COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13271

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13271

# MEMCONS: How Contemporaneous Note-Taking Shapes Memory for Conversation

Sarah Brown-Schmidt,[a] Christopher B. Jaeger,[b] Melissa J. Evans,[a] Aaron S. Benjamin[c]

[a]*Department of Psychology and Human Development, Vanderbilt University*
[b]*Baylor Law School, Baylor University, Waco, USA*
[c]*Department of Psychology, University of Illinois*

Received 12 July 2022; received in revised form 21 December 2022; accepted 4 March 2023

**Abstract**

Written memoranda of conversations, or memcons, provide a near-contemporaneous record of what was said in conversation, and offer important insights into the activities of high-profile individuals. We assess the impact of writing a memcon on memory for conversation. Pairs of participants engaged in conversation and were asked to recall the contents of that conversation 1 week later. One participant in each pair memorialized the content of the interaction in a memcon shortly after the conversation. Participants who generated memcons recalled more details of the conversations than participants who did not, but the content of recall was equally and largely accurate for both participants. Remarkably, only 4.7% of the details of the conversation were recalled by both of the partners after a week delay. Contemporaneous note-taking appears to enhance memory for conversation by increasing the amount of information remembered but not the accuracy of that information. These findings have implications for evaluating the testimony of participants on conversations with major political or legal ramifications.

*Keywords:* Conversation; Recall; Memory; Contemporaneous notes

## 1. Introduction

The ability to recall details of prior conversational interactions plays an influential role in educational settings, interpersonal dynamics, dispute resolution, investigative reporting, and in governmental and legal proceedings (Davis & Friedman, 2007). Yet, empirical investigations of conversational memory show that the ability to accurately recall the details of what

Correspondence should be sent to Sarah Brown-Schmidt, Department of Psychology and Human Development, Vanderbilt University, 230 Appleton Place, Nashville, TN 37203, USA. E-mail: sarah.brown-schmidt@vanderbilt.edu

was discussed in conversation after delays of days or weeks is extremely limited and prone to bias (Neisser, 1981; Ross & Sicoly, 1979, *inter alia*). Because these limitations and biases of human memory are generally acknowledged, it is a common practice in professional settings to memorialize conversation through contemporaneous or near-contemporaneous notes, called *memcons*.

The political and historical significance of memcons in governmental affairs is evident in the many databases that contain memcons detailing conversations between various heads of state and other notable figures, including conversations concerning international affairs and national security from the Ford,[1] Nixon,[2] and Clinton[3] administrations. These conversations include many historically important ones, including President Ford and Henry Kissinger discussing U.S. foreign policy toward Israel and President Nixon welcoming the Soviet Women's Gymnastics Team to the Oval Office. In some cases, the memcon was created on the basis of notes taken during the meeting itself. In other situations, the memcon was created by a conversational participant shortly after the conversation took place, in which case it represents a person's written recall of the conversation after a brief delay. The latter type of near-contemporaneous note-taking after the conversation is over may be more common in cases where note-taking might interfere with ongoing discussion, or in situations where the writer wishes to conceal from their conversational partner their intent to memorialize the conversation.

As one example, the National Security Archive makes available records of back-channel meetings between Henry A. Kissinger, when he was Assistant to the President, and Anatoly Dobrynin, who was the Soviet Ambassador to the United States. Remarkably, the two men both created memcons after some of the same conversations, allowing for comparison of their individual recollections (see Burr, 1999; U.S. G.P.O., 2007).[4] For example, on September 25th, 1970, Kissinger and Dobrynin held both a morning meeting and an evening meeting in the Map Room at the White House. In between the two conversations, Kissinger consulted with President Nixon. A portion of the evening conversation between Kissinger and Dobrynin, as recalled after the fact by each man independently, is reproduced in Table 1.

Noteworthy are the multiple commonalities in the content of the memories, including both parties recalling that the dates of June or September 1971 were acceptable times for the summit. While both men recall discussing that August would *not* work, Kissinger recalls this discussion occurring in the morning meeting, whereas Dobrynin recalls it occurring in the evening meeting. In addition, while Dobrynin's recall is much more detailed, Kissinger provides more meta-commentary (e.g., "his face was ashen"). Without a recording of the conversation, it is impossible to know if, at the conclusion of the topic of the summit, Dobrynin in fact said "this was very good news" (as recalled by Kissinger), or if instead Dobrynin said he could not give "any promises" (as recalled by Dobrynin), or something else entirely.

Notes written during or shortly after forensic interviews can play an important role in legal proceedings when audio or video recordings of the interview are unavailable. However, these notes are often incomplete and thus an imperfect substitute for a genuine recording or transcript. For example, Lamb, Orbach, Sternberg, Hershkowitz, and Horowitz (2000) analyzed contemporaneous notes taken by professional investigators during real investigative interviews, comparing the notes with audiotaped recordings of the same interviews. Even these

Table 1
Edited excerpt from Documents 82-82, Kissinger and Dobrynin meeting, 09-25-1970

| IU | S | IU—Kissinger | IU | S | IU—Dobrynin |
|---|---|---|---|---|---|
| – | – | Summit. | – | – | – |
| – | – | When I saw Dobrynin in the Map Room his face was ashen. | – | – | – |
| 1 | K | I began the conversation by saying that I had the President's answer on the Summit | 1 | K | Kissinger, at his own initiative, again raised the question of a summit meeting. |
| 2 | K | and that the answer was as follows. | 2 | K | Citing the President's instructions, |
| 3 | K | In principle, the President was willing | 3 | K | Kissinger said that in Nixon's view […] |
| 4 | K | to consider a Summit. | 4 | K | the most acceptable time frame |
| 5 | K | Further, the President would consider | 5 | K | would probably fall |
| 6 | K | either June or September 1971 | 6 | K | during the period from |
| 7 | K | as appropriate dates | 7 | K | June through September of 1971. |
| am | **K** | **…in other words, whether it should be in June or in July or September,** | **8** | **K** | **Only August would be less desirable,** |
| am | **K** | **August probably being a vacation month for both sides**. | **9** | **K** | **since it is the traditional vacation time,** |
| am | **D** | **Ambassador Dobrynin stated that this was essentially correct**. | **10** | **K** | **although it would be possible to agree on that time frame too**. |
| 8 | K | and the U.S. Government was willing to consider | 11 | K | As for the possible venue of such a meeting, |
| 9 | K | Moscow as the site | 12 | K | since President Nixon |
| 10 | K | for such a meeting. | 13 | K | had already invited the Soviet leaders |
| 11 | D | Ambassador Dobrynin said this was very good news. | 14 | K | to a meeting on U.S. territory |
| – | – | But, he clearly had his mind on the Cuban problem. | 15 | K | in October of this year, |
| | | | 16 | K | he expressed his willingness |
| | | | 17 | K | to travel to the Soviet Union in 1971 […] |
| | | | 18 | D | I told Kissinger I could not give him any promises |
| | | | 19 | D | about the timing of our reply [….] |

*Note*: Idea Units (IUs) are numbered consecutively for each memcon excerpt, and alignment of the two memcons is approximate. Attributed source (S) of utterance is Kissinger (K) or Dobrynin (D). Note that Dobrynin's recall is much more detailed and some sections are replaced by […] for illustration purposes. Also note that Dobrynin's IU#8-10 (bolded) appear in Kissinger's recall of the *morning* meeting, not the afternoon meeting (denoted with AM in Kissinger's memcon). https://nsarchive2.gwu.edu/NSAEBB/NSAEBB233/9-25-70.pdf

professional memcons failed to include 25% of the forensically relevant details that were provided by the interviewee.

As we shall see, the experimental literature on conversational memory can be consulted in order to evaluate the value of memcons as archival evidence. Yet, focusing solely on whether memcons themselves are complete and accurate ignores a fundamental lesson from cognitive psychology—namely, that remembering is not simply a report of memory, but an act that changes memory (Benjamin & Pashler, 2015; Bjork, 1975; Roediger & Karpicke, 2006a). In this report, we investigate the completeness and accuracy of memcons created shortly after a conversation, the completeness and accuracy of substantially delayed recall of the same conversation by the generators of memcons and by others who did not memorialize the conversation, and the relationship between the two events. What emerges is a clearer picture of the impact of memcons in high-stakes recall, like sworn testimony.

### *1.1. Conversational memory*

Conversational memory is characterized by a tendency to remember the gist of what was said, rather than verbatim details (Sachs, 1974). A standard measurement approach in this area of research is to code the conversation in terms of IUs expressed in the conversation (Stafford, Waldron, & Infield, 1989). The conversational IUs are then compared to IUs in the recall to calculate the completeness and accuracy of recall. In contrast to much research on human memory, the information conveyed in conversation is considered "recalled" even if the recaller is only able to reproduce the gist of what was said, and not the verbatim word-for-word utterance. After delays of minutes to days, conversational participants can accurately recall only a small percentage of conversational IUs, with estimates ranging from 5% to 20% recall (Benoit & Benoit; 1988; Ross & Sicoly, 1979; Samp & Humphreys, 2007; Stafford & Daly, 1984). For example, Stafford, Burggraf, and Sharkey (1987) had participants engage in a 7-min conversation, watch a short distracting film about Hawaii, and then recall the conversation. Participants recalled 10% of the IUs immediately after the film, and only 4% on a second test 4 weeks later. Taken together, these findings indicate that conversational recall is highly incomplete, even after short delays.

