What the heck are we talking about?
Topic selection in conversation

Michael Yeomans & Alison Wood Brooks
Harvard University

September 7, 2018
*** PRELIMINARY DRAFT - - DO NOT DISTRIBUTE ***

Abstract: Conversation is ubiquitous, important, and uniquely human. It is composed of turns, during which people reason about and manage the topics of conversation. Is this topic enjoyable for me? For my partner? Should we stay on this topic or switch to a different one? Descriptively, how do people navigate this decision? Prescriptively, how should they? In this paper, we develop a topic selection framework, and across thousands of synchronous and asynchronous conversations in both face-to-face and online settings, we uncover how people navigate topic-switching in cooperative conversation. Compared to a wide array of natural language processing algorithms, humans fall short in detecting others' interest in topics, and they are overly reluctant to switch topics—with straightforward improvements that can be achieved by empowering conversationalists to switch topics more readily.

Disclosure. None of the authors have any potential conflicts of interest to disclose in relation to this research. For each study, we report how we determined our sample size, all data exclusions, all manipulations, and all measures. All data, analysis code, stimuli, and preregistrations from each study are available as Online Supplemental Material, stored at https://osf.io/t4h3z/

Keywords: Decision-Making; Natural Language Processing; Discourse Modeling

Introduction

Conversation— turn-by-turn linguistic interaction—is one of the most common and important human behaviors. People reveal their preference to converse over and over again in life—*we talk to each other all the time*—and every human relationship can be understood as a sequence of conversations. Historically, language emerged as a way for people to coordinate in their pursuit of common goals (Enquist & Leimar 1993), and large parts of the human brain are devoted to the efficient computation of listening to others, understanding what they are saying, and generating timely responses (Pickering & Garrod, 2004a; 2004b). Many aspects of modern human flourishing depend on the pursuit of social goals—with close others, and with complete strangers—often through conversation.

But conversation is not only functional—it can also be *enjoyable*. Previous work has defined phatic conversations as those that happen primarily to pursue social goals (rather than informative or competitive ones). But these goals need not be strategic —talking to other people can be intrinsically rewarding (Dunbar et al., 1997; Epley & Schroeder, 2014). And many other choices we make are tied up in what we talk about with other people (Berger, 2014; Kumar & Gilovich, 2015).

In this paper, we focus on cooperative conversations in which all participants aim to mutually enjoy their time together. Of course, there are many other conversational goals, such as information gathering, deal making, persuasion, impression management, and so on—but these goals often involve people with competing interests, and tradeoffs, cheap talk, and stalemates are more likely (Crawford & Sobel, 1982). Even when people share a seemingly simple, integrative

goal such as mutual enjoyment, we expect substantial room for improvements in their conversational behavior together.

As a decision-making environment, conversation is uncertain, fast-paced, and high-dimensional. Though many isolated aspects of language generation can be understood in terms of efficient information transmission (Zajonc, 1960; Grice, 1975; Misyak et al., 2014), the *compounded* decisions necessary for conversation are quite difficult. Even the simplest conversations are a deluge, with every turn prompting a judgment about what was just said, and a decision about what to say next. Furthermore, the utility function for conversations is just as complex—what it means to "say the right thing" depends on the people involved, their goals, emotions, and beliefs; the rules and norms of the groups to which they belong; the location and context; and on and on. We are interested in the heuristics that people use in conversation to navigate this environment together.

Our research focuses on one of the most common and important choices in conversation: topic selection. *Every* turn in *every* conversation requires the speaker to make a choice: should we stay on the current topic or switch to a different one? In computational linguistics, topics are commonly thought of as latent semantic clusters embedded in a massive web of meaning. This structure is typically estimated from word co-occurrence in large corpora such as books, news articles, journal articles, or websites (Blei, Ng & Jordan, 2003; Blei & Lafferty, 2007). And topic segmentation strategies for single-authored documents typically rely on semantic shifts across topic boundaries - for example, whether the distinctive phrases of adjacent paragraphs are drawn from similar clusters.

But topic selection in conversation is different—often more explicit—than that. In almost every kind of conversation, people try to coordinate around mutually understood topics, to better communicate with each other and build a shared frame of reference (Sacks, Schegloff & Jackson, 1979; Schegloff, 2007; Echterhoff, Higgins & Levine, 2009). The topic of discussion can change frequently—or rarely—during a conversation, and topic boundaries can be deliberately managed by the participants themselves, using cues to transition away from the current topic and to start a new one (Passonneau & Litman, 1997; Galley, McKeown, Fosler-Lussier & Jing, 2003; Nguyen et al., 2014; Bonin, Campbell & Vogel, 2015). Topic selection is integral in the structure of conversations, and we want to understand how these decisions are made, what people are thinking while they make them, and whether there are ways to help people manage topic search, selection, and shift decisions more effectively.

Although humans are hard-wired for some aspects of conversation (e.g., turn-taking, laughter), common limitations in human decision-making may hinder effective topic selection. For example, people may have difficulty understanding their partners' topic preferences in conversation because we know people struggle with perspective-taking in general (e.g., Eyal, Steffel & Epley, 2018). However, compared to many perspective-taking tasks used in prior research, in conversation, judges have a rich set of information from which to evaluate their target—listening to what someone has to say about a topic provides rich data to learn their true preferences. On the other hand, signals of topic preferences may be complex: speakers may not be explicit about their topic preferences because they fail to communicate clearly, for strategic purposes, or simply to be polite (Brown & Levinson, 1987). And even when they do speak candidly, they may overestimate how transparent they are to others (Gilovich, Savitsky &

Medvec, 1998). Listeners may also have their own troubles understanding the speaker's perspective (Epley, 2008). In particular, people may egocentrically overweight their own preference as a cue of others' topic preferences (Epley, Keysar, van Boven & Gilovich, 2004; Goel, Mason & Watts, 2011; Tamir & Mitchell, 2013).

Many of these basic judgment biases are likely to compound in the dynamic context of a live, unfolding conversation. For most turns, the current topic often acts as a (helpful) coordinating default ("we can just stay on this topic"). However, this type of topic stagnation may represent a form of status quo bias, in which *availability* (current topics are easily accessible) narrow the scope of considered topics (Kahneman, Knetsch & Thaler, 1991; Berger & Schwartz, 2011), leaving many, potentially-better topics undiscussed.

Introducing a new topic may also be perceived as risky or uncertain, while the opportunity cost of undiscussed topics are likely to be overlooked (Frederick et al., 2009). As people update their beliefs about their partners, they may naively overestimate each other's true preference for topics that happen to have been discussed. Over time, this tendency can lead to perseveration around stale, contextually-cued topics (Gilbert & Malone, 1995; Shiller, 1995). Taken together, these heuristics and biases are likely to make topic selection a decision fraught with incorrect beliefs, interpersonal misunderstandings, and missed opportunities.

Overview of the Current Work

In this paper, we examine two questions at the core of a generative topic selection model. First, how well do people detect others' topic preferences? We predict that people have substantial room for improvement in detecting their partner's topic preferences, and we seek to benchmark human topic detection against natural language processing algorithms (using the

same conversational input data to draw inferences). Second, do people's topic-selection choices influence their enjoyment of live conversations? We predict that people search, select, and shift topics sub-optimally—and that simple interventions may help them improve their conversational enjoyment by selecting topics more effectively.

To investigate these questions, we conducted four main studies. Studies 1-2 focus on topic preference detection, while Studies 3-4 focus on interventions to help people improve their topic selection decisions to increase enjoyment. Across all of our studies, we report how we determined our sample sizes, all data exclusions, all manipulations, and all measures. The exact data and code from each study are available as Online Supplemental Material, stored at https://osf.io/t4h3z/.

Studies 1-2: Topic Preference Detection

In Studies 1 and 2, we focus on topic preference detection. That is, can people tell whether their partner is interested in a topic based on what they say about it? In these studies, we asked some people to respond to topic questions and rate their own interest in continuing to talk about each topic. We then showed their responses to others, who tried to guess the writers' interest in each topic.

This paradigm simulates an important, frequent, familiar decision made in conversation —before someone takes a conversational turn, they must decide (explicitly or implicitly) whether they will stay on the topic of the most recent turn, or switch to a new topic. To make this choice well, beliefs about the partner's preferences for the current topic should enter into this calculation. Correspondingly, conversations are likely to go better when the members hold accurate beliefs about each another's topic preferences (all else equal).

Compared to many perspective-taking tasks, this is a complex one. On one hand, readers (listeners) get a rich and relevant signal about their partner's preference by reading his or her own words in response to several topic questions. Furthermore, many common demands on conversational topic preference detection are relaxed here—because our methods are *asynchronous* (not in real-time conversation), there is no time pressure, fewer demands on working memory, and readers do not have to generate their next response as in live conversation. On the other hand, text is high-dimensional information: the cues that accurately indicate one's topic preference may be scattered, and readers may not have the right mental models for how to deduce topic preferences, especially in the absence of nonverbal signals.

Study 1 Method

Participants in Study 1 were recruited for two separate stages. First, we recruited participants ("writers") to write responses to (and report their preferences for) twelve topic questions. Then, we used their written responses as stimuli for a second set of participants ("readers"), who read the writers' responses and guessed their topic preferences.

The topic questions were selected from a larger pool of 50 conversation questions drawn from previous research on conversation (Aron et al., 1997; Huang et al., 2017). This selection was done empirically, using a pilot study that closely resembled the writer study (see Appendix A for full details, and Appendix B for the initial list of 50 topics). Based on the results of the pilot study, we selected a set of 12 conversation topics that had a medium average preference rating (i.e., neither the most nor the least preferred topics) with high heterogeneity in ratings across the population (see Table 1 for the final list).

