Behavioral Sciences and the Law
Behav. Sci. Law (2015)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/bsl.2184

Tell Me Everything You Discussed: Children's Memory for Dyadic Conversations after a 1-Week or a 3-Week Delay

Monica Lawson* and Kamala London†

In child abuse investigations, children are often asked to recount previous conversations related to the allegations (i.e., "conversational testimony"). To explore children's ability to provide conversational testimony, we staged a semi-structured novel dyadic conversation between an adult researcher and 8-year-old children ($n = 90$). Children's gist recall and recognition memory for their own statements, their conversational partner's statements, and question–answer pairs were tested after either a 1-week or a 3-week delay. The results revealed that children recounted a minority of the conversation, although children recalled more after a short delay (7%) than after a long delay (4%). A majority of children's free recall statements were accurate (68%); however, approximately one-third of their free recall statements were incorrect. Children almost exclusively recounted their own statements, and rarely recalled any of the adult's statements or the question–answer pairs during free recall. Reports of the adult's statements and question–answer pairs increased with cued recall questioning, but remained minimal. During recognition testing, children were able to distinguish between true and false recognition items for their own statements and the adult's statements, but performed at chance level on recognition items concerning question–answer pairs. Forensic implications of the results are discussed.

Children's recollection of conversations often plays a crucial role in both child abuse investigations (Lyon & Ahern, 2010; Malloy, Brubacher, & Lamb, 2013; Schaeffer, Leventhal, & Asnes, 2011) and criminal trials (Lyon & Stolzenberg, 2014; Stolzenberg & Lyon, 2014). Child abuse professionals often question children about whether they remember anything the alleged perpetrator said to them. Investigators are also interested in whether and to what extent children have talked to others about the allegations under investigation. Although children's ability to accurately recount conversations plays an important role in child abuse investigation, children's memory for conversations has gone ignored by researchers. In the present study, we examined children's memory for dyadic conversations (i.e., verbal interaction where the child actively converses with a partner) following a 1-week or 3-week delay.

During the investigative process, children frequently are asked to recount prior conversations that might have occurred during the commission of the suspected maltreatment (Stolzenberg & Lyon, 2014). For example, to determine whether genital

*Correspondence to: Monica Lawson, University of Toledo, 2801 W. Bancroft St. MS# 948, Toledo, OH 43606, USA.
E-mail: Monica.Lynn.Lawson@gmail.com
†University of Toledo, Toledo, OH.

touching involved routine caregiving versus sexual intent, investigators may ask children to recount any statements made by the alleged offender during questionable interactions. Children may also be asked if the alleged perpetrator made any threats or otherwise influenced them not to tell anyone about the maltreatment.

Investigators are encouraged to explore alternative hypotheses to abuse allegations (American Professional Society on the Abuse of Children (APSAC) Task Force on Investigative Interviews in Cases of Alleged Child Abuse, 2002; Anderson et al., 2010; State of Michigan Governor's Task Force on Children's Justice and Department of Human Services, 2004), including whether any person could have exerted tainting influence upon the child. The National Institute of Child Health and Human Development (NICHD) interviewing protocol, a developmentally appropriate and empirically supported protocol, includes an "information about the disclosure" section where children are asked to provide information regarding who they previously told about the allegations (Lamb et al., 2007). If the child mentions a conversation, interviewers are instructed to follow up by stating, "Tell me everything you talked about" (Lamb et al., 2007, p. 1230). In forensic practice, the majority of interviewers trained to follow the NICHD protocol include these questions when interviewing child witnesses (Malloy et al., 2013). Furthermore, legal professionals frequently ask about maltreatment-related conversations when children are on the witness stand (Stolzenberg & Lyon, 2014).

The reliability of children's conversational testimony relies on their memory for the conversation in question. A great deal of research has evaluated children's memory for events (for a review, see London & Kulkofsky, 2009). However, as previously noted, children's memory for conversation is largely neglected in the forensic developmental literature. The cognitive demands in remembering conversations cannot be assumed as equivalent to the cognitive processes involved in memory for experienced events (see Duke, Lee, & Pager, 2007; Lyon & Stolzenberg, 2014). Duke et al. (2007) posited that conversational memory might be more susceptible to erroneous encoding, rapid decay, and source monitoring errors than event memory due to the high frequency of daily conversations and the complexity of their nature (e.g., semantics, pragmatics). Researchers have explored children's memory for stimuli peripherally related to conversations, such as memory for written or verbal prose, including for individual sentences (Hedelin & Hjelmquist, 1998; Reyna & Kiernan, 1994) or more elaborate stories (Olson & Hildyard, 1981). However, little is known about children's memory for dyadic conversation (Davis & Friedman, 2007; Davis, Kemmelmeier, & Follette, 2005; Lyon & Stolzenberg, 2014).