Memory for conversation also exhibits a bias such that memory for what one said tends to be superior to what was heard (Fischer, Schult, & Steffens, 2015; Isaacs, 1990; McKinley, Brown-Schmidt, & Benjamin, 2017; Miller, deWinstanley, & Carey, 1996; Zormpa, Brehm, Hoedemaker, & Meyer, 2019). This reflects the more general phenomenon that the generation and production of information promote memory (MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010; Slamecka & Graf, 1978). This bias can even shape assessments of one's centrality in a conversation: though John Dean was found to have accurately remembered the gist of historically important conversations with President Richard Nixon, he consistently tended to overstate the importance and centrality of his own role in those conversations (Neisser, 1981).

Beyond explicit recall, memory for the content of a conversation also reveals itself in the shaping of linguistic form, including the subsequent use of definite versus indefinite expressions (the bagel vs. a bagel), the length and descriptiveness of an expression, and whether an

expression is repeated (Clark & Wilkes-Gibbs, 1986; Duff et al. 2006; Knutsen & Le Bigot, 2021; Yoon & Stine-Morrow, 2019). Memory for what has been discussed is also broadly assumed to support representations of common ground—representations of what information and beliefs are mutually known to conversational partners (Clark & Marshall, 1978). It is this knowledge of shared conversational history that allows conversational partners such as Kissinger and Dobrynin to refer to "the summit" and mutually understand what they are referring to.

Finally, conversational memory includes not only what was discussed in the past, but also includes memory for who said something, a type of source memory, and who it was said to, a type of destination memory (Gopie & MacLeod, 2009). Accurately tracking source and destination memory in conversation are essential components of everyday conversation, supporting audience design processes, such as clarifying which "summit" you are talking about to a new conversational partner who was not present for the prior discussion (Brennan & Clark, 1996; Wilkes-Gibbs & Clark, 1992). Similarly, tracking source memory is critical to accurately attributing individual contributions to a conversation, such as recalling who thought of the solution to a group problem (Foley, Foley, Durley, & Maitner, 2006; McKinley et al., 2017).

### *1.2. Effects of retrieval practice on memory*

The facts that conversational recall is limited and that completeness may drop as time passes prior to a memory report (Stafford et al., 1987) point to the utility of writing contemporaneous notes as a way to memorialize conversation. A key question, then, is whether taking contemporaneous notes can mitigate rapid memory loss for information from the conversation.

While we know of no evidence that directly addresses this question in the domain of unscripted conversation, the literature on *retrieval practice* makes the clear prediction that attempts to recall information enhance long-term memory (Roediger & Karpicke, 2006a, 2006b; Siler & Benjamin, 2020). When a memcon is created following a conversation, the act of retrieving the contents of the conversation from memory is very likely to improve memory for that information given the robustness of the retrieval practice effect. The fact that creating a memcon also involves the *production* of that information (i.e., in writing) also points to a likely beneficial effect of memcon generation on memory due to the production effect (MacLeod et al., 2010). The timing of when a memcon is created may be relevant to its impact on memory. For example, Aiken, Thomas, and Shennum (1975) reported superior memory for the gist of a video after 2 days in a group that recorded notes after the video compared to groups that either took no notes or took notes during the video itself. Such findings might indicate that the act of creating memcons during a meeting could have some disadvantages compared to delayed memorialization (see Einstein, Morris, & Smith, 1985).

A key difference, however, between these paradigms and interactive conversation is that conversational participants actively and interactively produce and receive information (Brennan, Galati, & Kuhlen, 2010) in a contextually sensitive manner. It is unclear whether retrieval practice would yield comparable benefits to conversational memory given that conversation

has a more complex structure and higher levels of social involvement than reading a passage or watching a video.

### *1.3. The present research*

What is the impact of taking contemporaneous notes on memory for conversation? To answer this question, we designed an experiment in which pairs of participants conversed on a topic, with one participant in each pair asked to recall that conversation in a written record (memcon) 5 min later. Memory for conversation was tested after 1 week for all participants using an oral recall procedure. We intentionally examine situations where a person memorializes the conversation in writing, followed by a delayed oral recall, in order to mimic real-world scenarios where a person generates a memcon to memorialize a conversation, and is later asked to recall that conversation orally, for example, in a forensic interview or in oral testimony. The present research represents the first attempt in the literature to examine the impact of contemporaneous notes on memory for conversation in a way that balances the benefits of precise experimental control with the high levels of ecological validity necessary to capture the dynamic, interactive properties and natural complexity of unscripted conversation.

## 2. Experiment

### *2.1. Method*

This study was run in 2018 at the first author's laboratory at Vanderbilt University. This study was preregistered at https://osf.io/s8vwx.

#### *2.1.1. Participants*

The preregistered study design called for 20 pairs of participants. This sample size was chosen to balance power to detect effects of reasonable size with the feasibility of collecting and coding the massive quantity of data associated with conversation in a reasonable amount of time. Participants were recruited through the Vanderbilt University research participant pool and were compensated with partial course credit or $30 for participation. In order to recruit pairs of participants who were not close friends or cohabitating, participants signed up for the study individually, with two participants scheduled for each session.

A total of 33 pairs of participants were recruited to participate; however, 12 pairs were excluded from analysis due to recording equipment failure ($n = 5$), to participants revealing that they lived together or were close friends and would, therefore, have an opportunity to discuss the study before the final recall ($n = 4$), or to experimenter error in administering the study (conversation lasted too long, or testing delayed by 2 weeks, $n = 3$). Of the 21 pairs of participants included in the final analysis, one participant from each of two pairs failed to return for the 1-week delayed recall sessions, and the data from a third participant's delayed recall session were lost. As a result, it was not possible to calculate the similarity of conversational recall for these three pairs. While our final sample of 21 pairs oversamples by one pair

compared to the analysis plan, this was considered acceptably close to the preregistration due to the loss of data from three participants at the 1-week delay.

### 2.1.2. Procedure

Pairs of participants completed a two-session study (see Supplementary Material for a timeline of the procedure, Table S1). At the first session, after providing informed consent, participants were seated in "Room 1" in the laboratory and instructed to engage in a 10-min conversation task. The conversation was audio recorded and participants were given two topics to get the conversation going: current events and living in Nashville. Most pairs discussed both topics during the 10-min period. After conversing for 10 min, the participants were separated and walked around the building, chatting with an experimenter for 5 min.[5] This 5-min filled delay was intended to mimic a real-life scenario in which a person engages in a conversation, then walks to another location before writing a memcon. Next, the two participants were directed to separate rooms (Rooms 2 and 3) and asked to fill out a Likert-type scale reflecting their interest in the prior conversation with the other student. Within each pair, one participant was randomly assigned to the role of Participant A and the other participant was assigned the role of Participant B. Participant A was then given a 15-min surprise recall task. Participants were instructed to "Please write as detailed of an account of your conversation with the other participant as you can. It is important to identify *who* said *what* where you are able to do so, so to the best of your ability write it like a play." Meanwhile, Participant B was prompted to write about career, family, and hobbies for 15 min. All participants typed their responses on laboratory computers. These tasks concluded Session 1.

Following a 1-week delay, participants were brought back to the lab individually for a second session in a new room (Room 4). Participants first filled out a Likert-type question indicating how well they remembered the conversation from a week prior. Then, the videorecorder was turned on, and they were given 10 min to recall out loud the prior conversation.

Data in this study included Likert-style judgments about the conversation and memory for it, as well as detailed analyses of the conversation and how it was recalled following delays of 5 min (Partner A only), and a 1 week (both Partner A and Partner B). Our analyses focus on memory for the ideas expressed in the conversation, as well as the relationship among these measures of memory within and across individuals.[6]

### 2.1.3. Coding of conversations and recalls

Audio recordings of each conversation were transcribed word-by-word, with utterances labeled for who said what (Participant A or B). Following prior work in conversational memory, utterances were broken up into IUs (Ross & Sicoly, 1979; Stafford & Daly, 1984; Stafford et al., 1987; 1989). An IU corresponds to "the smallest unit of meaning that has informational or affective value; it represents the gist of each thought expressed by the interactants" (Stafford et al., 1989, p. 600). Typically, an IU corresponds to a simple phrase that expresses an idea. The following deidentified excerpt from one of the conversations illustrates the IUs in the original conversation, as well as how they were eventually recalled (or not). The snippet of conversation in Table 2 illustrates a conversation where participants are discussing Nashville hot chicken:

Table 2
Example conversation and recall

| Speaker | Conversation | A: written recall | A: oral recall | B: oral recall |
|---|---|---|---|---|
| **B** | **I think it tastes gross** | but I actually think its pretty gross. | um she said it was just just gross | - |
| **A** | **Really oh?** | – | – | – |
| **A** | **I guess it's kinda–** | – | – | – |
| **A** | **that's sorta fair,** | – | – | – |
| **A** | **it's definitely hard on like the body,** | Its definitely hard on the body | – | – |
| **A** | **you know what I mean?** | – | – | – |
| **A** | **It's like fried chicken** | – | – | – |
| **A** | **dipped in like spicy grease** | with the grease and spice. | you know it's spicy, it's buttery, it's uh greasy | |
| **B** | **Well I—it doesn't have a lot of flavor** | I just think it has no flavor | – | it has like no flavor |

*Note*: A deidentified excerpt from the conversation is broken into idea units (second column from left), and matched to idea units at each of the three recalls (right three columns). Dash marks (–) indicate idea units that were not recalled.