All writers and readers were recruited from Mechanical Turk and completed attention checks before starting or being counted in our sample (see our OSF repository for exact stimuli). We recruited 400 writers who passed those attention checks, but excluded one for writing nonsense in the text boxes, and seven for giving near-identical answers to the 12 topic preference questions ($\sigma<1$), leaving a final sample of 392. We then intended to recruit enough readers so each written text would receive at least three readers. 693 readers passed the attention checks, but 38 did not complete the survey, and one person made the same guess for every person ($\sigma<1$), leaving 654 readers in the final sample.

Writer Study: All participants in the writer study were told to imagine they were in a conversation with someone. Writers were randomized into one of two between-subjects conditions: half of the writers were told to imagine they were conversing with someone they speak to often (*"close target"*), while the other half were told to imagine they were conversing with someone they had never met (*"distant target"*). They were presented with the twelve topics, one at a time, and responded to "How would you respond to this question in conversation?" (see Appendix C for the full prompt). After writing their response, they rated their interest in staying on the topic (+10) versus switching to a different topic (-10), using a slider tool on a single scale (which was initially positioned at zero, and required a response before continuing). Finally, they completed some demographic questions.

Reader Study: Each reader was randomly assigned to read responses on only six of the twelve topics, written by four writers from the writer study, for a total of 24 topic preference judgments. The writers were assigned so each reader only saw writers from one condition (close or distant) but there were no additional randomizations in the reader study. At the beginning of

the study, they were shown exactly what the writers saw (i.e. matching the writers' condition), as well as a histogram of all the writers' ratings, to get a sense of how people used the stay/switch preference scale in aggregate. Readers were also incentivized for performance, with the ten most accurate participants receiving a bonus of $2.

For each of the 24 topic responses, the reader saw the topic question, the writers' response in text, and made three judgments (see Appendix D for full prompts). First, they predicted the writer's preference for the topic, using the same -10 to +10 slider that the writers used (*"prediction"*: this was the primary measure, and was used to determine the incentives). Then, they reported their own preference for staying on that topic in conversation with the writer herself (*"paired preference"),* also on the -10 to +10 scale. Third, they reported whether they might themselves switch topics or stay on topic after this writer's turn in conversation ("*intent to switch*") on a 1 to 7 likert scale.

Readers were shown all six responses for each writer in a single block, before moving to the next writer. So at the end of each block, readers evaluated their overall impression of that writer, using common warmth and competence scales. They also predicted the gender of the writer. After all of the writer judgments, they also gave their own preference for all twelve topics without any partner in mind (*"generic preference"*). Finally, they completed some demographic questions.

Study 1 Results and Discussion

The distribution of writers' topic preferences is shown in Figure 1. Notably, the ratings were not normally distributed. Furthermore, there were large person-level differences in average ratings (representing mood, general interest in conversation, response bias). Accordingly, we

evaluated detection accuracy non-parametrically, and *within-writer*. That is, the question was: can a predictor (human or algorithm) read two responses written by the same person and determine the more preferable topic? This metric aligns with decisions in conversations, where people typically have a fixed partner and a wide consideration set of topics.

Overall, readers' predictions were correlated with the writers' true preferences (*average kendall's* τ = .142, bootstrapped 95% CI = [.127, .157]). It is clear that readers engaged in some perspective-taking, as they were more accurate than if they had merely used their own topic preferences as a perfect proxy for the writer's preferences (τ = .097, 95% CI = [.083, .112], paired t(653)=6.8, p<.001). However, readers did not learn much from the writers' text. For example, readers did not perform significantly better than a simple algorithm that guessed the population average for each topic (τ = .128, 95% CI = [.115, .141], paired t(653)=1.4, p=.155).

How much better could they have done? Was there a missed opportunity for perspective-taking? And can we attribute these mistakes to the readers (who may be biased, or miss cues in the text) or to the writers (whose text may not contain cues about their topic preference)? To better understand the judgment process, we used natural language processing to map features of the text onto the readers' predictions of the writers' topic preferences, and onto the writers' actual preferences.

Biases in Human Judgement: Perhaps the most available cue of a writers' topic preference for readers is their very own preference for the topic. Readers' own topic preference was a weak but useful cue (τ = .034, 95% CI = [.020, .049]), since it can contain some information about the average preference in the population (Hoch, 1987; Gilbert, Killingsworth, Eyre, & Wilson, 2009). But, as in many other preference domains, people fell prey to the false

consensus effect: their predictions for others were overly correlated with their own preferences for the topics (τ = .128, 95% CI = [.111, .145]).

This correlation is even stronger when we compare readers' predictions to their specific preference to talk more about a topic with the writer (τ = .531, 95% CI = [.511, .551]). However, it is not clear this is an error, as causality may run in reverse. That is, people may reasonably have an explicit preference for topics their partner finds interesting. In fact, their belief about their partner's topic preferences was even more highly correlated with their stated intent to stay on topic with the writer (τ = .584, 95% CI = [.565, .603]). These results underscore the lay belief that topic preference detection is important for conducting conversation.

Aggregating multiple judges can mitigate egocentric bias, because the judges' idiosyncratic preferences cancel each other out, if they are uncorrelated. Indeed, the readers' performance was better when all 3-6 predictions for the same text were aggregated into a crowd-average prediction (τ = .185, 95% CI = [.161, .210]). However, it is not clear how meaningful this aggregated benchmark can be for individuals, who often need to determine their conversation partners' topic preferences alone (in one-on-one conversation).

Some biases were also consistent across all human judges, and thus persisted even after aggregation. In particular, several simple cues in the text tended to be overweighted. For example, the word count of the writers' responses: the more someone wrote, the more they liked the topic. However, aggregated human judgments put more weight on word count than was warranted A similar pattern emerged with sentiment analysis (Mohammad & Turney, 2013). Though positive sentiment indicated writers' greater interest in topics, aggregated reader judgments overweighted positive sentiment as a cue for the writers' topic interest. These

comparisons are plotted in Figure 2 and Figure 3. In particular, human judges tend to underestimate how much people have to say about a topic even when they do not want to talk about it (perhaps underestimating the motivating power of politeness norms that urge or require people to speak, even on topics they dislike).

Comparing Human Judgment to Machine Judgment: In this study, topic preference detection can be thought of as a text classification problem: we extracted features from a subset of the open-ended text responses, trained an algorithm empirically to predict the outcome variable (via cross-validation), and applied the model to predict topic interest (outcome variable) on held-out text responses. We used a relatively simple machine learning algorithm (LASSO regression) and evaluated accuracy using a 20-fold nested cross-validation (Varma & Simon, 2006). Using this framework, we compare predictions from simple feature extraction from the writers' responses with several other more-sophisticated approaches.

We calculated two feature sets focused on semantic content: we applied word2vec embeddings pre-trained on web crawler data (Mikolov et al., 2017) to project every response into a single 300-dimensional semantic space; and we estimated an unsupervised 20-topic latent dirichlet allocation model within each topic, which learns clusters of words from co-occurence within our own data (Roberts et al., 2014). We also calculated two sets of syntactic features. One was from a set of 34 social cues nominally gathered to detect politeness in conversation (Yeomans, Kantor & Tingley, 2018). The other tried to capture common superficial features, by calculating all third-order polynomial combinations of word count, sentiment, and emotionality (i.e. absolute magnitude of sentiment words). Finally, we calculated an omnibus feature set that combined all of the other feature sets together, which we use as our main benchmark algorithm.

In developing these models, one question was: are cues of topic preference topic-generic or topic-specific? The answer would determine whether a topic preference model could perform transfer learning across topics, or whether it required the training data and the held-out data to have common support over topics. We performed early testing using weighted regression that systematically adjusted the importance of in-topic data. We found that there was potential for transfer learning across topics, with all feature sets predicting above chance in the pure hold-out case. However, the most accurate version of each algorithm seemed to put as much weight as possible on the in-topic data.

Based on these analyses, we decided to train entirely within topic as our benchmark for machine judgment, using nested 20-fold cross-validation to estimate the models and make held-out predictions. In Table 2, we compare the results of this procedure using several different NLP feature sets, as well as the aggregated human judges (readers). Although there are some differences from topic to topic, and across feature sets, our main result is that the omnibus NLP model (τ = .167, 95% CI = [.180, .154]) significantly outperformed individual humans (paired t(653)=2.7, p=.007). Aggregated groups of human judges performed similarly to the NLP model (paired t(391)=1.3, p=.186), and a linear ensemble that combined both predictions was the most accurate (paired t(391)=3.2, p=.001), suggesting that the algorithm was able to pick up on cues the humans did not. Given the limited training set (n~370 per outer fold) and simple algorithm (LASSO), we believe that this result represents a conservative benchmark for algorithmic performance on topic preference detection.

Close vs. Distant Target: Table 2 also shows separate accuracy results for both conditions from the writer study (as they imagined conversing with a close other versus

stranger). Human judges were worse at detecting topic preferences of people writing to close others than to strangers ($t(390)=1.76$, $p=.079$). And interestingly, the NLP algorithm showed an even larger drop in performance ($t(390)=2.8$, $p=.006$). This result could indicate several things about how we converse with close others. It is possible that we are more direct with people we do not know, in general, because we do not assume they know us as well. However, it is possible that close others simply communicate in subtle or idiosyncratic ways that may not be detectable by an algorithm trained on other peoples' responses, or by human judges who themselves do not know the speaker. To test these hypotheses, we needed to collect paired data from writers and readers who actually know each other, which we pursue in Study 2.

