 2022; Tsimperidis and Arampatzis, 2020); therefore it is important in future studies to control these other factors in order to independently understand keystroke patterns.

## 4.2 Summary of Collected Conversation Data

The final experimental data was reviewed to ensure that anonymity was maintained and that participants remained engaged throughout the conversation. After all of the collected data was sanitized in this manner, it comprised 102 conversations, 4,895 messages, and 355,408 individual keystrokes. The average conversation contained 48 messages, with a median conversation length of 42 messages. A distribution of the conversation lengths can be seen in Figure 4.4.

The average message was made up of 10.7 words, with a median message length of 8 words. The distribution of different message lengths can be seen in Figure 4.5.

Interestingly, the message count between each partner did not differ significantly; however the *word* count between partners was strongly dependent on the role of the participant, as it related to whether they were providing or receiving recommendations. To verify this, a Welsh Two Sample t-test showed that the difference in the number of messages was almost non-existent ($p = 0.94$); on the other hand, the average word count of a recommendation provider's message was 11.1 words while a recommendation receiver's average message was only 10.3 words ($p < 0.01$). The difference

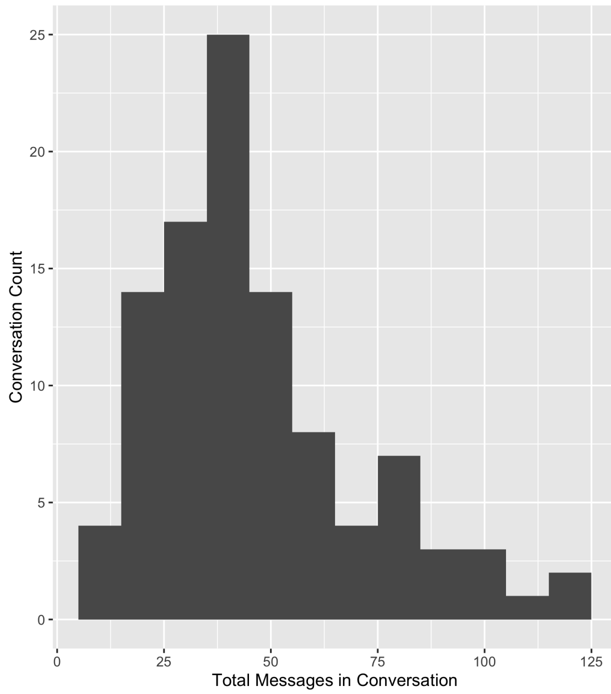


**Figure 4.4**
Distribution of conversation lengths, as measured by the total number of messages sent by both participants.