To our knowledge, only one study has evaluated children's ability to recall dyadic conversations. Stolzenberg and Lyon (2015) evaluated 5- to 9-year-olds' reports for details communicated during a conversation that co-occurred with a play event. They also tested children's memory for an interview that occurred after the play session. Stolzenberg and Lyon (2015) concluded that children were able to recall details from the conversations, and the majority of recalled details were accurate after a 1-week delay. However, due to the initial inclusion of the play session along with the conversation, it is impossible to distinguish between children's memory for *what was said* during the conversations and children's memory for *what they did* during the play session. For example, if the children recounted talking about a broken toy, it is unclear if the children truly remembered talking about a broken toy or if the children remembered seeing the broken toy and, thus, concluded that they must have discussed the broken toy.

DOI: 10.1002/bsl

In the present study, we evaluated children's memory for a dyadic conversation independently from children's memory for co-occurring actions or physical objects. In this way, the current study approximates children's pure memory for dyadic conversation. The goal of the present study was to examine children's memory reports of a prior dyadic conversation after a 1- or 3-week delay. We were interested in the amount of the conversation children reported and the characteristics of their reports in terms of accuracy and utterance type. Children's recall and recognition memory were examined regarding three aspects of the conversation: child-generated utterances (i.e., statements made by the child during a target conversation), adult-generated utterances (i.e., statements made by the adult researcher during the target conversation), and question–answer pairs (i.e., two utterances from the target conversation; an eliciting utterance from the adult along with the child's response). Children's memory for conversation was assessed by free recall, cued recall, and recognition testing in order to evaluate children's conversational reports across various eliciting formats. Our hypotheses were largely guided by the adult memory for conversation literature and are discussed in the following paragraphs.

We expected that children would report a low percentage of the conversation when prompted to recall the conversation with free recall questioning (e.g., "Tell me what you talked about"). In terms of gist recall, adult participants in Stafford and Daly's (1984) research reported approximately 10% of an introductory conversation after a 5-minute delay. Bruck, Ceci, and Francoeur (1999) found similar results when examining mothers' memory for a conversation with their 4-year-old children that had occurred about three days earlier. In terms of verbatim-like recall, mothers reported approximately 5% of total utterances from the conversation.

Additionally, we expected that the quantity of children's reports would decrease across time (Ebbinghaus, 1885/1964). Stafford, Burggraf, and Sharkey (1987) re-evaluated participants' memory for conversations from Stafford and Daly's (1984) study a month after the conversation occurred and found that recall dropped from 10% to 4% of the conversation. We hypothesized that children in the 1-week delay condition would recall significantly more of their conversations than children in the 3-week delay condition, but all free recall reports were expected to be incomplete.

Although children's conversational reports were expected to be limited in extent, we hypothesized that their reports would be highly accurate. Adult's reports of previous conversations tend to be fairly accurate, although incomplete (Lamb et al., 2000). Additionally, so long as children have not been exposed to misinformation, children's event reports are largely accurate when elicited with open-ended questioning (for a review, see London & Kulkofsky, 2009). As such, we hypothesized that children's conversational reports would be mostly accurate regardless of delay condition.

We also expected that children's memory for child-generated and adult-generated utterances would be better than children's memory for question–answer pairs across free recall, cued recall, and recognition testing, regardless of delay. Children's memory for question–answer pairs was expected to be poor, as memory for question–answer pairs requires recounting two consecutive utterances and depends on verbatim-like recollection. Adults often fail to recount question–answer pairs and frequently misrepresent the formatting of questions that they do manage to recall (e.g., recount a suggestive question as being open-ended; Bruck et al., 1999; Lamb et al., 2000; Warren & Woodall, 1999). Specific hypotheses were not established regarding the directionality of children's reports for child-generated versus adult-generated utterances, as the

DOI: 10.1002/bsl

adult memory for conversation literature has inconsistent findings concerning the superiority of memory for self versus other utterances (Miller, deWinstanely, & Carey, 1996; Stafford & Daly, 1984).