Study 2 Method

In Study 2, we made some modifications to the topic preference detection paradigm from Study 1. To collect data from people who knew each other, we recruited writers and readers simultaneously as pairs to a behavioral lab (e.g., spouses, partners, friends, coworkers, etc.). In this design, there were no randomized conditions, and each person in the pair was both a writer and a reader for their partner. First, each person wrote their response and rated their preference for each of the twelve topic questions at separate computers. Then, they both switched seats, and tried to guess their partners' preferences.

Every person predicted their partner's preferences for all twelve topics, but this was done in two blocks (of six items each). In one block, they read the topic question and their partner's response, and predicted their preference once. In the other block, they made two predictions for each topic. First, they made a prediction without seeing their partner's text (i.e. based on what they already know about their partner before the experiment). Then, they saw their partner's text,

and were allowed to adjust their initial prediction. The assignment of topics to block, and the order of the two blocks, was determined randomly for every participant. All predictions were made on the same -10 (switch topics) to +10 (stay on topic) slider.

One major addition to this study was that we also measured readers' confidence in their predictions (Moore & Healy, 2008). After every prediction they made (with or without the text) they reported the probability that their guess was within three scale points (above or below) of their partner's true rating. Additionally, at the end of the study, we asked them to make a summary judgment about how many of their twelve with-text predictions were within three scale points of the correct answer. Finally, we collected demographics, as well as information about how well the partners knew each other.

We intended to collect 200 participants (i.e. 100 pairs), and in our preregistration we decided to only exclude a pair based on a research assistant's judgment that one or both of the participants were not following instructions, or did not finish the survey. XXX people were excluded, leaving a final sample of XXX.

Study 2 Results and Discussion

[TBD!!!]

Comparing Human Judgment to Machine Judgment

For this study, we preregistered the prediction algorithms developed in Study 1. In this planned analysis, we merged the data from both Study 1 and Study 2 together, and re-trained the same algorithm. We again used 20-fold nested cross-validation to make held-out predictions throughout the data, estimating a separate model for each topic. And we again measured prediction accuracy non-parametrically, and within-person....

Studies 3-4: Topic Management in Conversation

In Studies 3 and 4, we shift to examine how people manage topics in live conversation. Do people's topic-switching strategies influence enjoyment? Can simple topic-switching interventions improve enjoyment? We conducted two experiments—one with face-to-face interactions, and one using online chat—to investigate how people choose and shift topics during cooperative conversations. In particular, we examine the frequency with which people switch topics together. Theoretically, many decision-making heuristics and biases may lead people to over-perseverate on a small number of topics, at the expense of conversational enjoyment. And practically, topic switching is an intuitive conversation strategy that can easily be implemented by participants themselves. Therefore, in these studies, we encouraged people to switch topics more often as an intervention designed to help our participants have more enjoyable conversations.

Study 3 Method

We recruited adults to a behavioral laboratory individually (participants did not know each other). First, they read the list of twelve topics from Studies 1-2 on separate computers and rated their interest in each topic, just like the writers in Study 1. Then they went to meet their conversation partner—another participant they had never met before—in another room, and spent ten minutes discussing the twelve topics.

All participants were given a sheet of paper during the conversation with the list of the twelve topics in a random order, matched within the pair, and some instructions (see Appendix

E). Participants were all given the single, clear, explicit goal to have fun and enjoy their conversation together.

We also randomized pairs into one of two conditions. Half of the pairs were encouraged to switch topics frequently during the conversation ("*switch often*"), while the other half received no additional instructions ("*natural*"), and were free to choose whatever topic-switching frequency they thought would be most enjoyable.

After their conversation, both people went back to their computers. First, they re-evaluated their own preferences for staying on the twelve topics with their partner, and predicted their partner's preference for staying on the topic with them. Next, they reported how much they enjoyed the conversation, and their impression of their partner (liking, warmth, competence, etc; see Appendix F for full measures).

All of the conversations were recorded on video, and the audio tracks were transcribed by a third-party vendor. Research assistants then watched the videos to correct the transcripts (spelling, speaker disambiguation, etc.), and to annotate the transcripts by labeling every turn. All transcripts were looked at by at least two annotators (with disagreements resolved via discussion). Every turn in even conversation was given one (and only one) of sixteen labels, including any one of the twelve topics, an introduction topic (e.g. hellos, exchanging names), an ending topic (e.g. goodbyes); an "off-topic" label for diversions from the list (e.g. sports, movies, current events), or a "switch" label for turns in which a speaker finished one topic and started the next topic in the same turn. After the initial annotations, each switch turn was split by hand into two fragments that each covered a single topic.

Study 3 Results and Discussion

The annotations allowed us to confirm that our manipulation worked as intended. Dyads in the switch-often condition discussed more of the topics on average (m=9.27, SD=3.15) than dyads in the natural condition (m=6.57, SD=3.23; t(194)=4.2, p<.001). This distribution was right-skewed, but still less than half of the dyads (44%) touched on all twelve topics for at least one turn (compared to 15% in the natural condition). This suggests that participants did not view our intervention as a mandatory checklist, as they seemed to prioritize enjoyment over 12-topic completion.

Conditional on being brought up at all, the average topic lasted for 75.6 seconds (SD=65.9) and 10.8 total turns (SD=9.8). Because switch-often dyads covered more topics than natural dyads, they spent fewer turns (m=9.6, SD=8.2 vs. m=12.7, SD=11.7; cluster-robust t(1566) = 2.7, p=.008) and less time (m=65.4, SD=50.0 vs. m=92.0, SD=82.9; cluster-robust t(1566) = 3.3, p<.001) per topic. But overall, there was a similar amount of conversation in both conditions—the switch-often dyads did not use significantly more turns (m=177, SD=80 vs. m=166, SD=95; t(96) = 0.6, p=.538) or significantly more words (m=1716, SD=289 vs. m=1633, SD=258; t(96) = 1.5, p=.140) than the natural dyads. This suggests our treatment induced more topic switching, but not more talking, in the switch-often condition.

Conversation Outcomes: Our primary outcome measure of the conversation was a standardized index of the five-item enjoyment questions (Chronbach's α=.89). The distribution of this outcome by condition is plotted in Figure 5. Conversation enjoyment was significantly higher among frequent-switching dyads (m=.14, SD=.71) than among natural-switching dyads (m=-.16, SD=.94; cluster-robust t(194)=2.4, p=.019). Although secondary, we also collected a four-item index of interpersonal liking (Chronbach's α=.80) and found that there was a

marginally significant effect, such that switch-often dyads liked each another (m=.10, SD=.73) more than did natural dyads. (m=-.11, SD=.85; cluster-robust t(194)=1.8, p=.080).

Topic Outcomes: We measured participants' topic preferences both before the conversation (with no particular partner in mind) and after the conversation (specifically with their partner), for all twelve topics. On average, topic preferences were lower after conversations (m=1.52, SD=5.79, than before them (m=1.99, SD=6.03; paired t(2210)=2.6, p=.010), suggesting some general topic satiation. More importantly, this allowed us to estimate how the conversation itself affected preferences for staying on topic while controlling for endogenous initial preferences (and clustering standard errors to account for correlated errors within-person and within-dyad; Zeileis, 2004).

Our topic-switching manipulation had a direct effect on participants' preferences for the topics themselves. That is, participants who were told to switch often wanted to keep talking about the twelve topics with their partner, compared to participants who did not receive those instructions (β=1.122, SE=.382; cluster-robust t(2350)=2.9 p=.004), and this result holds controlling for initial topic preferences (β=.839, SE=.342; cluster-robust t(2349)=2.5, p=.014). This provides converging evidence that, rather than distracting from a boring list of topics, our manipulation helped participants to enjoy the conversations they were having.

Human Preference Detection: After the conversation, we asked participants to predict their partner's topic preferences. As in Study 1, we again evaluated prediction accuracy non-parametrically, *within-person.* That is, how well can a person predict their partner's preference for one topic over another? Human judges were successful at detecting their partner's post-conversation topic preferences (τ = .203, 95% CI = [.165, .240]), somewhat outperforming a

simple baseline that just guessed the average rating for each topic (τ = .167, 95% CI = [.137, .197], paired t(195)=1.6, p=.102). These predictions were also somewhat correlated with pre-conversation topic preferences (τ = .141, 95% CI = [.109, .171]), though that may have been because the generic and specific preferences for the same topic were themselves highly correlated (τ = .408, 95% CI = [.375, .440]).

Surprisingly, human predictions of partner preference were less accurate in the switch-often condition (τ = .168, 95% CI = [.117, .219]) than in the natural condition (τ = .243, 95% CI = [.189, .295], t(194)=2.0, p=.049). This was primarily because ignoring a topic is often a good signal of someone's topic preference (τ = .179, 95% CI = [.137, .222]), and switch-often dyads ignored fewer topics. As a sanity check, no one in any condition was particularly good at distinguishing preferences for topics they did not discuss (τ = -.007, 95% CI = [-.076, .062]).

Machine Preference Detection: We had less data to train on in this study than in Study 1, for two reasons: the sample size was smaller, and most dyads did not discuss all topics. We first tried to use the Study 3 data as a pure held-out test set. That is, we trained the NLP algorithms from Study 1 on only the Study 1 data, and then applied them to the text here. To focus on dialogue acts that were comparable to the responses in Study1, we stripped out all backchannels and questions from the text of the transcripts, leaving just the statements each speaker made while talking about the topic.