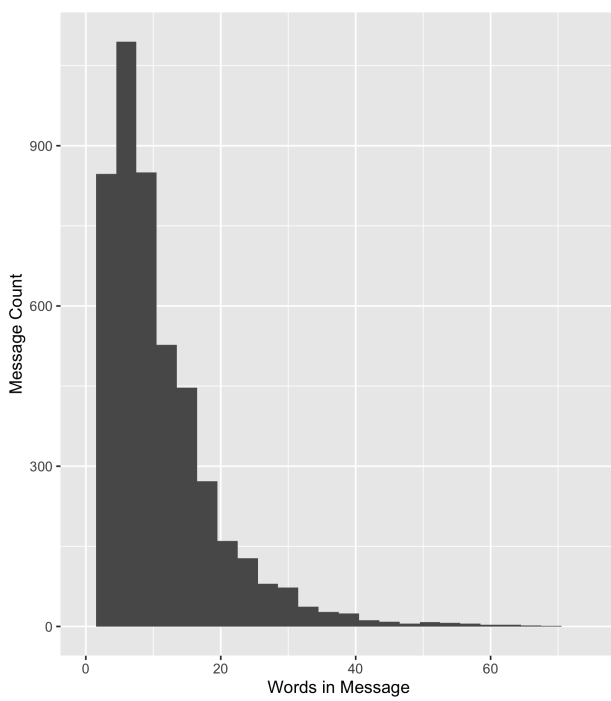


**Figure 4.5**
Distribution of message lengths, as measured by the number of (whitespace-delimited) words.

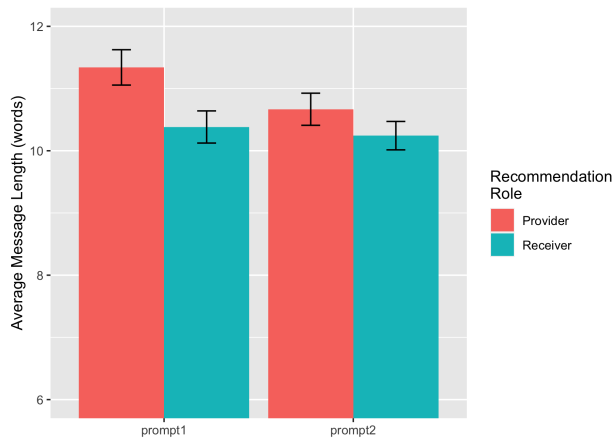


**Figure 4.6**
The average word count for each role type, as role relates to providing or receiving recommendations.

is illustrated in Figure 4.6, and broken up by prompt to illustrate that the differences existed in both halves of the conversation.

This will be expanded upon in Chapter 7, but it is interesting to note how this conforms with Herbert Clark's notion that language is a form of "joint action" (Clark, 1996), where every request or question from a recommendation recipient is responded to with a reaction or answer. On the other hand, as shown in Study 3, while the quantity of messages is equivalent, the contents of those messages as well as production patterns differ significantly.

## 4.3 Prolific Crowdsourcing

Because of the COVID-19 pandemic, collecting data in a traditional manner, such as having university students visit a lab for in-person experiments, was unfeasible and unsafe. Therefore, it was decided to use an online crowdsourcing platform instead. For my data collection I decided to use Prolific (Palan and Schitter, 2018) to run my experiments, rather than more popular crowdsourcing platforms such as Amazon Mechanical Turk (MTurk, Paolacci et al., 2010).

The benefits of Prolific over Amazon's Mechanical Turk are numerous. See Appendix B for an enumeration of Prolific's benefits. Studies such as Peer et al. (2022) demonstrate that the average data quality on Prolific is higher than on similar platforms, and users of Prolific are less likely to be using it as their primary source of income, which could make them less desperate to complete a task quickly.

## 4.4 Experimental design details

### 4.4.1 IRB Approval

All experiments were approved by the Northwestern Institutional Review Board (IRB) before any data was collected for analysis. The IRB ruled the study exempt from further review, based on two qualifications: the experiments only involve “tests, surveys, interviews, or observation,” and are only “benign behavioral interventions.” Both of these are considered low-risk by the IRB. Further, all participants were required to be located in the United States, so that there was no conflict with international privacy laws such as the EU’s General Data Protection Regulation (GDPR).

A critical element of my experiment was that keystrokes were collected without participants knowing that I was also recording them at this level of granularity. This was important because I wanted to capture “naturalistic” conversations and language production, without participants feeling self-conscious of the way they were typing. Because of this, my experiment was considered to be deceptive, or at least providing "incomplete information" at the beginning. The advertisement for the experiment is reproduced in Figure A.1. Participants were initially led to believe that the purpose of the study was to understand why people prefer or disprefer certain movies, genres, etc., and what their rationale was. No mention was made before the experiment that the study was also investigating keystrokes and typing patterns.

However, keystroke data is private information; after the experiment we disclosed the full objectives of the study, and provided the participants the option to withdraw from the study but still receive full compensation. In actuality, none of the participants chose to withdraw consent after full disclosure.