METHOD

Participants

Because child–adult conversations change dramatically as a function of children's development (Bloom & Tinker, 2001; Bohannon & Bonvillian, 2001), we limited our study to 8-year-old children in order to bolster the similarity (e.g., length, linguistic complexity) of the target conversations across child participants. Ninety 8-year-old children ($M_{age} = 8.07$ years, SD = 0.31) were recruited from a mid-sized Midwestern city. The sample consisted of 43 females and was largely Caucasian ($n = 81$), followed by African American ($n = 7$), and other ($n = 2$). Parental consent was obtained for all children prior to participation.

Design

All children participated in two separate sessions. In the first session, children engaged in a target conversation. During the second session, children's memory for the target conversation was assessed. Children were randomly assigned to a 1-week ($M = 7$ days, SD = 0.83) or a 3-week ($M = 23$ days, SD = 2.60) delay between the conversation and the memory test.

Materials

A digital voice recorder was used to audio record both the target conversation and the memory test. All recordings were then transcribed using Express Scribe software, and a second transcriber verified the accuracy of the transcriptions. During Session 1, participants were presented with a fabric box that contained a stuffed monkey and a baseball. These objects served as props for the target conversation. Additionally, the researcher wore a tall yellow stovepipe hat during Session 1 to increase the salience of the conversational event.

Procedure

Session 1: Target Conversation

Children participated in a 5-minute ($M = 5.15$ minutes, SD = 1.50) conversation with an adult researcher. During the conversation, the child and the researcher worked together to verbally compose a story about a monkey and a baseball. The researcher presented herself to the child as a storyteller who likes to create stories with children. She put on her yellow hat and explained the rules of the storytelling game: the child will select two objects from the storytelling box, and she will ask the child questions to help them create a story about the two selected items. After this explanation, all children listened to a 2-minute pre-recorded story regarding a race car driving dinosaur. The purpose of this story was to provide the children with an example of how to play the

DOI: 10.1002/bsl

storytelling game. Upon completion of the pre-recorded story, children were instructed to select two items from the storytelling box. Every child selected a monkey and a baseball, which were the only objects inside of the container. The adult researcher started the conversation by asking the child, "What is your monkey's name?" and the conversation ensued.

The goal of the storytelling game was to encourage a unique dyadic conversation between the child and the researcher while maximizing experimental control across conversations. To guide the flow of the conversations, researchers were highly trained to deliver 25 scripted questions about the monkey and/or the baseball during the target conversations. Invitation prompts such as "Tell me more about that" were used to encourage children to provide additional input for initially limited responses. Children's responses during the conversation were unique to each individual child and varied across conversations. Researchers were instructed to keep the conversation as natural as possible while adhering to the scripted questions. The only physical components of the storytelling game were the storytellers' yellow hat, the storytelling box, the monkey stuffed animal, and the baseball. Because the conversational event had limited co-occurring physical objects and actions, it can be assumed that children's later memory reports for the target conversation were primarily based on their ability to recount *what was said* during the conversation rather than what they did or saw.

Session 2: Memory Test

Free and cued recall. After the specified delay time (either 1-week or 3-weeks), an unfamiliar researcher (one of three different trained research assistants, all blind to study hypotheses) conducted a memory test with each child. There were three sections to the memory test: free recall, cued recall, and recognition testing. First, children were asked to recall their story (i.e., the target conversation), "Exactly the same way it happened from the very beginning to the end." Invitations such as "Tell me more about that" were used to encourage an exhaustive report from the child. During the cued recall section, children were asked if the adult researcher asked questions during the target conversation. Children were prompted to provide their original answer for every question they recalled from the target conversation. Additional cued recall questions were not asked if children reported the adult researcher did not ask them questions.

Recognition questions. Next, children were asked 30 forced-choice (i.e., yes/no) questions concerning the content of the target conversation. The recognition test consisted of three sections: child-generated utterances, adult-generated utterances, and question–answer pairs. Child-generated recognition questions were derived from statements the child made during the target conversation. Adult-generated recognition questions were derived from statements made by the adult during the target conversation. Question–answer pairs contained two consecutive statements from the target conversation: an eliciting question from the adult and the child's answer. Each recognition test section consisted of 10 questions. For each of the three sections, half the statements were made during the conversation (i.e., true questions), and half the statements were not made during the conversation (i.e., false questions). Children were instructed to either confirm or deny that the designated declarant said the utterance or question–answer pair during the target conversation.

True and false adult-generated question items and false child-generated question items were standard across all recognition tests. The content of true and false question–answer pair items and true child-generated question items varied according

DOI: 10.1002/bsl

to the child's original reply in the target conversation. However, the selection of statements used for the true and false question–answer pair and true child-generated items were standard across all tests. Children's responses to a predetermined set of eliciting questions during the target conversation were used to create these recognition test items. For example, the child's response to, "Is your monkey a girl or a boy?" during the target conversation was used to create the recognition item, "Did you tell the storyteller that 'your monkey was a [boy/girl]?'"