As a fair comparison to the humans' predictions, we tested the model using the participants' partner-specific preferences as the outcome variable. We found modest accuracy (τ = .109, 95% CI = [.080, .137]). We also found that the algorithms' predictions lined up somewhat more closely with the generic preference ratings given by participants (τ = .131, 95% CI = [.

100, .162]). This is perhaps not surprising, given that the outcome variable in Study 1 was a generic topic preference, as well. But this suggests that the cues to partner-specific preferences will likely be in context of the conversation.

Although our topic-generic model in Study 1 was not overly successful, we tried it again here. There is reason to be more optimistic here because face-to-face conversation may carry many more explicit social cues to communicate topic preferences during the conversation, compared to the asynchronous interactions in Study 1. This might increase the amount, and the informativeness, of cues we might recover in a natural language processing model.

We again calculated the set of social cues included in the politeness detection package, and also added an array of conversational behaviors that would be relevant in live conversation (but not in Study 1). The transcripts explicitly identified laughter,and interruptions. And the topic annotations indicated which partner mentioned a topic first. Also, we labeled one-to-three-word utterances in the midst of two longer turns by one's partner as backchannels. We also parsed the kinds of questions people asked one another, into one of three types, using the topic annotations: "switch" questions, which introduced a topic; "follow-up" questions, which asked a person to elaborate on a statement they made; and "mirror" questions, which asked a person to themselves answer a question they had just asked their partner. Switch questions were identified in the first two turns of a new topic (as labeled), and questions on other turns were labeled as mirror or follow-up using the question type classifier from Huang et al. (2017).

We compared the prevalence of these conversational features across all topics that were discussed for at least one turn (n=1564) in Figure 6. We find that there are many common topic-generic signals to topic preference. Some may not be surprising—backchannels and follow-up

questions are common signals of responsiveness and interest in a partner, which is likely to signal (and/or induce) a partner's preference for a topic. When someone mentions a topic first, they are likely to be interested in it, whereas when their partner mentions it first, they are less likely to be interested in it. In addition, laughing is a common sign of enjoyment. One perhaps surprising feature is interruptions: some previous papers have found interruptions as evidence of competitive status-seeking behavior (Mendelberg, Karpowitz & Oliphant, 2014; Jacobi & Schweers, 2017). We find here that in phatic conversation, interrupting one another is a sign of a dynamic, bubbling discourse, in which the participants are interested and listening attentively (Collins et al., working; Fay, Garrod & Carletta, 2000).

To estimate a full model of topic preference, we put all these features (as well as the full set of politeness features from Study 1, and the average topic ratings) into a nested cross-validation. The inner loop of the cross-validation contained 20 folds, as before. But here we partitioned the outer folds into 10 folds per topic (120 total), so that for any given loop, all the held-out documents were from the same topic. The training set was thus a mix of some in-topic and all out-of-topic documents, with equal weight on each document. This model was quite accurate, outperforming even the targets' own human partners (τ = .260, 95% CI = [.223, .294], t(195)=2.7, p=.008). And like the human judges, the algorithm was less accurate for people in the switch-often condition (τ = .221, 95% CI = [.171, .267]) than in the natural condition (τ = 303, 95% CI = [.249, .352], t(194)=2.3, p=.025).

For good measure, we built an ensemble model that combined five datapoints for each document: the topic-specific model trained on Study 1 data; the topic-generic model trained on social cues within Study 3; the humans' predictions for their partner; the average topic rating; and

whether the topic was discussed or not. We performed a leave-one-out nested cross-validation (with 20 inner folds per loop) to make out-of-sample predictions using the same LASSO regression, and found no improvement from the topic-general model (τ = .263, 95% CI = [.229, .294]).

Although these results are promising, we do not think we are close to ceiling performance for topic preference detection in conversation. For one, more training data, from more speakers, about different topics, will surely improve the generalizability of the model. Additionally, it is clear that context can moderate the detectability of topic preferences. In our case, the randomized treatment mattered, but surely other contextual variables will also matter. More sophisticated machine learning (e.g., better feature representations or different estimators) might also improve our algorithm's performance. Although the algorithm may not be as good as it could be, it is still good enough to demonstrate that humans are not as good as they could be, either.

Study 4 Methods

We conducted a pre-registered replication of Study 3 with three important changes. First, participants conversed through online chat (via ChatPlat, see chatplat.com) rather than face to face. Second, we removed the 12-topic constraint entirely (including preference ratings and predictions for the twelve topics). Third, we added a new condition. In Study 3, pairs were always in the same condition—both encouraged to switch frequently, or both switching naturally. In Study 4, one third of pairs were composed of one partner who was encouraged to switch frequently, and one person who was not. This difference was not made explicit to the partners, though all pairs in all conditions were told their partner might receive different instructions.

We recruited participants from Mechanical Turk. We intended to collect a sample size of at least 100 participants per condition. However, our pre-registered protocol made it difficult to anticipate recruitment exactly, since there was attrition as people waited for partners, or after they were already matched, on top of several attention checks, and a research assistant checking to filter people who did not earnestly converse. Although 1308 people began our survey, 1156 people were paired in dyads, and 658 people (329 dyads) passed all the post-chat exclusion criteria.

Study 4 Results and Discussion

Text chatting has a different cadence than spoken conversation, with longer delays for people to think and plan the next turn. Accordingly, we found that there were markedly fewer turns per person (m=13.7 SD=6.73), and fewer words per turn (m=10.65, SD=9.15) compared to Study 3. But again, there did not seem to be much difference across conditions in the amount of turn-taking, or in the word count of the average turn (all $p>.3$).

The difference in enjoyment between conditions is plotted in Figure 7. Replicating the results of Study 3, we found that dyads in which both people were told to switch topics frequently enjoyed their conversation more (m=.096, SD=.829) than dyads in which neither person was told to switch topics (m=-.076, SD=.836; $t(656)=2.0$, $p=.050$). However, in the new condition in which people in the same dyad were given different instructions, they were no better or worse than either of the other conditions, no matter if they were the partner told to switch often (m=-.033, SD=.838) or the one who did not receive those instructions (m=-.003, SD=.796; all $p>.3$).

Future Directions: The transcripts are in the process of being annotated by research assistants for topic management strategies. In this study, annotation is more difficult because the topic boundaries were not defined in advance. Instead, turns will be labeled on a four-point scale indicating either an explicit topic switch (+2), a topic drift (+1), an on-topic, neutral, or ambiguous turn (0), or else an explicit encouragement of the current topic (-1). Every conversation will be given at least two sets of labels from independent annotators, resolving disagreements by discussion. These annotated boundaries should help us enrich our model of topic transitions to better understand how good conversationalists can switch topics effectively.

General Discussion

Topic selection in conversation can be a difficult task, even in enjoyable, cooperative (phatic) conversation. To make good topic decisions on the fly, people must balance their own preferences, beliefs about their partner's preferences, the consideration set of topics, the timeliness and responsiveness of what we say, all while listening. In this paper, we examine this process across thousands of interactions—live and asynchronous, face to face and in text, among strangers and close others.

We identify two areas for improvement. First, people are biased in their assessments of their conversation partners' preferences. Even when they are given someone's perspective on a topic, they may not take it. Second, people may be overly reluctant to switch topics in friendly conversations with strangers. And they may be more likely to enjoy those conversations when they switch topics more often together.

Although we focus mainly on enjoyment, we think that topic management is also critically important to pursue other conversational goals, such as sharing informations, managing meetings at work, or navigating conflict. Topic management often reflects status

These psychological questions are increasingly relevant for the computer science literature on automated dialogue agents (Jurafsky & Martin, 2017). While progress has been clear for specific task- or topic-constrained applications (e.g. personal assistants, question-answering), and some of the domain-general subcomponents (e.g. automated speech recognition), open domain chatbots are from achieving anything like human performance.

Our results suggest that one avenue for progress may be more sophisticated topic management strategies, going beyond utterance-level satisfaction detection (e.g. Fang et al., 2018). More broadly, our results suggest that because human performance in conversation can itself be biased or mistaken, expert-level machine language generation will often require data from human subjects experiments, that test the effectiveness of both utterance-level and topic-level conversation decisions. In turn, improvements in machine dialogue may help us develop interventions that can improve humans' ability to converse with one another.

Recent research has begun to uncover who, when, why, how, and by what metrics of failure and success people falter when they converse. For example, people make errors of *commission*, saying things they shouldn't, as well as errors of *omission*, not saying things they should. These examples demonstrate how challenging it can be to make the right choices in conversation. Much of linguistic theory is built on understanding examples of efficient communication (Grice, 1975). We believe that examples like these of inefficient communication

are perhaps even more interesting - for understanding the human mind, and for developing conversational interventions to help people get along better together.

References

1. Aron, A., Melinat, E., Aron, E. N., Vallone, R. D., & Bator, R. J. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. Personality and Social Psychology Bulletin, 23(4), 363-377.

2. Bawa, M., Manku, G. S., & Raghavan, P. (2003, July). SETS: search enhanced by topic segmentation. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 306-313). ACM.

3. Bearman, P., & Parigi, P. (2004). Cloning headless frogs and other important matters: Conversation topics and network structure. Social Forces, 83(2), 535-557.

4. Berger, J. (2014). Word of mouth and interpersonal communication: A review and directions for future research. Journal of Consumer Psychology, 24(4), 586-607.

5. Berger, J., & Schwartz, E. M. (2011). What drives immediate and ongoing word of mouth? Journal of Marketing Research, 48(5), 869-880.