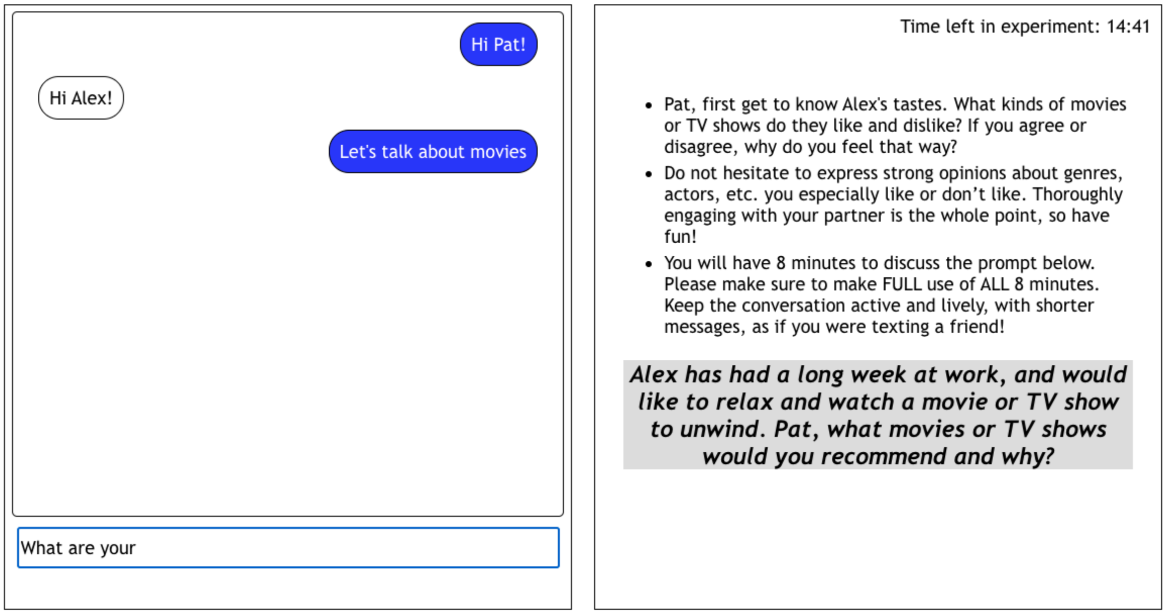


**Figure 4.7**
The experiment chat interface. This apparatus or web interface is what participants were viewing in Phase II of my experiment (see Figure 4.1).

This method of consent was also approved by the IRB. For further details of the IRB approval, see Appendix A.

### 4.4.2 Experiment Design Iterations

The experimental apparatus described in Section 4.4.3 is the finalized version that was used for data collection in my thesis. However, prior to this in both peer-testing and a pilot study, the experimental setup was improved through an iterative design process. Using feedback from both colleagues as well as participants in the online pilot study, adjustments were made to the timer displayed during the experiment, age constraints of the participants, and the wording of the experimental prompts. Appendix D provides the rationale for these changes and how they improved the data being collected.

### 4.4.3 Experiment Apparatus

The experiments were hosted on an Amazon Elastic Compute Cloud (Amazon EC2) instance. Because of the small footprint of the experimental apparatus, I was able to use a free-tier `t2.micro` instance, with only 1 GB of memory, running on Amazon's own Linux distribution.[2]

The actual experimental apparatus was written in JavaScript, HTML, and CSS. It was mostly built around a React (Meta Platforms, 2022) and `socket.io` framework (Rauch, 2013), in order to display, transmit and record every keystroke as soon as it was activated. The experiment worked on almost every major browser. The only exception, perhaps because of a CSS issue, was that the experiment did not work properly on Firefox. Participants were informed of this before starting the experiment.

The backend of the experiment automatically created new chat rooms for every two participants. In other words, when the first participant joined they were automatically routed to Room 1. The second participant was also routed to Room 1. When the third participant joined they were automatically routed to Room 2.

As seen in Figure 4.7, the experiment had two main components. On the left-hand side was the chat interface. This operated similar to a messaging interface on a phone, where the participant could see their own text in real-time as they entered it. In my interface, autocorrect, autocomplete and spellcheck were all turned off. Turning off autocomplete prevented participants from "entering" a long text string just by clicking their mouse. Turning off autocorrect meant that only the exact text entered would be transmitted, rather than a corrected version of the keystrokes entered. Finally, since spellchecks can differ from browser to browser and person to person, eliminating this option removed this variable from the experiment.