*2.1.3.1. Coding of IUs:* For each conversation, there were three separate recalls, two from Participant A (written recall at 5-min delay; oral recall at 1-week delay), and one from Participant B (oral recall at 1-week delay). The three recalls were broken up into IUs in the same manner as the conversation and compared to the conversational IUs to characterize the accuracy and completeness of recall. After the initial coding was complete, a second coder checked all of the coding for accuracy. To calculate intercoder reliability, a third coder independently coded two conversations and associated recalls, masked to the original coding. Of all IUs produced in these two conversations, the coders were in agreement 92% of the time as to whether that IU was later recalled or not, corresponding to a Kappa of .798 (*SE* = .015) and indicating substantial agreement (calculated using GraphPad QuickCalcs Web site, accessed January 2022: https://www.graphpad.com/quickcalcs/kappa1/). Of all IUs that were produced at recall in the two conversations, the coders were in agreement 83% of the time as to whether that IU reflected an accurate recall of something said in the conversation, corresponding to a Kappa of .634 (*SE* = .025), and again reflecting substantial agreement. Next, the third coder reviewed the coding for the entire dataset, revealing a total of 336 coding disagreements out of 10,367 total IUs produced in recall sessions (3%). These disagreements were resolved through discussion, and the resulting dataset was used in the final analysis.

*2.1.3.2. Measures of remembering:* For each of the IUs expressed in the conversation, we coded whether this idea was accurately produced in each of the three recall tests. Following the prior literature, this coding used gist recall, meaning that an IU was coded as being recalled if the gist meaning of the original IU was recalled, even if the precise wording was not recalled. For example, the original conversational utterance "*dipped in like spicy*

*grease"* was counted as correctly recalled by Participant A in the written recall when they wrote "…*with the grease and spice*," and was also counted as correctly recalled by Participant A 1 week later when the said "*you know it's spicy, it's buttery, it's uh greasy."* The calculation of the number of IUs from the original conversation that were later recalled (either after 5 min or 1 week) allows us to calculate the probability that an individual IU expressed in the initial conversation was recalled. This is a measure of the completeness of recall.

We also coded whether each produced component of recall was an accurate reflection of a specific IU in the conversation (correct recall), a new IU (incorrect recall), or a commentary that was germane to the conversation but not a recall of a specific IU expressed in the conversation. Correct recalls reflect the gist or meaning of what was expressed in the conversational IU, even if the wording is different. Incorrect recalls reflect information that was not present in the original conversation. For example, during a different portion of Participant A's oral recall from Table 2, Participant A asserted that "*she said she's from like the southwest*." This information was not expressed in the original conversation, and it was coded as incorrect. Commentaries reflect meta-comments, opinions, extrapolations, and broad statements that do not reflect recall of specific IUs in the conversation, but rather commentary and extensions from it. For example, after relaying part of the conversation in the oral recall, Participant A offered an opinion about something B had said "*which was kind of funny*." Likewise, Participant B accurately recalled discussing a pancake restaurant but then stated "*it sounded like he had eaten there before*," which reflects an extrapolation or inference from a portion of the conversation, rather than a recall of something that was actually stated. The proportion of IUs expressed in the recall that are an accurate representation of an idea expressed in the original conversation, out of all correct and incorrect IUs (but excluding commentaries) is a measure of accuracy. We also calculated the accuracy of the source memory judgments, specifically how often participants correctly attributed the correct contributor of an IU, when it was recalled.

*2.1.3.3. Concordance in memory reports:* Concordance in recall is important for understanding the degree to which individuals hold a common understanding of a conversation and how those ideas diverge. Concordance in recall was first characterized by descriptive statistics that traced whether a conversational IU was recalled in the memcon, and if so, whether those IUs were recalled by Participants A and B a week later. The concordance of recall can be quantified using a measure of mnemonic similarity that counts the number of IUs recalled at both recall sessions or not recalled at either recall session out of the total IUs in the conversation (Coman, Momennejad, Drach, & Geana, 2016). For Participant A, we separately measured reminiscence (the proportion of all produced items that were recalled on the oral test, after a delay, but not on the immediate written recall) and oblivescence (the proportion of IUs that were recalled on the immediate written test but not on the oral test, after a delay; e.g., Erdelyi, 2010). Lastly, inferential statistics address our preregistered research questions regarding the question of whether writing the memcon promotes A's memory for the memorialized IUs, and whether the resultant memory benefit extends to the conversation more generally.

### 2.2. Results

We first present an analysis of the judgment data. We then turn to the primary analysis, which focuses on conversational memory. For each analysis below, we specify whether it was or was not included in the preregistration analysis plan. In the years between the preregistration and the data analysis, a number of new empirical discoveries, and our rediscovery of older empirical findings motivated us to include a number of analyses that were not preregistered.

#### 2.2.1. Analysis of judgment data

Analysis of the judgment data was not a part of our preregistration nor a focus of our analyses; the descriptive statistics are presented here for completeness.

Following the conversation and 5-min delay, participants responded to a Likert scale question, "What was your level of interest in the conversation you had with the other study participant?" on a scale of 1 to 7, with higher values indicating more interest. Excluding one Participant B who did not fill out the rating form, participants found the conversations to be moderately interesting on average, $M_A = 4.74$ ($SD = 1.04$), $M_B = 4.85$ ($SD = 1.09$).

After the 1-week delay, participants responded to a Likert scale question, "How well do you remember the conversation you had with the other study participant last week?" on a scale of 1 to 7, with higher values indicating better memory. Excluding the two participants who did not return for the second session (one Participant A and one Participant B), participants found the conversations to be moderately well remembered, $M_A = 3.30$ ($SD = 0.98$), $M_B = 4.25$ ($SD = 1.25$), with somewhat *lower* memory ratings by the participant (A) who generated the memcon.[7] Participant A's lower estimates of their own memory may relate to the difficulty that participants experienced while recalling the conversation after only 5 min.

#### 2.2.2. Conversation descriptive statistics

Across the 21 conversations, speakers produced a total of 9258 IUs, corresponding to approximately 42,025 words. On average, there were 441 IUs ($SD = 79$) and approximately 2001 words ($SD = 285$) per conversation. The written recalls, which were produced by Participant A after a 5-min delay, included a total of 2928 IUs ($M = 139$ per recall, $SD = 26$). Oral recalls by Participants A and B after a 1-week delay included 4075 (204 per recall, $SD = 74$) and 3364 IUs (177 per recall, $SD = 60$), respectively.

#### 2.2.3. Completeness of recall

Fig. 1 plots the probability that each of the conversational IUs was recalled, separately for the three recall sessions (see Fig. S1 for an illustration of order effects). Overall, the percent of the conversational IUs that were recalled by Participant A in the initial written recall was higher (24%) compared to A's oral recall a week later (18%) and B's oral recall after a week (13%). Note this study was designed with intentional format differences between the written (15-min period of typing) and the oral recall (10-min period of oral recall); the amount of forgetting between the two tests is confounded with this format manipulation and cannot be precisely estimated.

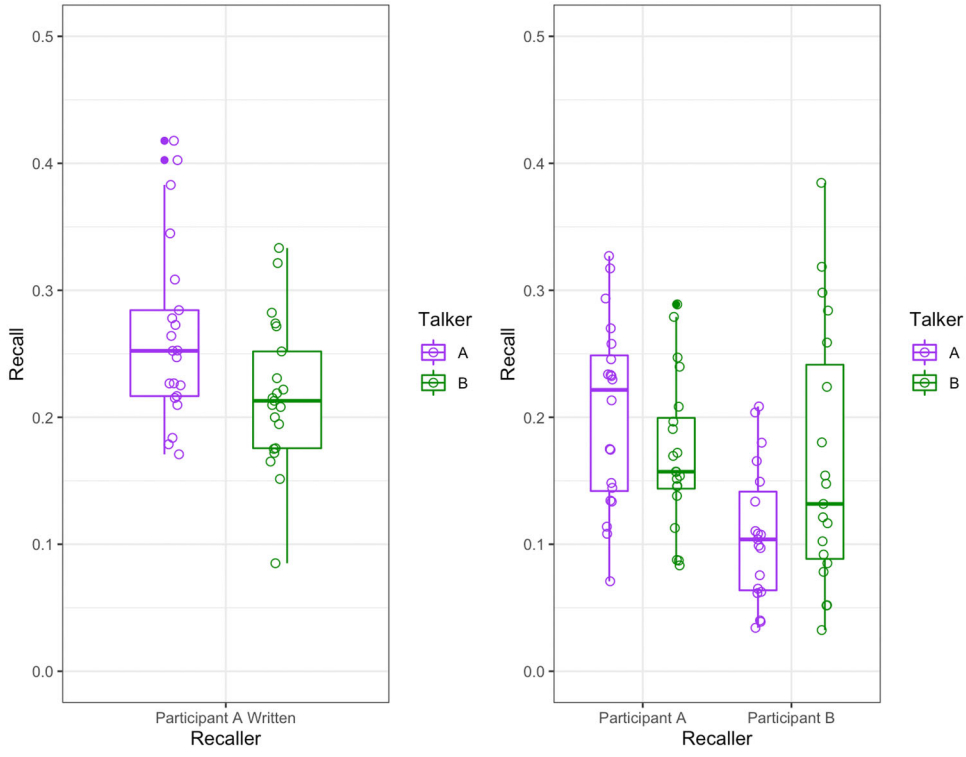


Fig. 1. Proportion of 9258 conversational IUs that were recalled by Participant A in the written recall (left panel), and by Participant A and Participant B in the oral recall at a 1-week delay (right panel), split by which Talker (A or B) uttered that IU. Distribution of data illustrated with box-plot, and circles representing individual participant means. Data plotted using ggplot2 (Wickham, 2016).