6. Bitterly, T.B., Brooks, A.W., & Schweitzer, M.E. (2017). Risky business: When humor increases and decreases status. Journal of personality and social psychology, 112(3), 431-455.

7. Blei, D. M., & Moreno, P. J. (2001, September). Topic segmentation with an aspect hidden Markov model. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 343-348). ACM.

8. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

9. Bonin, F., Campbell, N., & Vogel, C. (2015). The discourse value of social signals at topic change moments. In Sixteenth Annual Conference of the International Speech Communication Association.

10. Brooks, A.W. (2014). Get excited: reappraising pre-performance anxiety as excitement. Journal of Experimental Psychology: General, 143(3), 1144-1158.

11. Brooks, A.W., Dai, H., Schweitzer, M.E. (2014). I'm sorry about the rain! Superfluous apologies demonstrate empathic concern and increase trust. Social psychology and personality science, 5(4), 467-474.

12. Brooks, A.W., Gino, F., Schweitzer, M.E. (2015). Smart people ask for (my) advice: Seeking advice boosts perceptions of competence. Management Science, 61(6), 1421-1435.

13. Brown, P., & Levinson, S. C. (1987). Politeness: Some universals in language usage.

14. Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. Econometrica: Journal of the Econometric Society, 1431-1451.

15. Choudhury, T., & Basu, S. (2005). Modeling conversational dynamics as a mixed-memory markov process. In Advances in neural information processing systems (pp. 281-288).

16. Cooney, G., Gilbert, D. T., & Wilson, T. D. (2017). The novelty penalty: Why do people like talking about new experiences but hearing about old ones? Psychological science, 28(3), 380-394.

17. Dunbar, R. I., Marriott, A., & Duncan, N. D. (1997). Human conversational behavior. Human nature, 8(3), 231-246.

18. Eagle, N., Singh, P., & Pentland, A. (2003, August). Common sense conversations: understanding casual conversation using a common sense database. In Proceedings of the Artificial Intelligence, Information Access, and Mobile Computing Workshop (IJCAI 2003).

19. Echterhoff, G., Higgins, E. T., & Groll, S. (2005). Audience-tuning effects on memory: The role of shared reality. Journal of personality and social psychology, 89(3), 257.

20. Epley, N. (2008). Solving the (real) other minds problem. Social and personality psychology compass, 2(3), 1455-1474.

21. Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. Journal of personality and social psychology, 87(3), 327.

22. Epley, N., & Schroeder, J. (2014). Mistakenly seeking solitude. Journal of Experimental Psychology: General, 143(5), 1980.

23. Eyal, T., Steffel, M., & Epley, N. (2018). Perspective mistaking: Accurately understanding the mind of another requires getting perspective, not taking perspective. Journal of personality and social psychology, 114(4), 547.

24. Eyster, E., & Rabin, M. (2010). Naive herding in rich-information settings. American economic journal: microeconomics, 2(4), 221-43.

25. Fang, H., Cheng, H., Sap, M., Clark, E., Holtzman, A., Choi, Y., Smith, N. & Ostendorf, M. (2018). Sounding Board: A User-Centric and Content-Driven Social Chatbot. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations(pp. 96-100).

26. Fay, N., Garrod, S., & Carletta, J. (2000). Group discussion as interactive dialogue or as serial monologue: The influence of group size. Psychological Science, 11(6), 481-486.

27. Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy?. Trends in cognitive sciences, 8(1), 8-11.

28. Gilbert, D. T., Killingsworth, M. A., Eyre, R. N., & Wilson, T. D. (2009). The surprising power of neighborly advice. Science, 323(5921), 1617-1619.
29. Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: biased assessments of others' ability to read one's emotional states. Journal of personality and social psychology, 75(2), 332.

30. Goel, S., Mason, W., & Watts, D. J. (2010). Real and perceived attitude agreement in social networks. Journal of Personality and Social Psychology, 99(4), 611.

31. Goldstein, N., Vezich, I., & Shapiro, J. (2014). Perceived perspective taking: When others walk in our shoes. Journal of personality and social psychology, 106(6), 941.

32. Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political analysis, 21(3), 267-297.

33. Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. Journal of Personality and Social Psychology, 53(2), 221.

34. Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1373-1378).

35. Huang, K., Yeomans, M., Brooks, A.W., Minson, J., & Gino, F. (2017). It doesn't hurt to ask: Question-asking increases liking. Journal of personality and social psychology, 113(3), 430-452.

36. Jurafsky, D., & Martin, J. H. (2017). Speech and language processing (Vol. 4). London:: Pearson.

37. Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. Journal of Economic perspectives, 5(1), 193-206.

38. Kumar, A., & Gilovich, T. (2015). Some "thing" to talk about? Differential story utility from experiential and material purchases. Personality and Social Psychology Bulletin, 41(10), 1320-1331.

39. Lewis, M., Yarats, D., Dauphin, Y., Parikh, D., & Batra, D. (2017). Deal or No Deal? End-to-End Learning of Negotiation Dialogues. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2443-2453).

40. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in Pre-Training Distributed Word Representations. arXiv preprint arXiv:1712.09405.

41. Misyak, J. B., Melkonyan, T., Zeitoun, H., & Chater, N. (2014). Unwritten rules: virtual bargaining underpins social interaction, culture, and society. Trends in cognitive sciences, 18(10), 512-519.

42. Mohammad, S. M., & Turney, P. D. (2013). NRC emotion lexicon. National Research Council, Canada.

43. Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. Psychological review, 115(2), 502.

44. Nguyen, V. A., Boyd-Graber, J., Resnik, P., Cai, D. A., Midberry, J. E., & Wang, Y. (2014). Modeling topic control to detect influence in conversations using nonparametric topic models. Machine Learning, 95(3), 381-421.

45. Passonneau, R. J., & Litman, D. J. (1997). Discourse segmentation by human and automated means. Computational Linguistics, 23(1), 103-139.

46. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

47. Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and brain sciences, 27(2), 169-190.

48. Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science, 58(4), 1064-1082.

49. Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In Studies in the organization of conversational interaction (pp. 7-55).

50. Schegloff, E. A. (2007). Sequence organization in interaction: Volume 1: A primer in conversation analysis (Vol. 1). Cambridge University Press.

51. Shiller, R. J. (1995). Conversation, information, and herd behavior. The American Economic Review, 85(2), 181-185.

52. Tamir, D. I., & Mitchell, J. P. (2013). Anchoring and adjustment during social inferences. Journal of Experimental Psychology: General, 142(1), 151.

53. Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. Advances in experimental social psychology, 35(35), 345-411.

54. Yeomans, M. (2018). Some Hedonic Consequences of Self-Expression in Recommending. *Journal of Consumer Psychology*, forthcoming.

55. Zajonc, R. B. (1960). The process of cognitive tuning in communication. The Journal of Abnormal and Social Psychology, 61(2), 159.

56. Zeileis A (2004). "Econometric Computing with HC and HAC Covariance Matrix Estimators." *Journal of Statistical Software*, **11**(10), 1–17. doi: 10.18637/jss.v011.i10.

57. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too?

Table 1: The Topic Space

These are the twelve conversation topics that were chosen in the pilot study for the main experiments.

1. What do you do for work? What do you like about it?
2. Why do you do these kinds of studies?
3. Are you a religious person? Why?
4. Do you have any fruit trees, plants, or a garden?
5. What's the strangest thing about where you grew up?
6. What is the cutest thing you've seen a baby or child do?
7. Would you like to be famous? In what way?
8. When did you last sing to yourself? To someone else?
9. If you were able to live to the age of 90 and retain either the mind or body of a 30-year-old for the last 60 years of your life, which would you want?
10. If you could change anything about the way you were raised, what would it be?
11. What do you value most in a friendship?
12. Your house, containing everything you own, catches fire. After saving your loved ones and pets, you have time to safely make a final dash to save any one item. What would it be? Why?

Table 2: Accuracy of various preference detection benchmarks, across all writers and divided by writer condition. Numbers represent mean per-write r kendall's tau, with the bootstrapped 95% confidence interval in brackets.

Predictor	Accuracy		
	All Conditions	**Close Target**	**Stranger Target**
Humans			
Single	.142 [.127, .157]	.122 [.101, .143]	.164 [.142, .185]
Groups	.185 [.161,.210]	.164 [.131,.197]	.208 [.171,.243]
NLP			
Basic	.125 [.104, .145]	.119 [.088, .150]	.131 [.158, .103]
Politeness	.157 [.133, .179]	.130 [.095, .164]	.184 [.153, .214]
word2vec	.161 [.138, .184]	.139 [.105, .172]	.184 [.153, .215]
LDA	.127 [.105, .148]	.109 [.078, .140]	.145 [.115, .175]
All NLP	.167 [.145, .190]	.136 [.103, .170]	.199 [.169, .228]
Ensemble	.216 [.191, .240]	.190 [.157, .223]	.243 [.206, .276]