Once a participant pressed ENTER, the text was transmitted to their partner and displayed in the window on the top left. Similar to many text-based messaging interfaces, an "is typing..." indicator was added, so that participants would know when their partner was active. Importantly, the "is

[2] All of the code is publicly available. The front-end code is available at https://github.com/angoodkind/dialogue-keystrokes; the back-end code is available at https://github.com/angoodkind/keystrokes-collablab-backend.

typing“ indicator had a lag of two seconds, so that if a participant stopped typing for a moment, the indicator would not continuously disappear and reappear. The exception to this, though, was that the indicator would disappear as soon as a message was transmitted. This was important so that reaction times to messages would not be affected by a participant thinking their partner was still typing, and therefore waiting to respond.

The right-hand side displayed the conversation prompts as well as a countdown timer. The details of the prompt texts is discussed in 4.4.4. The experiment included two prompts so that in the first prompt one participant could provide recommendations, and in the second prompt that participant would be the receiver of recommendations. The timer on the top kept track of how much time was remaining in the experiment, so that the participants knew when they needed to wrap up.

Upon joining the room, participants were assigned the name either Pat or Alex. The purpose of assigning names was to help protect anonymity while still allowing each participant to refer to the other by their “name.” The names Pat and Alex were chosen because they are relatively gender-nonspecific, at least within the United States.

After each prompt was displayed for eight minutes, the experiment ended and the participants were automatically redirected to a post-experiment questionnaire, as well as a full disclosure about the experiment (see Section 4.4.1).

All keystrokes and transmitted messages were sent to a Firebase Real Time Database (Moroney, 2017). For every keystroke the database logged details of the key, e.g. not only that SHIFT was used, but whether it was the SHIFT key on the right-hand or left-hand side of the keyboard. The database also logged the exact time, using millisecond-level UNIX timestamps, when the key was pressed and released. For each transmitted message, the exact message as it was send to a partner was recorded, along with the time it was transmitted (approximately equal to when the ENTER key was pressed).

### 4.4.4 Conversational prompts

As mentioned previously, the entire experiment consisted of two prompts. The reasoning behind this was that in the first prompt, Subject 1 would recommend movies and TV shows to Subject 2, and in the second prompt Subject 2 would recommend movies and TV shows to Subject 1. This balancing between recommendation provider or recipient was to ensure that each participant had a chance to lead the conversation and respond to the conversation, assuming different "roles" in each section. If this balance had not been taken, then a possible confound in the experiments would be that a certain typing pattern is specific to one role (recommendation provider or recipient) rather than a behavior related to the overall social dynamics that I am interested in studying.

The topics of movies and TV shows were chosen because they are almost universal within the United States. While every person does not watch the same amount of TV and movies, it seems safe to assume that by age 18 almost every person has watched a fair amount of entertainment. Although I did not pre-screen for this or ask participants about this, this will be added in future iterations of my experiment so that viewership frequency can be factored in.

Some dialogues conveyed that one participant knew a lot about movies or TV and was possibly an expert in the subject. For example, one participant said of their partner, "my partner was well versed and knowledgeable about [movies and TV shows]," whereas another participant said "[my partner] did not have the necessary information to provide accurate recommendations." This variation is to be expected and is an accurate reflection of everyday encounters where people interact with others of varying subject matter expertise.

The two conversational prompts were:

| **Prompt 1** |
| --- |
| Alex has had a long week at work, and would like to relax and watch a movie or TV show to unwind. Pat, what movies or TV shows would you recommend and why?<br>Pat, first get to know Alex's tastes. What kinds of movies or TV shows do they like and dislike. If you agree or disagree, why do you feel that way? |

**Prompt 2**

Pat is bored, and would like to watch a really thought-provoking or stimulating movie or TV show. Alex, what movies or TV shows would you recommend and why?

Alex, now get to know Pat's tastes first. What kinds of movies or TV shows do they like and dislike? If you agree or disagree, why do you feel that way?

**Both prompts**

You will have 8 minutes to discuss the prompt below. Please make sure to make FULL use of ALL 8 minutes. Keep the conversation active and lively, with shorter messages, as if you were texting a friend! Do not hesitate to express strong opinions about genres, actors, etc. you especially like or don't like. Thoroughly engaging with your partner is the whole point, so have fun!