The five *true question–answer pair* items comprised a researcher's question from the target conversation and the child's original answer. For the five *false question–answer pair* recognition items, questions that were directed by the researcher during the conversation were presented as open-ended, and statements that were open-ended in the conversation were conveyed as directed in the recognition test. For example, the researcher's question, "How many brothers and sisters does your monkey have?" during the target conversation was incorrectly represented during the false question–answer recognition item as, "Did your monkey have one brother and two sisters?"

All recognition tests were counterbalanced so that each section (i.e., child-generated utterances, adult-generated utterances, and question–answer pairs) had an equal opportunity of appearing first during the recognition test. True and false recognition items within each section were randomly ordered via a random number generator. The ordering of items within each section was standard across all tests.

Data Coding

Target Conversation (Session 1) Coding

All transcripts were coded in a series of steps. First, false starts (e.g., um, ah) were eliminated from all transcripts. Then, transcripts were parceled into utterances. An utterance was operationally defined as a verb bounded by a pause (Bruck et al., 1999). For example, a statement such as, "The monkey ran and he jumped" represents two utterances [i.e., the monkey ran (1) and he jumped (2)]. However, occasionally, an utterance was composed of one word (e.g., "Yes") (see Bruck et al., 1999). Utterances were classified as either child-generated (i.e., said by the child) or adult-generated (i.e., said by the adult). Question–answer pairs (i.e., an eliciting utterance from the adult and the child's response) were also identified. The number of total utterances, child-generated utterances, adult-generated utterances, and questions-answer pairs stated during the target conversations were tallied.

Memory Test (Session 2) Coding

Transcripts from the memory test were coded in a series of steps. Similarly to coding of the target conversations, false starts (e.g., um, ah) were eliminated from transcripts. Additionally, questions posed by children for clarification regarding various memory test tasks were disregarded. Second, transcripts were divided into utterances. Utterances produced during free and cued recall were operationally defined as recalled utterances. All recalled utterances were classified as either a child-generated recalled utterance (i.e., a statement the child said during the target conversation) or adult-generated recalled utterance (i.e., a statement the adult said during the target conversation). Recalled question–answer pairs were identified when the child recalled an

DOI: 10.1002/bsl

eliciting utterance stated by the adult and the child's original response from the target conversation. Third, recalled topics were classified as a special category of recalled utterances. Recalled topics were operationally defined as a recalled utterance that reflected a semantic summary of multiple utterances stated during the target conversation. For example, the broad recall, "We talked about a monkey and a baseball," was coded as a recalled topic as this utterance summarizes what was generally discussed during the target conversation rather than recalling the content of an individual utterance that occurred in the conversation. Recalled topics were their own category and were not combined with recalled utterances in the following analyses. The number of topics within each conversation was limitless, as there are multiple variations in how several utterances could be summarized.

In some instances, children repeatedly reported the same information during memory testing. Repeated recalled utterances were disregarded in the analyses. The number of total recalled topics, recalled utterances, child-generated recalled utterances, adult-generated recalled utterances, and recalled questions–answer pairs were tallied.

Data Scoring

Free and Cued Recall Scoring

Recalled topics. Recalled topics were judged for accuracy by assessing overall topics within the participants' target conversation. Accurately recalled topics correctly summarized multiple utterances from the target conversation (e.g., "We talked about a monkey."). Inaccurate recalled topics misrepresented what was discussed during the target conversation (e.g., "We talked about the moon."). If any part of an utterance was inaccurate, then the utterance was categorized as inaccurate. Accurately and inaccurately recalled topics were tallied. Recalled topics were analyzed separately from recalled utterances and recalled question–answer pairs.

Recalled utterances. Recalled utterances were compared with utterances from the target conversation to determine accuracy. Accurately recalled utterances contained at least the gist of an utterance from the target conversation. Gist recall was operationally defined as a semantic recall of an individual utterance from the target conversation (adapted from Brainerd & Reyna, 2004). For example, "My monkey had a lot of siblings" was coded as an accurate recollection of the original utterance, "My monkey has three sisters and two brothers."