Figure 1: Distribution of Topic Preferences

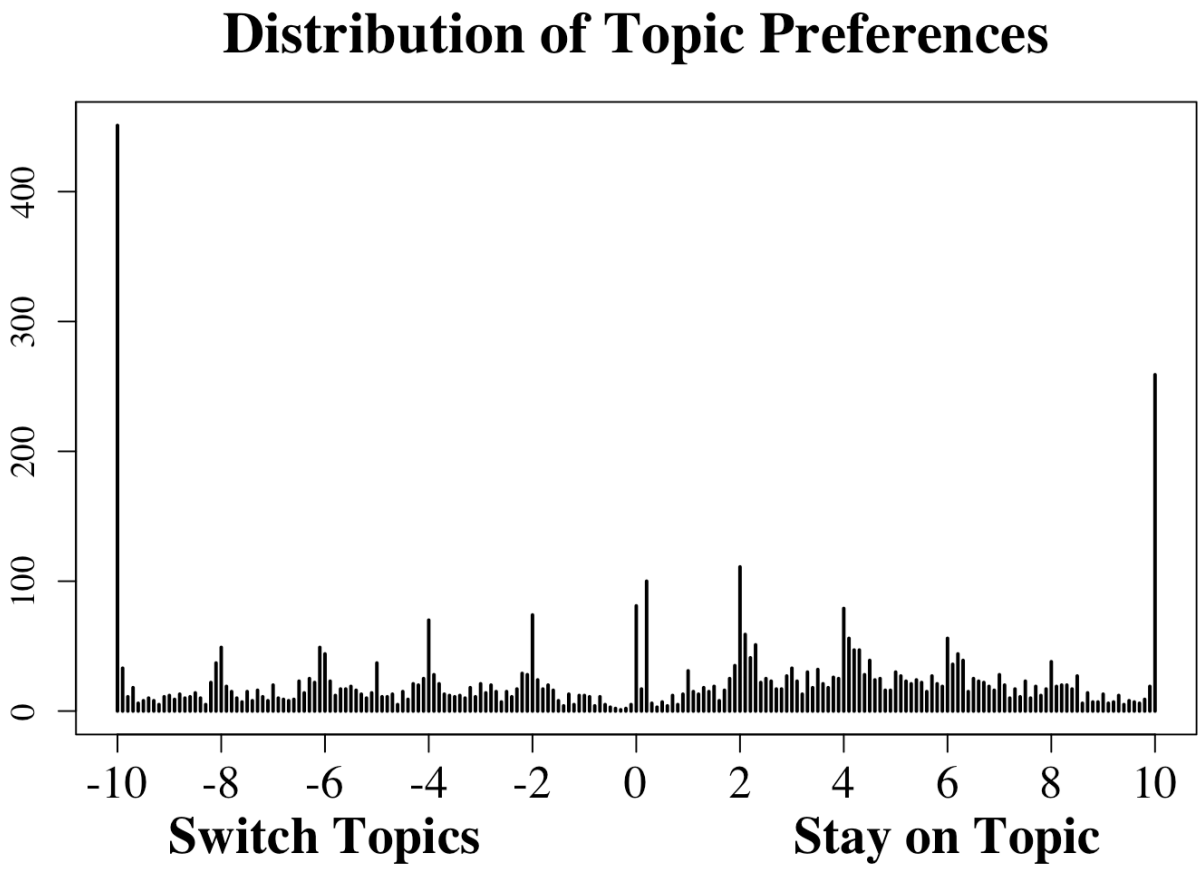


Figure 2:

Predicted and actual relationships between word count and the writer's topic preferences. Human readers tend to over-rate the importance of how much people have to say about a topic, especially at the low end.

Figure 3:

Predicted and actual relationships between word count and the writer's topic preferences. Human readers tend to over-rate the importance of sentiment.

Figure 4: Politeness features indicating topic preferences in Study 1.

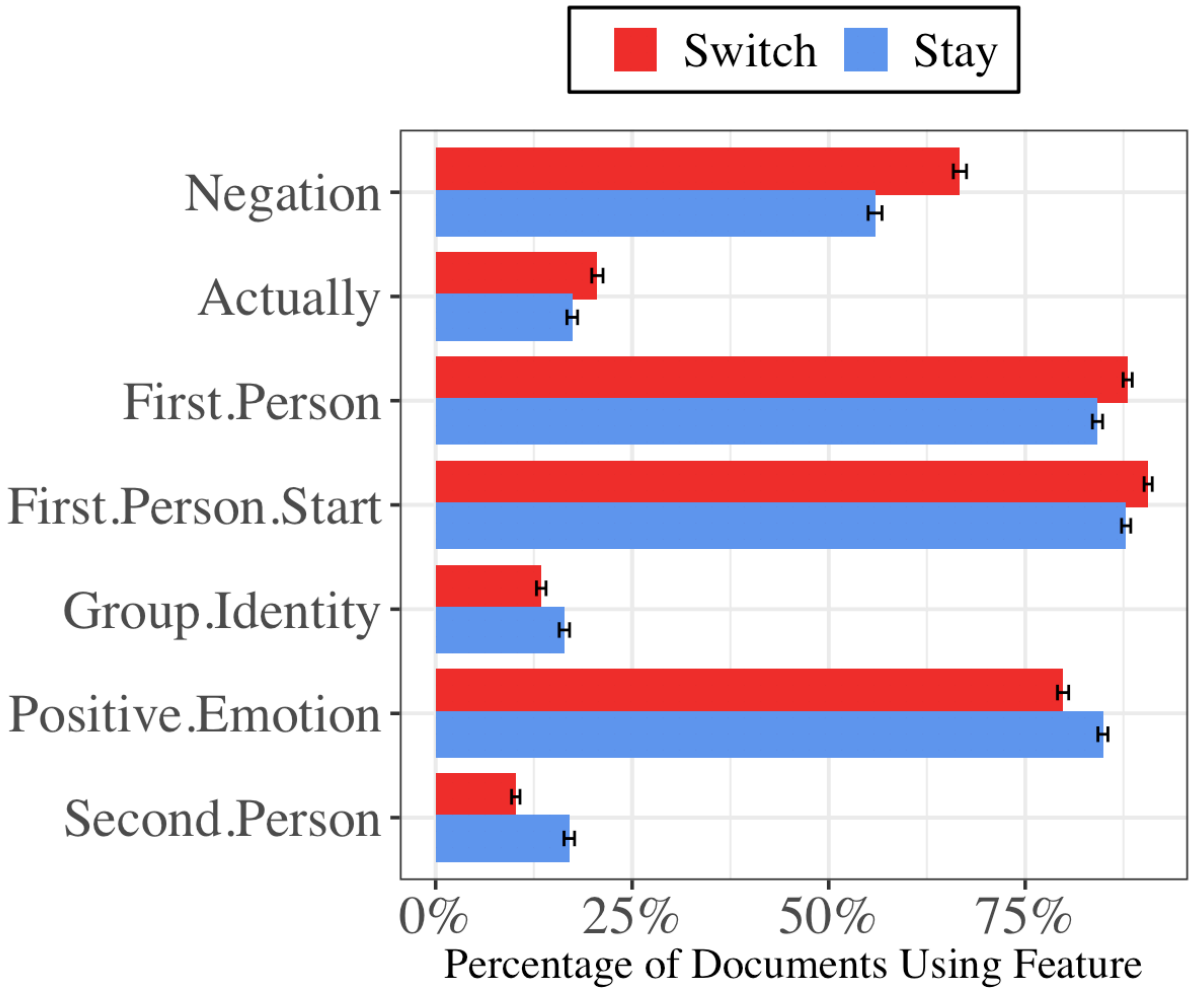


Figure 5:

Distribution of enjoyment index from Study 3, by condition.

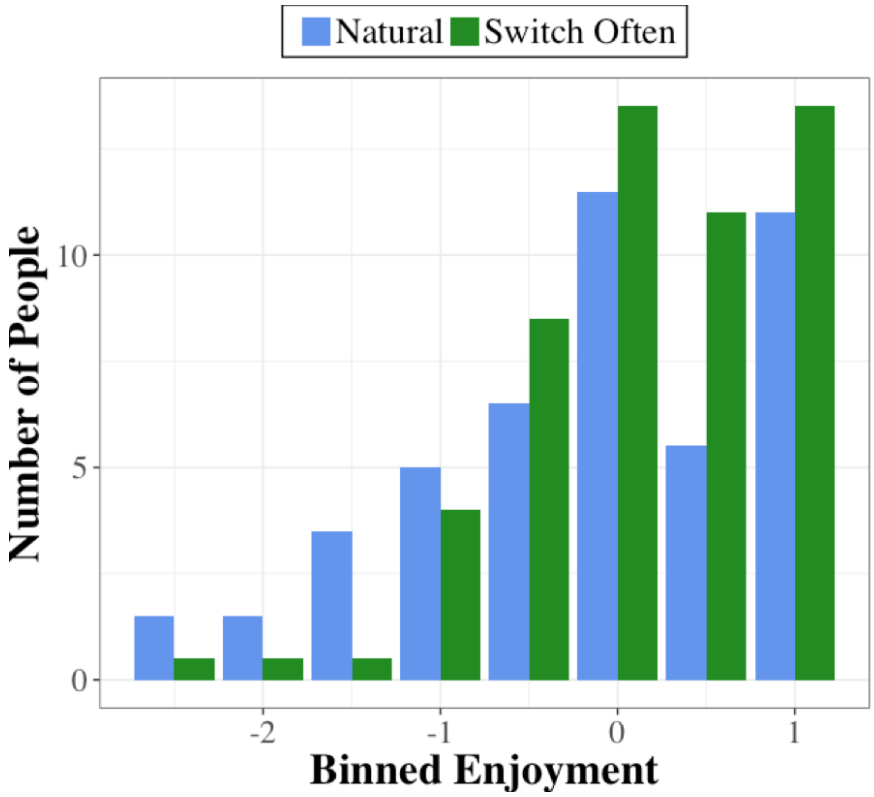


Figure 6:

Conversational features corresponding to topic enjoyment in Study 3.