The types of movies to be recommended were also changed between prompts in order to help the conversation stay dynamic. If the movie types were the same in both prompts, then it is possible that the second half would simply become a reflection of the first half, where Subject 2 simply recommended the same entertainment that Subject 1 had recommended in the first half.

Moreover, as stated in the experiment's instructions, discord was also encouraged. Although most conversations stayed positive, I also wanted to see how a person's typing behavior would change when reacting to a statement with which they strongly disagreed.

Finally, in order to maintain the conversational nature of the experiment, participants were encouraged to keep messages short. In my pilot study, participants would sometimes type paragraph-length messages where they stated all of their likes and dislikes, or all of the movies they would recommend. Since the goal of my overall thesis, though, is to look for typing analogs to conversational speech, I wanted the typed dialogues to more closely mimic the short, spontaneous utterances produced during a face-to-face conversation.

### 4.4.5 Post-experiment Questionnaire

Upon completion of the conversational phase of the experiment, participants were redirected to a questionnaire about the conversation in which they participated.

These questions were based on multiple related prior studies. Liebman and Gergle (2016a) and Liebman and Gergle (2016b) also studied human-human text-based dialogue, in which two participants had to resolve a moral dilemma. Similar to my own studies, Liebman and colleagues were also investigating the nature of computer-mediated communication. My questions were also based on those from Pecune et al. (2019), which investigated human-AI interactions for movie recommendations. Since Pecune et al. (2019) was also interested in what made a recommendation appealing, many of its questions were relevant to my own studies.

For the questions below, participants had to provide ratings from 1-7, based on a Likert scale (Joshi et al., 2015). The final two questions were open-text responses.

1. To what degree did you enjoy the conversation? [1=Not at all, 7=I enjoyed it a lot]
2. To what degree do you think your partner enjoyed chatting with you? [1=Not at all, 7=They enjoyed it a lot]
3. To what degree did the conversation go smoothly? [1=Not smooth at all, 7=Very smoothly]
4. Hypothetically, how much do you think you'd enjoy watching a movie with your partner? [1=Not at all, 7=I would definitely enjoy it]
5. How would you rate the level of rapport established between you and your partner? [1=No rapport, 7=Lots of rapport]
6. How likely do you think it is that you'll end up watching one of the movies your partner recommended? [1=Not likely at all, 7=Very likely]
7. In a few sentences, how would you describe your partner and the overall conversation?
8. Do you have any additional comments or questions for the study authors? [optional]

The responses to these questions were skewed very positively. Figure 4.8 shows this distribution. The studies in my thesis will discuss these distributions in more detail, but it seems that most subjects had high opinions of their partner as well as the overall conversation.

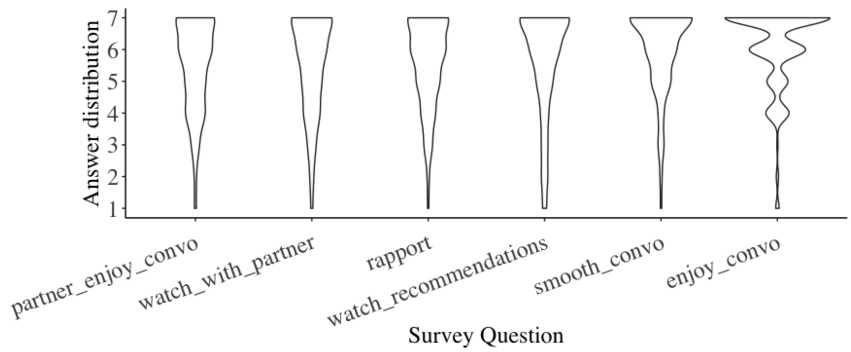


**Figure 4.8**
The distribution of responses to survey questions. All responses were heavily skewed towards positive ratings.

## 4.5 Keystroke Collection

From a methodological standpoint, it is also necessary to mention the nature of the keystroke timestamps that I collected. The majority of prior keystroke experiments were conducted in a lab setting, often with a desktop computer (e.g. Brizan et al., 2015; Killourhy and Maxion, 2012; Kuzminykh et al., 2020). This is done both to reduce latency between a keystroke and when that keystroke appears, as well as to establish consistency between each run of an experiment.