The dependent measure for the comparison of oral recalls is whether (1) or not (0) each of the 9258 conversational IUs was recalled by Participants A and B at the 1-week oral recall. A mixed-effects logistic regression analysis was used to analyze the recall data using the glmer function in lme4 (Bates et al., 2015) in R version 4.1.2 (R Core Team, 2021). A parsimonious model structure was determined using the buildmer function (Voeten, 2020; see Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). The initial input to buildmer included effects of Talker (A or B), Recaller (A or B), and their interaction; both were effects coded (A = .5 and B = –.5). In reviewing our preregistration, we realized we neglected to include Talker as a preregistered factor. It is well-known that conversational memory tends to be better for what one has said themselves (e.g., Ross & Sicoly, 1979); thus, we included this factor in the final model. In addition, we included a measure of IU length in terms of the number of words (centered and scaled/10 to avoid convergence issues), Idea Unit serial order (centered separately for each group and scaled/100), and a quadratic function of Idea Unit serial order as control variables. Although these control variables were not a part of the preregistered

Table 3
Mixed-effects logistic regression analysis of recall completeness including 17,114 binary observations, 21 participant groups, and 39 participants

| **Fixed effects** | **Estimate** | ***SE*** | ***z*-value** | ***p*-value** | |
|---|---|---|---|---|---|
| (Intercept) | –1.947 | 0.121 | –16.142 | <.0001 | |
| IU order | –0.300 | 0.060 | –5.006 | <.0001 | |
| Recaller | 0.435 | 0.101 | 4.327 | <.0001 | |
| IU length | 0.048 | 0.204 | 0.236 | 0.814 | |
| Talker | –0.094 | 0.073 | –1.288 | 0.198 | |
| IU orderˆ2 | 0.023 | 0.039 | 0.576 | 0.565 | |
| Recaller*Talker | 0.704 | 0.147 | 4.806 | <.0001 | |
| **Random effects** | **Variance** | ***SD*** | **Corr.** | | |
| Group (intercept) | 0.234 | 0.484 | | | |
| IU length | 0.636 | 0.797 | 0.15 | | |
| Participant (intercept) | 0.074 | 0.273 | | | |
| IU order | 0.121 | 0.348 | 0.09 | | |
| IU orderˆ2 | 0.046 | 0.215 | –0.33 | 0.50 | |
| Talker | 0.120 | 0.346 | –0.50 | 0.14 | 0.05 |

*Note*: Fixed effects of Recaller and Talker were effects coded and control variables of IU number and IU word count were mean-centered and scaled.

analysis plan, we made the decision to include them as control variables prior to evaluating the effects of their inclusion (or exclusion).

The final model (Table 3) included a negative intercept term ($b = –1.95$, $p < .0001$), indicating that IUs were more likely to be not recalled than recalled. A negative effect of IU order ($b = –0.30$, $p < .0001$) indicates that IUs earlier in the conversation were more likely to be recalled than those later in the conversation. An effect of Recaller was due to Participant A recalling more IUs on the week-delayed oral test than Participant B ($b = .43$, $p < .0001$), demonstrating a memory benefit from having generated a memcon. A significant interaction between Talker and Recaller ($b = .70$, $p < .0001$) was due to a bias to selectively remember one's own contributions to the conversation for both Participant A ($b = .26$, $p = .009$) and Participant B ($b = –.45$, $p < .0001$).[8]

A supplemental analysis which was not preregistered compared completeness of recall for Participant A in immediate written recall compared to oral recall using the same model-fitting procedures described above. In addition to the previously described findings, this model revealed that A's written recall was more complete than the oral recall after a 1-week delay ($b = 0.41$, $p < .0001$), corresponding to increased odds of an IU being recalled at the earlier time-point of 1.51. Of course, these findings must be interpreted with caution given the previously noted differences in format between the two tests.

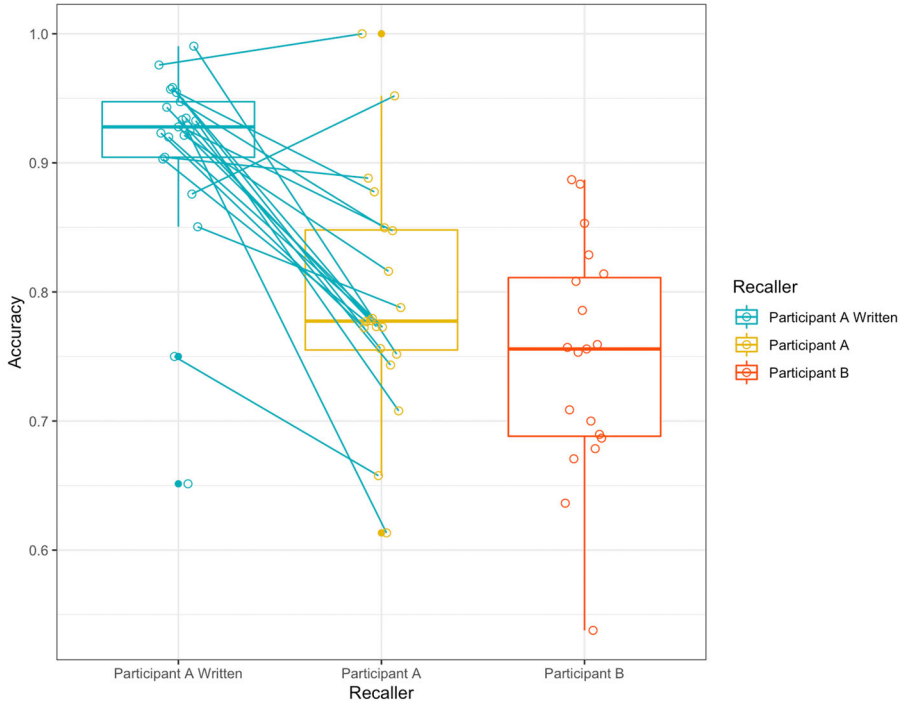


Fig. 2. Proportion of 6806 recall IUs that were accurate by Participant A in the written recall, and by Participant A and Participant B in the oral recall at a 1-week delay. Distribution of data illustrated with box-plot and data-points representing individual participant means. Lines connect individual Participant A's written and oral recall.

### 2.2.4. *Recall accuracy*

The second major analysis focuses on the accuracy of recalled information. As mentioned above, the recall sessions included a large number of commentaries, expressed opinions, and other remarks that were clearly not an attempt to recall content from the conversation. Across the three recall sessions, there were a total of 6806 recall IUs (written: 2564; A oral: 2478, B oral: 1764) out of 10,367 total IUs expressed in the recall, with the remaining recall units reflecting opinions and other commentary. Our analyses focus on these recall IUs. Overall, recall IUs were more accurate than not (Fig. 2), with accuracy rates of 91% in Participant A's initial written recall, dropping to 78% and 76% accuracy for Participant A and B's oral recalls after 1 week, respectively.

As described in the preregistration, we first compared the accuracy of recalls between Participants A and B at the 1-week delay. The dependent measure was whether (1) or not (0) each of the 4242 recall attempts by Participants A and B at the 1-week oral recall was an accurate reflection of something said in the conversation. A mixed-effects logistic regression analysis was constructed as before. The initial input to buildmer included a fixed effect of Recaller (A or B) only. Note that for inaccurate recalls, the additional variables included in the completeness analysis (Talker, IU length, and IU order) are undefined, and, therefore, not included in

the present analysis. The buildmer function returned a model that included an intercept that varied randomly by participant, with no other fixed or random effects. For completeness, we added an effect of Recaller to the final model (see Supplemental Materials, Table S2). Participants were more accurate than not, with the odds of accurate to inaccurate responses of 3.59 ($b = 1.277$, $p<.0001$). The effect of recaller was not significant ($b = .267$, $p = .124$), indicating that accuracy was similar regardless of whether the participant had completed the written recall task week prior.

A supplemental analysis which was not preregistered compared the accuracy of recall for Participant A in the immediate written recall compared to the oral recall. Using the same model-fitting procedures described above, the final model revealed that the initial recall was significantly more accurate ($b = 1.09$, $p <.0001$), corresponding to an increased odds of an accurate response at the earlier time-point of 2.98. This result supports the intuitive view that underlies the decision to memorialize conversations shortly after they take place: memcons contain both more information, and more accurate information, than memory is able to provide later.

### *2.2.5. Source memory accuracy*

The source attribution data had a polytomous structure, with 15% of the relevant productions containing either an ambiguous attribution or no attribution to a particular talker (e.g., "we started talking about taco places"). The remaining IUs were attributed to either Participant A or Participant B (e.g., "she said she didn't own cowboy boots"; "I applied to be a dog-walker too"). An exploratory analysis of source memory accuracy was conducted on the 3281 accurately recalled IUs by Participants A and B during oral recall (see Fig. S2). The source data were coded in polytomous form, distinguishing accurate source attributions, inaccurate attributions, and cases where the participant did not attribute the IU to an individual. These data were recoded in binary form at two nodes, and analyzed using a multinominal processing tree GLMM (see Cho, Brown-Schmidt, De Boeck, & Shen, 2020). Node 1 in the model distinguished cases where the participant made an attribution (1) versus not (0). At Node 2 in the model, nonattributions are not modeled, as Node 2 distinguishes between accurate attributions (1) versus inaccurate attributions (0). In addition to a Node covariate, node-specific fixed effects included Recaller, Talker, number of words in the IU, IU order, IU order-squared, and the interactions. A parsimonious model structure was determined using the buildmer function as before. The results of that process are shown in Table 4.

At Node 1, in this model, attributions occurred more than not ($b = 2.34$, $p<.0001$). An effect of Recaller at Node 1 was due to more attributions by Participant A than B ($b = 1.11$, $p = .014$). The buildmer process also included both the linear and quadratic terms for IU order; only the latter was significant ($b = –0.45$, $p = .002$).