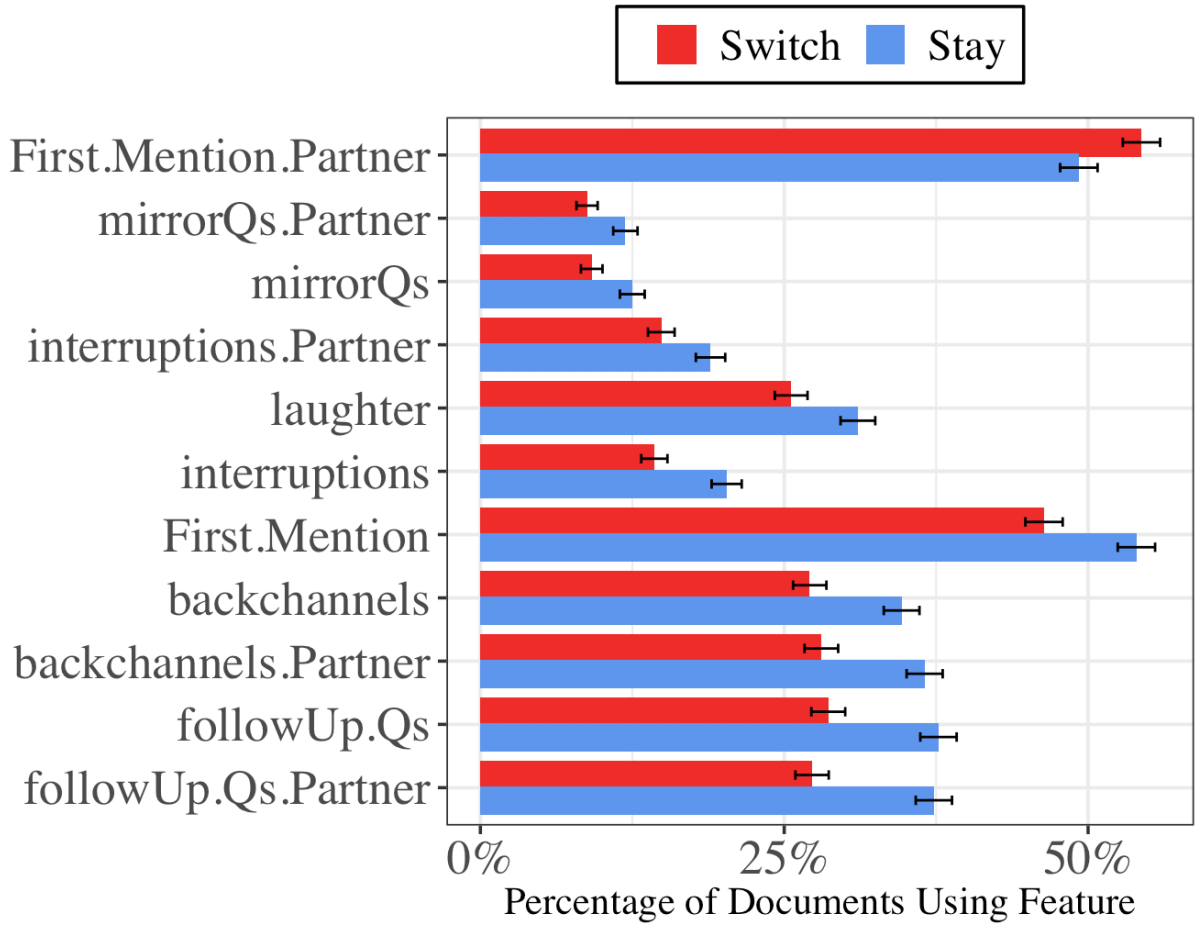
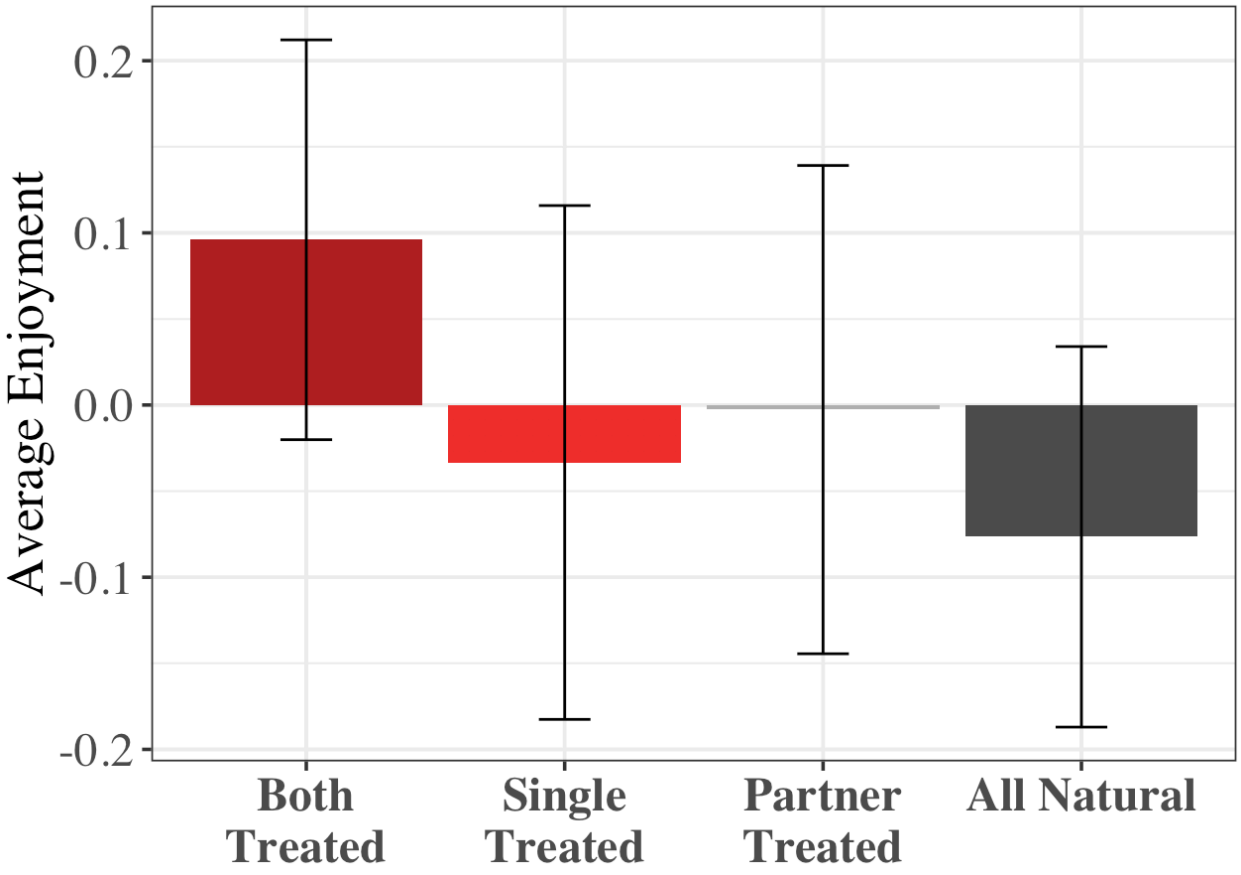


Figure 7:

Effect of condition on enjoyment in Study 4.



Appendix A: Topic Space Pilot Study

For our main experiments, we needed to determine a finite set of topics to present to our participants. However, the universe of possible conversation topics is very high-dimensional. So we decided to use a pilot study to narrow down to a small set of stimuli for our experiments.

We first created a larger pool of potential topic questions. We drew some examples from the well-known Fast Friends procedure (Aron et al., 1997). We also took topic-switching questions from actual conversations that were conducted in a study that had run previously in the same location (Huang et al., 2017). We ended up with a pool of 50 topic questions - for each participant in the pilot, we drew ten topics at random for them to evaluate.

The paradigm of the pilot was almost identical to the writers to our topic preference detection studies (see Appendix C for stimuli). Participants wrote text responses to ten topic questions in a random order. Afterwards, they indicated their interest in staying on topic, or switching to a new topic, for each of the topic questions (in the same order).