In my case, however, there existed variation in computer type (laptop, desktop, tablet with a physical keyboard), browser choice (Chrome, Safari, etc.), and internet connection (slow, fast, wired, and wireless). Future studies will pinpoint the differences between browsers and connections. In addition, the dataset released alongside my thesis includes details about the computers each participant used, so that participants can be partitioned in future studies.

### 4.5.1 Collected Features

The raw features sent to each of the database are enumerated in Tables 4.2 and 4.3. For the keystroke database, all information was collected for every individual keystroke. Although a separate entry was recorded for the key-press and key-release, these were immediately combined in the database

Keystroke Database

| Feature | Description | Example |
|---|---|---|
| Experiment ID | An identifier for the session with two participants | E001 |
| Subject ID | An identifier for the individual participant. This was used as the key to link keystrokes to personal demographics and questionnaire answers. | S001 |
| Prompt Number | Which prompt the keystroke occurred within, so that I can reference whether the participant is the recommendation provider or receiver | Prompt2 |
| Utterance ID | An identifier for the utterance within which the keystroke occurs. This was used as the key to link keystrokes to messages. | E001-S001-1 |
| Raw Keystroke Char | A raw representation of a keystroke. For example, even if SHIFT was being held down, the E-key would be recorded as e rather than E. In addition, non-printing keys such as SHIFT would be recorded as such. | b |
| Visible Keystroke Char | The visible representation of a keystroke. A capital letter that is the result of a held-down modifier key would be recorded as such. Non-printing keys receive a null entry. | B |
| Key-press Time | A UNIX timestamp recorded when a key was pressed down | 1642087297175 |
| Key-release Time | A UNIX timestamp recorded when a key was released up | 1642087297278 |
| Existing Text In Message | The text that currently exists in the participant's textbox, that has not been sent yet. | "A " |

**Table 4.2**
In the schema for the keystroke database, imagine that the participant is typing the B in the message "A Bee"

into a single entry. For the message database, all information was collected for every individual transmitted message.

## 4.6 Engineered Features

The studies in this thesis all require engineered features beyond the raw features that were collected during data collection. A complete list is provided below. The tables are broken down by engineered features of each keystroke, an individual word, a complete message, a participant overall, and an entire conversation. Each study/chapter will also make mention of the specific features used for that particular study.

Message Database

| Feature | Description | Example |
| --- | --- | --- |
| Experiment ID | An identifier for the session with two participants | E001 |
| Subject ID | An identifier for the individual participant. This was used as the key to link keystrokes to personal demographics and questionnaire answers. | S001 |
| Prompt Number | Which prompt the keystroke occurred within, so that I can reference whether the participant is the recommendation provider or receiver | Prompt2 |
| Utterance ID | An identifier for the utterance within which the keystroke occurs. This was used as the key to link keystrokes to messages. | E001-S001-1 |
| Sent Text | The full message transmitted | A Bee |
| Time Sent | A UNIX timestamp recorded when the message was transmitted | 1642087297278 |

**Table 4.3**
In the schema for the message database, imagine that the participant has sent the message `"A Bee"`

Each of the features in Table 4.4 was measured for every individual keystroke. This includes both printing and non-printing keystrokes, as well as printed characters that were deleted and not in the final transmitted text.

Since my thesis also considers linguistic context, I also calculated features based on the complete word within which each keystroke took place. These engineered features are enumerated in Table 4.5.

The features in Table 4.6 were applied at the utterance level. More specifically, the features only apply to utterances that were transmitted and therefore viewable by both partners.

The features in Table 4.7 were calculated for each participant. These features therefore reflect overall traits of a participant, rather than features that are specific to a single utterance or even subset of utterances.