At Node 2, accurate attributions were more common than inaccurate attributions ($b = 2.57$, $p<.0001$). An effect of IU length in words was due to greater accuracy with longer IUs ($b = .82$, $p = .002$). Exploration of a two-way interaction between Talker and Recaller at Node 2 ($b = –0.94$, $p = .002$) revealed that, for Participant A, accuracy of source attributions did not differ significantly as a function of talker ($b = –0.28$, $p = .15$). For Participant B, however, source attributions were more accurate for A's speech than their own ($b = .66$, $p = .004$).

Table 4
Results of logit-link multinomial processing tree DGLMM with two nodes for source memory data

| **Fixed effects at Node 1 (attribution)** | **Estimate** | ***SE*** | ***z*-value** | ***p*-value** | |
|---|---|---|---|---|---|
| Node 1 intercept | 2.342 | 0.225 | 10.423 | <.0001 | |
| Recaller | 1.105 | 0.450 | 2.454 | 0.014 | |
| IU order | 0.012 | 0.260 | 0.045 | 0.964 | |
| IU orderˆ2 | –0.447 | 0.143 | –3.133 | 0.002 | |
| IU length (words) | 0.379 | 0.207 | 1.835 | 0.066 | |
| Talker | –0.051 | 0.118 | –0.428 | 0.668 | |
| Recaller*Talker | –0.338 | 0.237 | –1.427 | 0.154 | |
| **Fixed effects at Node 2 (accuracy)** | **Estimate** | ***SE*** | ***z*-value** | ***p*-value** | |
| Node 2 intercept | 2.573 | 0.145 | 17.683 | <.0001 | |
| Recaller | 0.337 | 0.279 | 1.210 | 0.226 | |
| IU order | –0.152 | 0.157 | –0.966 | 0.334 | |
| IU orderˆ2 | –0.152 | 0.155 | –0.984 | 0.325 | |
| IU length (words) | 0.823 | 0.269 | 3.059 | 0.002 | |
| Talker | 0.189 | 0.149 | 1.263 | 0.207 | |
| Recaller*Talker | –0.935 | 0.300 | –3.120 | 0.002 | |
| **Random effects** | **Variance** | ***Std. Dev.*** | **Corr** | | |
| Participant | 1.611 | 1.269 | | | |
| Node 2 intercept | 0.920 | 0.959 | –0.87 | | |
| Node 1: IU order | 1.661 | 1.289 | 0.16 | –0.08 | |
| Node 2: IU order | 0.082 | 0.286 | 0.51 | –0.39 | –0.29 |

*Note*: Node 1 models whether the participant made a source attribution or not (accurate = 1, inaccurate = 1, no attribution = 0); Node 2 models whether a source attribution was accurate or not (accurate = 1; inaccurate = 0). The model includes 39 participants and 6068 binary observations across two nodes (3281 at Node 1, 2787 at Node 2; a smaller number of datapoints is modeled at Node 2 because IUs where the participant did not make an attribution are removed at Node 2).

This finding may be due to a tendency for B to not attribute a source when recounting their own speech (e.g., "we talked about…").

### *2.2.6. Memory concordance*

The similarity of recall, or concordance in memory, can be calculated among the three time points. Our preregistered analyses include an examination of the relationship between what is recalled in the memcon, and what is recalled a week later, by A and B. Before describing that analysis, we first present some descriptive statistics. The concordance analysis is restricted to the data from 18 pairs for whom the dataset is complete. A total of 7856 total IUs appeared in those 18 conversations (Fig. 3).

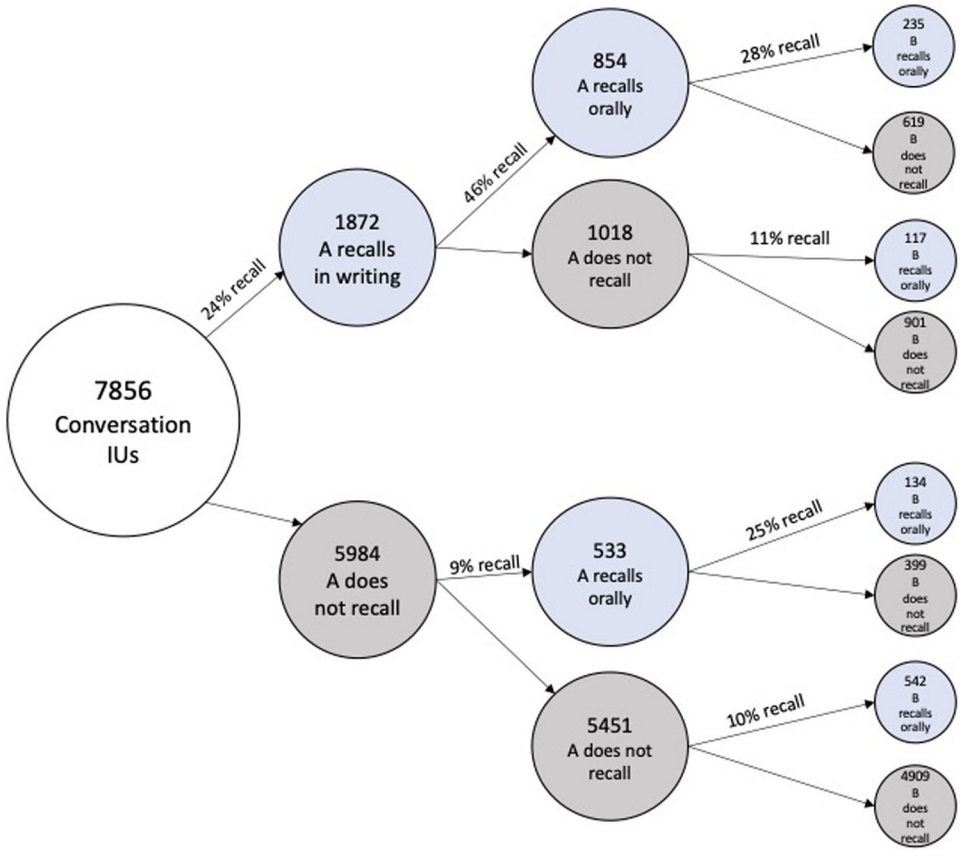


Fig. 3. Illustration of the relationship between oral recalls for the 18 complete pairs.

*2.2.6.1. Similarity in recall:* To characterize the similarity in recall, we assessed the degree to which an IU is likely to be recalled by two parties, or by a given Partner A at on the two recalls. We also calculate the mnemonic similarity (Coman et al., 2016), which takes into account similarity in recall successes and failures by counting the number of IUs recalled at both recall sessions (RR) or not recalled at either recall session (NN) out of the total IUs in the conversation. Specifically, we calculated the following similarity metrics for the similarity of A's written and oral recall ($MS_{AA}$), and A and B's oral recalls ($MS_{AB}$):

$$MS_{AA} : (RR + NN) / total = (854 + 5451) / 7856 = 80.3\%$$

$$MS_{AB} : (RR + NN) / total = (369 + 5810) / 7856 = 78.7\%$$

First, we examined the relationship between Participant A's written recall and their oral recall 1 week later. Of the 1872 IUs that were recalled in the memcon, 854 (46%) were recalled by A 1 week later. Of the 5984 IUs that were not recalled in the memcon, 533 were

recalled 1 week later (9%). Thus, as is typical, reminiscence was less common than oblivescence (e.g., Stanley & Benjamin, 2016). The fact that a large number of IUs were never recalled by A results in a fairly high measure of similarity.

Next, comparing A and B's oral recall, of the 1387 IUs that appeared in A's oral recall, 369 (27%) also appeared in B's oral recall. By contrast, of the 6469 IUs that did not appear in A's oral recall, only 659 (10%) appeared in B's recall. This suggests some similarity in which IUs A and B tended to recall in the week following the conversation. Note, however, that of the 7856 IUs that were uttered in the original conversation, only 369 (4.7%) were recalled by both A and B a week later. This result suggests that after a delay, only a small portion of a conversation is likely to be accurately recalled by both partners.

*2.2.6.2. Relationship between memcon and later recall:* Our preregistration specified two types of similarity analysis: First, for the person who created the memcon, we examine if the IUs that were recalled in the written recall affect which IUs are recalled at the 1-week delay. Participants in the A role recalled 1872/7856 conversational IUs (24%) in their memcons. A week later, of the 1872 IUs in the memcon, 854 (46%) were also recalled orally by A. By comparison, of the 5984 IUs which were not in the memcon, only 533 (9%) appeared in A's oral recall. This asymmetry (46% vs. 9%) emphasizes the protective power of the memcon on A's memory.

Second, we examine if taking notes affects memory for what is recalled in the notes specifically, or memory for the conversation in general. One way to answer this question is to ask if, after a week delay, A was more likely than B to recall IUs that did not appear in the memcon. If writing the memcon was generally beneficial to memory, we would expect A to outperform B on these IUs. The answer to this question appears to be no, as IUs that did not appear in the memcon were recalled 9% (533/5984) of the time by A after a week delay, compared to B who recalled 11% of them (676/5984).

Inferentially, we can model the completeness of A and B's oral recall using an additional predictor variable of whether or not that IU had been recalled in the memcon (Table 5). This model revealed, in addition to previously described effects, a significant interaction between Recaller (A or B) and whether the IU appeared in the memcon ($b = 1.39$, $p<.0001$). As predicted, Participant A was significantly more likely to recall a given IU at the 1-week delay if that IU had appeared in the memcon ($b = 2.11$, $p<.0001$). IUs that were in the memcon were also more likely to be recalled by Participant B, though this effect was smaller ($b = 0.72$, $p<.0001$). The effect of Recaller (A or B) was not significant ($b = –.139$, $p = .50$), indicating that IUs *not* in the memcon were recalled at similar rates by Participants A and B 1 week later (note the effect of Recaller is the simple effect when Written Recall = 0 as this was treated as the reference level in the analysis).