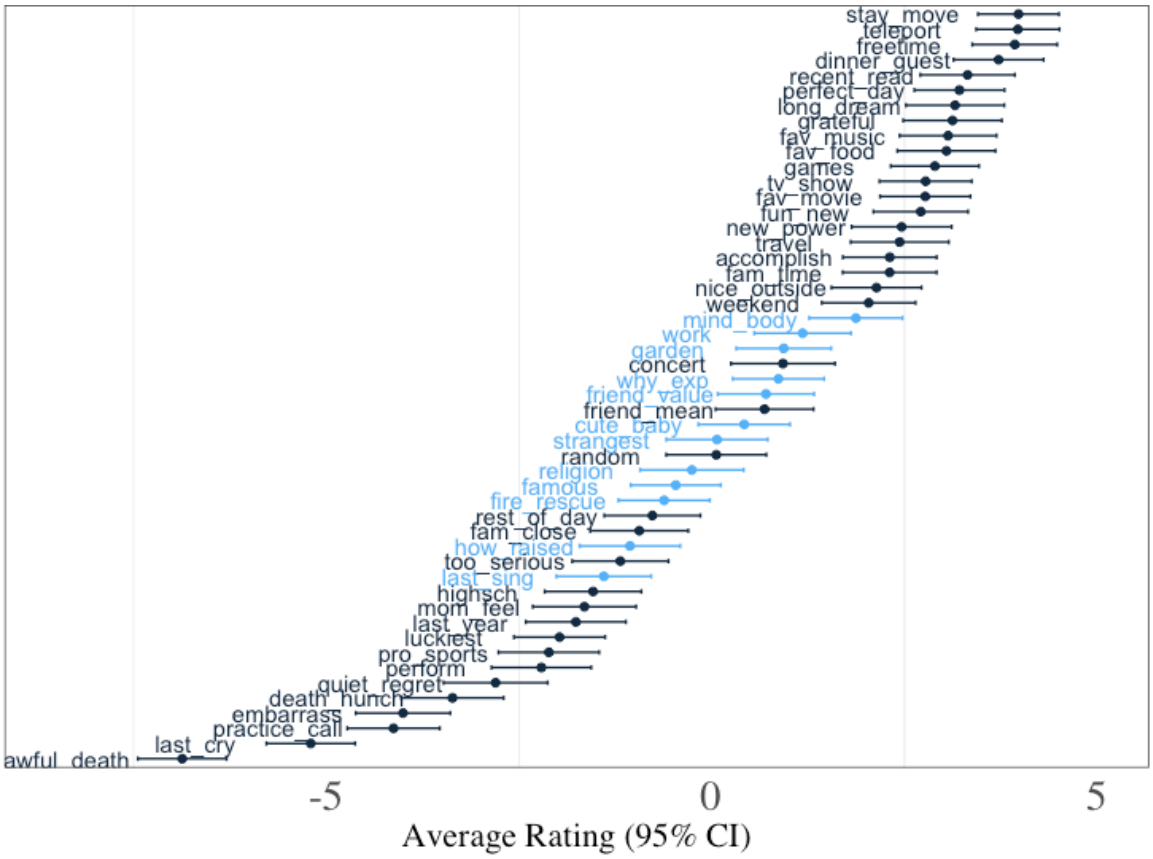
Before the study, we decided to select 12 topics, based on three main empirical criteria. In order of priority, we wanted to select topics with (i) a medium average preference rating, and (ii) high variance in preference ratings, as well as (iii) topics with higher average word counts. This was to increase the amount of preference heterogeneity among our participants. We thought that their perspective-taking ability would be better tested when there were no topics that everybody liked (disliked).

We recruited 300 mTurkers for this pilot. In Figure A1, we plot the average ratings (with 95% CI) of all 50 topics. This shows that while there are certainly some topics that are universally liked (e.g. “Do you like it where you live or do you want to move eventually?”) and

universally loathed (e.g. “Of all the people in your family, whose death would you find most disturbing? Why?”) there were many in a medium range, with meaningful variation across the population. Accordingly, we chose twelve topics that fit these criteria for the main studies in the paper.

Figure A1

This plot shows the mean rating (and 95% confidence interval) for all 50 topics used in the pilot study. The twelve that were chosen for the final study are in blue.



Appendix B: Topic Search Space

This is the large 50-topic list that was used in the pilot study, to determine the smaller 12-topic list for the main experiments.

1. What do you do for work? What do you like about it?
2. What do you enjoy doing in your free time?
3. Did you do any sports or clubs in high school?
4. Do you do these studies for fun, for money, or to advance science?
5. Do you have any plans for the weekend?
6. What's something random about you?
7. Have you read anything interesting recently?
8. Have you tried anything new recently that was particularly fun?
9. Are you a religious person? Why?
10. What games have you played in the past that are most memorable?
11. What is your favorite kind of music?
12. How do you most enjoy spending time with your family?
13. Do you have any fruit trees, plants, or a garden?
14. What's your favorite movie?
15. Do you have any plans for the rest of the day?
16. Do you like it where you live or do you want to move eventually?
17. Do you travel much?
18. What do you enjoy doing when the weather is beautiful?
19. Do you have a favorite type of food?
20. For what in your life do you feel most grateful? Why?
21. What was an embarrassing moment in your life?
22. What's the strangest thing about where you grew up?
23. Who is the luckiest person you know? Why?
24. If you could teleport by blinking your eyes, where would you go right now?
25. What is the last professional sports game or match you watched?
26. What is the last concert you attended? Why?
27. If you had to perform music in front of a crowd, what would you do?
28. What TV show have you watched lately?
29. What is the cutest thing you've seen a baby or child do?
30. Given the choice of anyone in the world, whom would you want as a dinner guest?
31. Would you like to be famous? In what way?
32. Before making a telephone call, do you ever rehearse what you are going to say? Why?
33. What would constitute a "perfect" day for you?
34. When did you last sing to yourself? To someone else?
35. If you were able to live to the age of 90 and retain either the mind or body of a 30-year-old for the last 60 years of your life, which would you want?
36. Do you have a secret hunch about how you will die?

37. If you could change anything about the way you were raised, what would it be?
38. If you could wake up tomorrow having gained any one quality or ability, what would it be?
39. Is there something that you've dreamed of doing for a long time? Why haven't you done it?
40. What is the greatest accomplishment of your life?
41. What do you value most in a friendship?
42. If you knew that in one year you would die suddenly, would you change anything about the way you are now living? Why?
43. What does friendship mean to you?
44. How close and warm is your family? Do you feel your childhood was happier than most other people's?
45. How do you feel about your relationship with your mother?
46. When did you last cry in front of another person? By yourself?
47. What, if anything, is too serious to be joked about?
48. If you were to die this evening with no opportunity to communicate with anyone, what would you most regret not having told someone? Why haven't you told them yet?
49. Your house, containing everything you own, catches fire. After saving your loved ones and pets, you have time to safely make a final dash to save any one item. What would it be? Why?
50. Of all the people in your family, whose death would you find most disturbing? Why?

Appendix C: Writer Prompts

The text of the topic writing and evaluation prompts from the pilot study and the preference detection studies. In all studies, participants wrote their response to all topics first, one at a time, and then circled back and evaluated their topic preference for all topics afterwards (also one at a time).

————————WRITING PAGE————————

Imagine you're meeting someone and having a friendly conversation, and s/he asks you the following question:

[topic question]

What would you say in response? Please write your answer in the box below. Your answer should be at least three sentences (and no more than ten sentences). To complete the task, you must write clearly, in full sentences with correct spelling/grammar/punctuation, as best you can.

————————EVALUATION PAGE————————

Please answer a few questions about the following conversation topic:

[topic question]

As a reminder, here is what you wrote in response to this question:

[written answer]

Imagine you were discussing this topic in a conversation. That is, imagine someone asked you the question above, and you responded with the same response you wrote for us. At this point in the conversation, would you want to talk more about this topic? Or would you want to switch to a new topic?

Please tell us your preference using the slider below, which ranges from -10 (strong preference to switch topics) to +10 (strong preference to stay on topic).

After my response, I would want to…

[—-————slider response————————]

Appendix D: Reader Prompts

The text from the screen for all for all twenty-four texts that each reader in Study 1 saw. They gave three judgments; predicting the writer's preference, stating their own preference, and predicting their behavioral response.

Imagine you're meeting this person and having a friendly conversation, and you asked them the following question:

[topic question]

Furthermore, imagine that their response to this question was:

[topic response]

What do you think their preference for this topic is? That is, do you think that they want to change topics, or continue talking about this topic?

[-10 to +10 slider : "switch to a new topic" to "stay on the current topic"]

What is your own preference for talking about this topic with this? That is, if the text above was a part of a real conversation you were having with this about this topic, would you prefer to keep talking about this topic, or would you prefer to switch to a new topic?

[-10 to +10 slider : "switch to a new topic" to "stay on the current topic"]

After this person says the text above, how would you respond? That is, would you keep talking about this topic with this person (e.g., by asking follow-up questions or adding your own thoughts) or would you try to change the topic (e.g., by asking a totally different question or starting to talk about something else)?

[1-7 likert response : "try to change the topic" to "keep talking about the topic"]

Appendix E: Instructions for Study 3

All participants in Study 3 were given a paper sheet, with one of two sets of instructions (switch often vs. natural, in italics). Below the instructions, they had all twelve topics listed in a random order. For any given pair, both people had exactly the same sheet.

For the next ten minutes, please have a friendly, open-ended conversation about the topics listed below.

*[**You and your partner should try to discuss all twelve topics during your conversation**. That is, your goal is to discuss all the topics before ten minutes are up. You can switch back and forth if you prefer, as long as you cover them all once.]*

*[**You and your partner can chat about as many or as few of these topics as you'd like**. That is, you don't have to discuss all the topics before the ten minutes are up. You could even stay focused on one or two of them for the whole conversation if you prefer.]*

During your chat, you should try to stick close to the topics below, though you can discuss the topics in any order you'd like. You should also feel free to discuss any questions, ideas, jokes, and stories related to these topics (we don't want you to feel overly constrained). Be natural. Be yourself. Have fun.

[list of 12 topics in random order]

Appendix F: Outcome measures in Studies 3 & 4

All participants in Studies 3 & 4 answered these blocks of questions after their conversations (and in Study 3, after they had evaluated the topics of conversation). All questions were answered on a likert scale from 1 (not at all) to 7 (very much). All participants answered each block of questions as written below, and then predict what their partner was going to say on the exact same questions. The order of questions was randomized within each block.

Liking Question Index

My partner is likeable.
I liked my partner.
I would enjoy spending time with my partner.
I dislike my partner.

Enjoyment Question Index

I enjoyed this conversation.
I thought this conversation was engaging.
I had an interesting conversation with this person.
I felt happy during this conversation.
I was watching the clock, wishing time would pass more quickly.

Impression Question Index

In general, I would describe my conversation partner as....

...caring
...sincere
...tolerant
...likeable
...good-natured
...confident
...intelligent
...competent
...independent
...competitive
...funny
...polite
...high status
...attractive
...warm