Finally, the features in Table 4.8 are applied to the entire conversation. Some features reflect the differences between participants, but because these are contrastive features they only apply to the pair of partners rather than an individual.

| Feature | Description |
| --- | --- |
| Time Since Conversation Began (minutes) | The conversation is considered to have "began" when the first user transmits their first message |
| Interval Between Keystrokes | The time gap between when the previous key was released and this key was pressed. Cam be negative in the case of SHIFT being held while a letter is entered. |
| Keystroke Duration | The time elapsed from when this key was pressed to when it was released |
| Flight Time | The time between when the previous key was pressed to when this key was pressed. This will always be positive and can be more reliable. |
| Time Since Utterance Start | The time between when the first key in the utterance was pressed to when this key was pressed |
| Position in Word | The keystroke's position relative to the entire utterance. Position $\in$ Beginning of Utterance, End of Word, Space, Beginning of Word, End of Utterance |
| Within Word | The word within which this keystroke occurs. Words are delimited by spaces, and may include nonsense words |
| Within Keystroke Sequence | The space-delimited sequence within which this keystroke occurs. This sequence will includes keystrokes such as DELETE |

**Table 4.4**
Individual keystroke-level features

| Feature | Description | Example |
|---|---|---|
| Word Length (characters) | The number of printed characters in a word | `Bee` = 3 |
| Lemma | The lemmatized version of a word, or the "canonical" form of a word. This allows for comparison of the same root word but in different forms, e.g. past tense or pluralized. Lemmatization was performed using the `hunspell` package in R (Ooms, 2022). | `drove` = `drive` |
| English Word | Checks whether the word is a valid English word, using the `hunspell` English dictionary (Ooms, 2022) | `drive` = TRUE<br>`droive` = FALSE |
| Semantic Category | Divides words into Function and Content words, where function words are words with a grammatical function, such as determiners and conjunctions and content words have specific meanings, e.g. nouns and verbs (Pennebaker et al., 2003) | `dog` = CONTENT<br>`the` = FUNCTION |

**Table 4.5**
Features added to each word, which is a space-delimited character segment.

| Feature | Description |
|---|---|
| Continuation | Denotes whether this message is a continuation of a turn, where the same participant sent the previous message as well |
| Utterance Length (words) | The number of space-delimited words in a transmitted utterance |
| Time Delay After Previous Message Received | The interval between when when a message was transmitted and composition of the next message begins. For a message that is a continuation of a turn by the same participant, this number will always be positive, i.e. a person cannot begin typing a new message before the current message is transmitted. For different participants, this interval can be negative if a participant begins composing a message before their partner has transmitted a message. |
| Sentiment Score | For each transmitted utterance, sentiment scores were calculated using the `VADER` package in R (Hutto and Gilbert, 2014). A score above 0 denotes positive sentiment, while a score less than 0 denotes negative sentiment. Neutral sentiment (score = 0) is also possible |
| Sentiment Difference | The difference between the sentiment score of this utterance and the previous utterance |

**Table 4.6**
Features added to each full transmitted message

| Feature | Description |
|---|---|
| Overall typing rate | The elapsed time of all utterances the participant typed, as opposed to the elapsed time of the entire conversation, divided by the total number of keystrokes |
| Intra-word typing rate | The average interval between each keystroke within a word, rather than before or after the word. This metric tends to be a more accurate measurement of motor-based typing ability as opposed to language skills (Logan and Crump, 2011). |
| Inter-word typing rate | The average pause time before a word is typed, or the interval between a SPACE and the first letter of a word. Unlike intra-word typing rate, this rate is a more accurate measurement of lexical recall, i.e. retrieving a word from memory (Logan and Crump, 2011). |
| Average Length of Utterances | The average length of each utterance, with separate features for the number of words and the number of keystrokes |
| Edit Rate | The average rate of BACKSPACE or DELETE keypresses per visible keystrokes |
| Average Pause Before Responding | For non-overlapping messages, when a participant responds to the other participant's message, this is the average gap between receiving a message and initiating a reply |
| Questionnaire Answers | These are summarized in Section 4.4.5 |

**Table 4.7**
Features calculated for a participant overall.

| Feature | Description |
|---|---|
| Conversation Length | Three separate features: the number of utterances, words, and characters in a conversation |
| Average Utterance Length | Two separate features: the number of words and characters in each utterance |
| Ratio of turns | A measure of how equally the participants produced contributions: the ratios of utterances, words, and characters |
| Average Questionnaire Scores | The average rating that each partner assigned to the other |
| Age Difference | The differences in ages between partners |

**Table 4.8**
Features calculated for an entire conversation.