Lastly, there was an unanticipated significant interaction between Talker (A or B) and whether an IU had been memorialized in the memcon ($b = –0.22$, $p = .028$), such that the effect of an IU being in the memcon on subsequent memory was smaller for Talker A's contributions ($b = 1.31$, $p<.0001$), compared to Talker B's contributions ($b = 1.53$, $p<.0001$). This effect may relate to item-specific features that made certain IUs produced by Talker B particularly memorable.

Table 5
Mixed-effects logistic regression analysis of whether conversational IUs were recalled at the 1-week delay oral recall, including 15,712 binary observations and 36 participants

| Fixed effects | Estimate | *SE* | *z*-value | *p*-value | |
|---|---|---|---|---|---|
| (Intercept) | –2.384 | 0.103 | –23.055 | <.0001 | |
| Written Recall | 1.416 | 0.084 | 16.798 | <.0001 | |
| Recaller | –0.139 | 0.207 | –0.670 | 0.503 | |
| IU order | –0.145 | 0.059 | –2.431 | 0.015 | |
| Talker | –0.095 | 0.064 | –1.491 | 0.136 | |
| IU order ^2 | 0.029 | 0.034 | 0.875 | 0.381 | |
| Written * Recaller | 1.391 | 0.150 | 9.305 | <.0001 | |
| Recaller * Talker | 0.675 | 0.100 | 6.757 | <.0001 | |
| Written * Talker | –0.222 | 0.101 | –2.194 | 0.028 | |
| **Random effects** | **Variance** | ***SD*** | **Corr.** | | |
| Participant (intercept) | 0.314 | 0.561 | | | |
| IU order | 0.106 | 0.326 | –0.12 | | |
| IU order ^2 | 0.027 | 0.165 | 0.06 | 0.59 | |
| Written Recall | 0.141 | 0.375 | –0.44 | 0.48 | –0.05 |

In sum, these similarity analyses indicate that writing a memcon strongly supports the writer's later memory for that material, primarily because the noted material is more likely to be recalled after a delay. We also observe some consistencies in which ideas tend to be recalled by both partners. Lastly, patterns in recall and failure to recall indicate a high degree of similarity in recall after a week, largely because most information in the conversation is not recalled by either participant. Notably, only 4.7% of the ideas expressed in the original conversation were recalled by both conversational participants after a week's delay.

## 3. General discussion

Writing memoranda of conversations, or "memcons," is a practice common in business, government, and myriad other settings. The ubiquity of this practice points to the need to better understand the impact of writing memoranda on later memory for conversation.

Our empirical findings show that participants who wrote detailed notes about what was said in conversation recalled significantly greater quantities of information from that conversation 1 week later, compared to participants who did not take notes. Accuracy—both in terms of the content of recall, and in terms of source attributions—was similar for participants who did and did not take notes, suggesting that note-taking enhances the quantity of information that can be accurately recalled after a delay, rather than increasing the quality of the recall. Composers of memcons did forget information between the two recall opportunities, but the enhancement to memory attributable to memcon composition led to a sizable advantage in

recall even in the face of forgetting. Although the memcons contained only 24% of the original content of the conversation, 46% of this produced material was reproduced on the delayed oral test. This material constituted 62% of the total IUs that the composer provided on that delayed test. Clearly, note-taking benefits memory and increases the completeness of reports of conversation at a delay.

Despite clear benefits of note-taking, when participants rated how well they remembered the conversation after a 1-week delay, participants who created a memcon provided lower estimates of their memory than participants who did not create a memcon. This result suggests an interesting metacognitive discrepancy between how much of the conversation participants *thought* they remembered and how much they actually remembered. It also suggests provisionally that self-assessments of the fidelity of one's memory may be disrupted by the generation of a memcon, and that those self-assessments should not be relied on in assessing the probative value of recall.

Another noteworthy outcome of this work concerns the amount of material that was co-remembered by conversational partners. After a week delay, the recalls of the two participants were similar, with participants both recalling or failing to recall, on average 78.7% of the conversational IUs. However, this similarity largely arose from common recall failures. In fact, only 4.7% of all IUs uttered in the original conversation were later recalled by both participants. This finding is surprising given that high fidelity for memory for the discourse history—that is, memory for what has been said in conversation—is posited to be critical to multiple well-established effects in the conversational literature, including effects of prior discourse on referential and syntactic form in language production (Reitter, Moore, & Keller, 2006; Yoon & Stine-Morrow, 2019), and memory for what has been said on subsequent language interpretation processes (Tooley & Traxler, 2010). Conversational partners tend to converge on shared conceptualizations and shared ways of referring to things that they refer to multiple times (Brennan & Clark, 1996; Clark & Wilkes-Gibbs, 1986; Horton & Gerrig, 2005). Partners also use collaboratively established referring expressions in conversation with the partner with whom they shared the experience of developing the term, but not with new partners, an effect attributed to a representation of common ground in the former case, and a lack of common ground in the latter (Wilkes-Gibbs & Clark, 1992). The fact that shared conversational recall is limited to less than 5% of the conversational details suggests that the effects of shared experience on referential form in conversation may be independent of the ability to recall the details of the conversation that gave rise to the similarity in form. Consistent with this interpretation are findings that persons with severe declarative memory impairment nonetheless develop shared labels with their conversational partner, for example, learning to call an abstract image "the bow tie" (Duff et al. 2006; Yoon, Duff, & Brown-Schmidt, 2017).

Although this study was not specifically designed to compare conversational recall after delays of different lengths, our findings are consistent with prior findings showing that conversational recall is likely to be more complete and accurate the shorter the delay between the conversation and the recall attempt (Stafford et al., 1987). In situations where the details of a conversation are in question, if written contemporaneous or near-contemporaneous notes are available, the notes are likely to be more complete and more accurate than a person's unaided recall of that conversation after a delay of a week or more.

### 3.1. Implications for theories of memory and conversation

The beneficial effect of note-taking on conversational memory is likely related to two effects that are well-established in the memory literature and typically studied with lists of unrelated materials, often words. The first is retrieval practice effects, which show that attempts to retrieve information from memory support delayed recall of that information (Roediger & Karpicke, 2006a, 2006b; Siler & Benjamin, 2020). The second is the production effect literature, which shows that producing information, such as naming a word aloud, promotes later memory for that information (MacLeod et al., 2010). The creation of contemporaneous notes involves both retrieval practice (recalling what was said from the recent conversation), and production (writing these ideas down). It is interesting to note that, despite the fact conversation is a major activity of daily living, and the writing of contemporaneous notes a common procedure in professional settings, no prior studies have examined the beneficial effects of retrieval practice and production for memory of conversation.

The finding that generating a memcon enhanced the amount of recalled information but not the accuracy of that information may reveal something about the highly structured and semantically coherent nature of conversation. The major benefits of retrieval practice lie in enhanced recall of the retrieved information, though secondary benefits appear to accrue to broad categorical knowledge of the material (Kang, McDaniel, & Pashler, 2011; Siler & Benjamin, 2019) and to memory for unrecalled but related material (Butler, 2010; Chan, 2010). In the case of unscripted experimental conversation, the structure of the content may not lend itself to these secondary effects. In contrast to conversation in service of a political or business goal, the material is only loosely related to one's ongoing goals and may not follow the kind of meaningful organizational or sequential structure that a conversation would if one party were trying to convince the other to engage in some important act. With materials that have an underlying semantic structure, but a weak one, each accurate recall of a component of the conversation may be as likely to lead to an incorrect as a correct inference about other components of the conversation. Indeed, after a week's delay, participants who generated a memcon recalled 1945 accurate and 533 inaccurate details (for an overall accuracy rate of 78%); participants who did not generate a memcon recalled 1336 accurate and 428 inaccurate details (an accuracy rate of 76%). As the number of accurate details increased, so did the number of inaccurate details. For example, one participant recalled the following:

> "We talked about…the weather which is better than where he's from which is Seattle"

Although the participant accurately recalled that they talked about the weather, the participant inaccurately recalled where his partner was from (it was not Seattle). With unstructured materials, such as word lists, inferences are rarely drawn and consequently rarely wrong (as evidenced for the typically low rates of intrusions in free recall; Anderson & Bower, 1972). And with highly structured material encoding with respect to underlying goals and motivations, we speculate that inferences are more likely to be accurate. Casual conversation of the sort emulated here likely exists between those two extremes.

Our paradigm was originally modeled after the types of situations that might arise in professional or business settings, in situations where a person wants to immediately memorialize a conversation after it is through. As such, the notes serve as an externalization of memory that is argued to be functionally equivalent to human memory mechanisms (Tollefsen, 2006), and that reflect a type of distributed cognition (Duff, Mutlu, Byom, & Turkstra, 2012; Hamilton & Benjamin, 2019). External memory aides offer utility in settings where the author of the memcon may wish to consult their own notes; indeed, the use of memory aids is a common strategy to manage memory impairment following brain injury (Evans, Wilson, Needham, & Brentnall, 2003). Viewed from this lens, externalization of one's experiences is not only functional in so far as it may support subsequent recall, but it also serves as an aid that can be subsequently consulted to refresh one's memory (e.g., before delivering testimony; when reminiscing about a vacation).

The online chronicling of daily life on social media is a common example of how people externalize their experiences, a process that is likely to shape our collective memory for events, only some of which are accurately chronicled (Stone & Wang, 2019). Collaborative recall introduces important biases to memory; for example, collaborative recall of word lists reduces the amount of information produced (compared to the expected output of two individual recallers) but increasesthe *similarity* of the partners' memories (Congleton & Rajaram, 2014). Thus, memory externalization in group settings, such as social media sharing, may serve to increase the similarity of individual recollections of shared experience.

### *3.2. Limitations and future directions*

A broader understanding of memory for conversation is critical for evaluating the probative value of memory reports in legal and political settings. The experiment reported here is the first to balance the benefits of experimental control with high levels of ecological validity needed to capture the variety and manner of genuine conversation. We did so by testing a convenience sample of individuals from a university community engaging in an unscripted conversation. The use of unscripted conversation affords exploration of the memory phenomena of interest in an interactive setting and offers advances over paradigms in which the to-be-remembered material is observed or scripted (cf. Fischer et al., 2015). A limitation, however, of the present approach is that participants engaged with relatively mundane topics (e.g., Nashville hot chicken), leaving open questions about how the findings might generalize to high-pressure situations or topics of great importance, such as the foreign policy discussions between Anatoly Dobrynin and Henry Kissinger described in the Introduction. Prior work suggests that more engaging, interactive, surprising, or inappropriate language is likely to be better remembered (Keenan, MacWhinney, & Mayhew, 1977; Kintsch & Bates, 1977). Given the positive effects of composing memoranda on conversational memory in the present study, it might be expected that such inappropriate remarks would have an increased chance at being memorialized in notes, and consequently be accessible to recall after longer delays.

Another limitation that merits discussion is that the written recall task was a surprise to participants. In real-world business and political settings, the author of a memcon often participates in the memorialized conversation with knowledge that he or she will prepare a memcon.

This knowledge might guide attention during the conversation, with effects on downstream memory, as expectation of recall tends to improve the likelihood of recall (Stafford & Daly, 1984). It is known that foreknowledge that a learner will need to teach to-be-learned material to others enhances memory (Nestojko, Bui, Kornell, & Bjork, 2014); similarly, the knowledge that material is likely to be of critical value and testimony may be called could exert a similar effect.

An additional consideration is that in real-world settings, the person who wrote a memcon may have the opportunity to use the memcon as a retrieval cue to assist recall of the conversation after some delay. For example, a witness in a legal proceeding may consult a memcon to refresh their recollection of a conversation before testifying about it (e.g., Niehoff, 2021). An unanswered question, then, concerns the memory quality of conversational recall after a delay of weeks or more, when one consults a memcon before recall. The present findings tentatively suggest that near-contemporaneous notes are likely to be more complete and accurate than attempts to recall a conversation after a significant delay. While it remains an open empirical question, we speculate that recalling a conversation with the use of a memcon as an external memory aid would produce a fuller accounting of that conversation. There is, however, some risk that repeatedly refreshing one's recollection by referring back to near-contemporaneous notes may distort the recaller's confidence in the underlying memory, and thus potentially skew perceptions of the recaller's credibility during a legal proceeding. An analogous concern has been studied in the context of eyewitness memory. Findings suggest that a witness's initial memory of seeing (or not seeing) a particular suspect can be contaminated by repeated lineup identifications during the investigative process, leading a witness who initially identified a suspect from a lineup with low confidence—or who initially failed to identify a suspect in a lineup at all—to become increasingly confident with each subsequent identification and ultimately testify that they saw the suspect with a high degree of confidence in the courtroom (Wixted, Wells, Loftus, & Garrett, 2021; Wells et al., 2020).

In the U.S. legal system, when a witness of one party to a lawsuit uses a written memcon to refresh their memory *while* testifying, the adverse party is generally entitled to have the writing produced, inspect it, cross-examine the witness about it, and introduce relevant portions into evidence (see Federal Rule of Evidence 612). If the witness uses the memcon to refresh their memory *before* testifying, however, courts retain discretion as to whether these options are available to the adverse party. A court is less likely to afford an adverse party these options if it does not believe the writing reviewed "ha[d] an impact upon" the testimony of the witness (see, e.g., Adidas Am. Inc. v. TRB Acquisitions, LLC, 2017). As a proxy for "impact," courts may focus on whether the witness reviewed the writing shortly before testifying. But earlier review or generation of the writing may also "impact" the witness's memory and testimony.[9] Indeed, while the time period between memcon generation and recall in our study (1 week) is much shorter than would be typical in a litigation setting, our findings suggest that creating a memcon may bolster subsequent recall even if the memcon is not reviewed. Future research might probe the extent to which reviewing a memcon at varying intervals before delivering mock testimony affects testimonial accuracy, testimonial completeness, and "witness" confidence. To further probe the risks described in the preceding paragraph, such research might also probe whether evaluations of the perceived credibility of testimony vary based on how recently the person testifying reviewed the relevant memcon. This research could also vary

whether the evaluator of credibility is told that the person testifying consulted a memcon before testifying (as factfinders in legal proceedings may or may not be aware that the witness used a memcon to refresh their memory before testifying).

In addition to memcons' influence on witness memory, it may be worthwhile for future research to investigate the value of memcons themselves as substantive evidence. Assume X is suing Y and that X's witness consults a memcon before testifying about the events described in the memcon. If X wants to admit the memcon as additional, documentary evidence that the events described therein are true, hearsay issues are raised. The hearsay rules generally prohibit the use of out-of-court statements (including memcons) to show that the assertions contained therein are true (Federal Rules of Evidence 801, 802). While there are numerous exclusions from and exceptions to the hearsay rules that can capture memcons in specific sets of circumstances (e.g., the prior consistent statement exclusion, the business record exception, the recorded recollection exception, and the present sense impression exception, among others, see Federal Rules of Evidence 801, 803), some memcons will not fall within any exception or exclusion. If a memcon does not fall within an exception or exclusion, X will generally not be able to put it into evidence to prove that the things that it says are true. Chief among the concerns justifying this approach is that juries will not be able to properly evaluate whether out-of-court statements—including memcons—are credible and reliable (see generally Tribe, 1974; Imwinkelried, 1989; Sevier, 2014). To this end, it would be interesting to empirically evaluate mock jurors' beliefs about the completeness and accuracy of memcons (see generally Sundby, 2022; Sevier, 2014; Jaeger, Levin, & Porter, 2017; Rachlinski, 2003). Do jurors' (mis-)understandings of memory lead them to overestimate the evidentiary value of memcons? Is the risk of overvaluation actually lower in circumstances in which hearsay exceptions apply?

## 4. Conclusion

The results of our experiment show that when a conversation is recalled in writing after a delay of 5 min, the act of taking written notes enhances later recall of the same conversation a week later. Conversational recall was significantly more complete for participants who engaged in written recall compared to those that did not. The groups were equally accurate—that is, the material they produced was just as likely to have been a genuine part of the prior conversation. The immediate written recall itself was more complete and more accurate than oral recall after a week. These findings suggest that the practice of writing memoranda of conversations is beneficial to later conversational memory, and that the memorandum itself is likely to be a better record of the conversation than what one can recall after a delay.

## Acknowledgments

This material is based on work supported by National Science Foundation Grant 15-56700 and 19–21492 to Sarah Brown-Schmidt. We thank Jordan Zimmerman, Kaitlin Lord, and many other research assistants in the Conversation Lab for their work in

collecting and coding these data. Thank you to Edith Beerdsen, Jeremy Counseller, David Faigman, Lisa Fazio, Ira Hyman, participants of the 16th Annual Conference on Empirical Legal Studies, and participants of the 63rd Annual Meeting of the Psychonomic Society for helpful comments and suggestions on this project and/or earlier drafts on this manuscript.

## Open practices statement

This study was preregistered at https://osf.io/s8vwx. The deidentified data and analysis scripts are available at https://osf.io/z8r9s/

## ENDNOTES

1 https://www.fordlibrarymuseum.gov/library/guides/findingaid/Memoranda_of_Conversations.asp#Ford
2 https://www.fordlibrarymuseum.gov/library/guides/findingaid/Memoranda_of_Conversations.asp#Nixon
3 https://clinton.presidentiallibraries.us/collections/show/255
4 Also see: https://nsarchive2.gwu.edu/NSAEBB/NSAEBB233/index.htm
5 For one Participant B, a second experimenter was not available so this person walked around the building on their own.
6 The preregistration additionally proposed analyzing the credibility of the oral recalls as credibility is relevant to impressions of witness testimony. Two coders used the Observed Witness Efficacy Scale (Cramer, DeCoster, Neal, & Brodsky, 2013) to rate all of the oral recalls. The average score across the 20 items in the OWES scale was similar for Participants A ($m$=8.53, $SD$=.31) and B ($m$=8.46, $SD$=.32); however, the correlation between the two coders' sets of ratings was low ($r$=.22), limiting the conclusions that can be drawn from this analysis. Development of a scale that is designed to evaluate credibility of ordinary recall (as opposed to testimony) may offer more utility for the materials used in the present study.
7 For participants who generated a memcon, recall completeness after 1 week was somewhat positively associated with interest ($r$=.327, $N$=20) and memory ratings ($r$=.539, $N$=20). For participants who did not generate a memcon, these associations were negligible for both interest ($r$= −.104, $N$=18) and memory ($r$=.082, $N$=19).
8 A supplemental model that excluded the nonpreregistered covariates and included only the effects of Recaller, Talker, and their interaction revealed a similar pattern of results, with an effect of Recaller ($b = .43$, $p < .001$) and a Recaller*Talker interaction ($b = .64$, $p < .0001$).
9 It is interesting to consider whether the act of *creating* a memcon, without subsequently reviewing it, would constitute "us[ing] a writing to refresh memory" before testifying under Federal Rule of Evidence 612.

## References


Adidas Am. Inc. v. TRB Acquisitions, LLC, 324 F.R.D. 389. (D. Or. 2017).

Aiken, E. G., Thomas, G. S., & Shennum, W. A. (1975). Memory for a lecture: Effects of notes, lecture rate, and informational density. *Journal of Educational Psychology*, *67*(3), 439.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, *79*(2), 97.

Bates, D., Mäachler, M., Bolker, B., Walker, S., (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benjamin, A. S., & Pashler, H. (2015). The value of standardized testing: A perspective from cognitive psychology. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 13–23.

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

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two minds, one dialog: Coordinating speaking and understanding. In B. H. Ross (Ed.),*Psychology of learning and motivation* (Vol. 53, pp. 301–344). Academic Press.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, *22*(6), 1482.

Burr, W. (1999). *The Kissinger transcripts: The top secret talks with Beijing and Moscow*. New York: New Press.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133.

Chan, J. C. (2010). Long-term effects of testing on the recall of nontested materials. *Memory*, *18*(1), 49–57.

Cho, S.-J., Brown-Schmidt, S., De Boeck, P., & Shen, J. (2020). Modeling intensive polytomous time series eye tracking data: A dynamic tree-based item response model. *Psychometrika*, *85*, 154–184.

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

Clark, H. H., & Marshall, C. R. (1978). Reference diaries (pp. 57.63). In D. L. Waltz (Ed.), Theoretical issues in natural language processing (Vol. 2). *New York*, NY: Association for Computing Machinery.

Coman, A., Momennejad, I., Drach, R. D., & Geana, A. (2016). Mnemonic convergence in social networks: The emergent properties of cognition at a collective level. *Proceedings of the National Academy of Sciences*, *113*(29), 8171–8176.

Congleton, A. R., & Rajaram, S. (2014). Collaboration changes both the content and the structure of memory: Building the architecture of shared representations. *Journal of Experimental Psychology: General*, *143*(4), 1570.

Cramer, R. J., DeCoster, J., Neal, T. M., & Brodsky, S. L. (2013). The Observed Witness Efficacy Scale: A measure of effective testimony skills. *Journal of Applied Social Psychology*, *43*(8), 1696–1703.

Davis, D., & Friedman, R. D. (2007). Memory for conversation: The orphan child of witness memory researchers. In *Handbook of eyewitness psychology*.

Duff, M. C., Mutlu, B., Byom, L., & Turkstra, L. S. (2012). Beyond utterances: Distributed cognition as a framework for studying discourse in adults with acquired brain injury. In *Seminars in speech and language33(1)*, 44–54

Duff, M. C., Hengst, J., Tranel, D., & Cohen, N. J. (2006). Development of shared information in communication despite hippocampal amnesia. *Nature neuroscience*, *9*(1), 140–146.

Einstein, G. O., Morris, J., & Smith, S. (1985). Note-taking, individual differences, and memory for lecture information. *Journal of Educational Psychology*, *77*(5), 522.

Erdelyi, M. H. (2010). The ups and downs of memory. *American Psychologist*, *65*(7), 623.

Evans, J. J., Wilson, B. A., Needham, P., & Brentnall, S. U. E. (2003). Who makes good use of memory aids? Results of a survey of people with acquired brain injury. *Journal of the International Neuropsychological Society*, *9*(6), 925–935.

Fischer, N. M., Schult, J. C., & Steffens, M. C. (2015). Source and destination memory in face-to-face interaction: A multinomial modeling approach. *Journal of Experimental Psychology: Applied*, *21*(2), 195.

Foley, M. A., Foley, H. J., Durley, J. R., & Maitner, A. T. (2006). Anticipating partners' responses: Examining item and source memory following interactive exchanges. *Memory & Cognition*, *34*(7), 1539–1547.

Gopie, N., & MacLeod, C. M. (2009). Destination memory: Stop me if I've told you this before. *Psychological Science*, *20*(12), 1492–1499.

GraphPad QuickCalcs Web site. (2022). Retrieved from https://www.graphpad.com/quickcalcs/kappa1/.

Hamilton, K. A., & Benjamin, A. S. (2019). The human-machine extended organism: New roles and responsibilities of human cognition in a digital ecology. *Journal of Applied Research in Memory and Cognition*, *8*(1), 40–45.

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

Imwinkelried, E. J. (1989). The importance of the memory factor in analyzing the reliability of hearsay testimony: A lesson slowly learnt–and quickly forgotten. *Florida Law Review*, *41*, 215–252.

Isaacs, E. A. (1990). Mutual memory for conversation (Doctoral dissertation, Stanford University).

Jaeger, C. B., Levin, D. T., & Porter, E. (2017). Justice is (change) blind: Applying research on visual metacognition in legal settings. *Psychology, Public Policy, and Law*, *23*(2), 259–279.

Kang, S. H., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, *18*(5), 998–1005.

Keenan, J. M., MacWhinney, B., & Mayhew, D. (1977). Pragmatics in memory: A study of natural conversation. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 549–560.

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

Knutsen, D., & Le Bigot, L. (2021). Estimating each other's memory biases in dialogue. *Discourse Processes*, *58*(2), 155–176.

Lamb, M. E., Orbach, Y., Sternberg, K. J., Hershkowitz, I., & Horowitz, D. (2000). Accuracy of investigators' verbatim notes of their forensic interviews with alleged child abuse victims. *Law and Human Behavior*, *24*(6), 699–708.

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 671–685.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McKinley, G. L., Brown-Schmidt, S., & Benjamin, A. S. (2017). Memory for conversation and the development of common ground. *Memory & Cognition*, *45*(8), 1281–1294.

Miller, J. B., deWinstanley, P., & Carey, P. (1996). Memory for conversation. *Memory*, *4*(6), 615–632.

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

Nestojko, J. F., Bui, D. C., Kornell, N., & Bjork, E. L. (2014). Expecting to teach enhances learning and organization of knowledge in free recall of text passages. *Memory & Cognition*, *42*(7), 1038–1048.

Niehoff, L. (2021). Recollections refreshed and recorded. *Litigation*, *47*(3), 1–4.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Rachlinski, J. J. (2003). Misunderstanding ability, misallocating responsibility. *Brooklyn Law Review*, *68*, 1055–1091.

Reitter, D., Moore, J., & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Roediger, H. L., III, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210.

Roediger, H. L., III, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.
Ross, M., & Sicoly, F. (1979). Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology*, *37*(3), 322.
Sachs, J. S. (1974). Memory in reading and listening to discourse. *Memory & Cognition*, *2*(1), 95–100.
Samp, J. A., & Humphreys, L. R. (2007). "I said what?" Partner familiarity, resistance, and the accuracy of conversational recall. *Communication Monographs*, *74*(4), 561–581.
Sevier, J. (2014). Testing tribe's triangle: Juries, hearsay, and psychological distance. *Georgetown Law Journal*, *103*, 879–931.
Siler, J., & Benjamin, A. S. (2020). Long-term inference and memory following retrieval practice. *Memory & Cognition*, *48*(4), 645–654.
Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592–604.
Stafford, L., Burggraf, C. S., & Sharkey, W. F. (1987). Conversational memory: The effects of time, recall, mode, and memory expectancies on remembrances of natural conversations. *Human Communication Research*, *14*(2), 203–229.
Stafford, L., & Daly, J. A. (1984). Conversational memory: The effects of recall mode and memory expectancies on remembrances of natural conversations. *Human Communication Research*, *10*(3), 379–402.
Stafford, L., Waldron, V. R., & Infield, L. L. (1989). Actor-observer differences in conversational memory. *Human Communication Research*, *15*(4), 590–611.
Stanley, S. E., & Benjamin, A. S. (2016). That's not what you said the first time: A theoretical account of the relationship between consistency and accuracy of recall. *Cognitive Research: Principles and Implications*, *1*(1), 1–11.
Stone, C. B., & Wang, Q. (2019). From conversations to digital communication: The mnemonic consequences of consuming and producing information via social media. *Topics in cognitive science*, *11*(4), 774–793.
Sundby, C. (2022). The neuroscience of evidentiary rules: The case of present sense impression. Unpublished manuscript. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3410089
Tollefsen, D. P. (2006). From extended mind to collective mind. *Cognitive Systems Research*, *7*(2–3), 140–150.
Tooley, K. M., & Traxler, M. J. (2010). Syntactic priming effects in comprehension: A critical review. *Language and Linguistics Compass*, *4*(10), 925–937.
Tribe, L. H. (1974). Triangulating hearsay. *Harvard Law Review*, 87(5), 957–974.
U.S. G.P.O. (2007). *Soviet–American relations: The detente years, 1969–1972*. Washington, DC: U.S. G.P.O
Voeten, C. C. (2020). Package "buildmer": Stepwise elimination and term reordering for mixed-effects regression. Retrieved from https://cran.r-project.org/package=buildmer
Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior*, *44*, 3–36.
Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag.
Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, *31*(2), 183–194.
Wixted, J. T., Wells, G. L., Loftus, E. F., & Garrett, B. L. (2021). Test a witness's memory of a suspect only once. *Psychological Science in the Public Interest*, *22*(1), 1S–18S.
Yoon, S. O., Duff, M. C., & Brown-Schmidt, S. (2017). Learning and using knowledge about what other people do and don''t know despite amnesia. *Cortex*, *94*, 164–175.
Yoon, S. O., & Stine-Morrow, E. A. (2019). Evidence of preserved audience design with aging in interactive conversation. *Psychology and aging*, *34*(4), 613.

Zormpa, E., Brehm, L. E., Hoedemaker, R. S., & Meyer, A. S. (2019). The production effect and the generation effect improve memory in picture naming. *Memory*, *27*(3), 340–352.

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

SUPPLEMENTARY MATERIAL

**Fig. S1**. The relationship between Conversational Idea Unit order and recall.

**Fig. S2**. Source attribution by Talker and Recaller