Recalled utterances that did not match the gist of an individual utterance from the target conversation were scored as inaccurate. Inaccurately recalled utterances were classified into one of four mutually exclusive categories: elaborations, inferences, factual errors, and speaker switches. Elaboration errors occurred when the content of an incorrect recalled utterance added additional information above and beyond what was said during the target conversation. Inference errors occurred when the content of the recalled utterance was inferred from an utterance in the target conversation, but was never actually said. Factual errors were inaccurate recalled utterances that retained a kernel of truth from an utterance provided during the target conversation, but the details were incorrect. Speaker switches occurred when the children incorrectly claimed one of the utterances stated by the adult as their own and vice-versa. Examples of errors are provided in Table 1.

DOI: 10.1002/bsl

Table 1. Examples of inaccurate recalled utterance classification

Classification	Utterance in the target conversation	Recalled utterance
Elaboration error	Recalled utterance does not match any utterance from target conversation.	The monkey likes music.
Inference error	He practices every day after school.	My monkey loves baseball.
Factual error	My monkey's name is Jim.	My monkey's name was Bob.
Speaker switch	How old is your monkey? [Stated by the researcher]	I asked how old the monkey was. [Child claims credit for researcher's utterance]

Recalled question–answer pairs. An accurately recalled question–answer pair required that the child accurately reported two utterances: the adult's question and the child's answer. Any inclusion of an inaccurately recalled utterance deemed the entire recalled question–answer pair as inaccurate.

Recognition Test Scoring

Accurate and inaccurate responses for true and false questions for each section of the recognition test were tallied.

Inter-rater Reliability

Two researchers, blind to the study hypotheses and the experimental delay conditions, coded and classified utterances from the target conversation transcripts. Another pair of researchers, also blind to hypotheses and conditions, coded and scored all memory test transcripts. Each member of the pair coded all transcripts. The interclass correlation (ICC) was selected as the inter-rater reliability measure because transcripts were coded twice and the data are continuous (Hallgren, 2012). The resulting ICCs were in the excellent range, ICC $\geq$ 0.99 (Cicchetti, 1994), reflecting a high degree of coder agreement.

RESULTS

Preliminary analyses were conducted to examine the utterance composition of the target conversations. The analyses revealed that, on average, the child and the adult contributed equally to the target conversations in number of utterances produced. The utterance composition of the target conversations did not significantly vary between delay conditions (p-values > 0.05). The storytelling game was successful in producing dyadic conversations where the child and the adult were verbally productive (Table 2).

Table 2. Means and standard deviations of utterances in the target conversations by utterance type and delay condition

	1-week delay ($n = 49$)	3-week delay ($n = 41$)	Total ($n = 90$)
Child-generated	54.31 (26.60)	47.49 (16.19)	51.20 (22.61)
Adult-generated	51.55 (13.50)	51.51 (11.84)	51.53 (12.70)
Total utterances	105.86 (32.66)	98.76 (22.42)	102.62 (28.53)

Note. All differences between the 1-week and 3-week delay conditions were non-significant (p-values > 0.05).

DOI: 10.1002/bsl

The results concerning children's reports of the conversation during free recall, cued recall, and recognition testing are provided in the following sections. First, we report children's recall regarding the topic of the conversation. Second, children's free recall reports are examined in raw numbers. Third, to gauge the completeness of children's free recall, their reports were reanalyzed as a function of the total number of utterances from the target conversation. Then, children's reports during cued recall are presented. Finally, children's performance on the recognition test is reported.

Free Recall Reports

Recalled Topics

Virtually all participants ($n = 87$ of 90) remembered, without specific prompting, that the target conversation occurred (i.e., in response to an open-ended question of "Did you talk to my friend, the storyteller?"). The majority of children ($n = 74$) spontaneously reported at least one topic of the conversation. All recalled topics were accurate. However, the production of recalled topics was limited and did not significantly differ between the 1-week ($M = 0.92$ recalled topics, $SD = 0.61$) and the 3-week ($M = 1.17$ recalled topics, $SD = 0.80$) delay conditions [$t(88) = -1.70$, $p = 0.09$]. On average, children accurately recounted 1.26 ($SD = 0.57$) topics of the conversation during free recall questioning. As noted in the methods section, each conversation had a limitless amount of topics as there are various ways in which utterances could be summarized. Thus, proportion scores regarding the percentage of topics that were recalled are not presented. The maximum number of recalled topics was four.

Raw Data

The following analyses of recalled utterances were computed with logarithm-transformed data following guidelines provided by Field (2013), Howell (2013), and Osborne (2002), as raw data violated the parametric assumption of normality. Geometric means were computed by taking the antilog of the logarithmic mean value and are used to interpret the results within this report. Effect sizes were computed with transformed means and standard deviations.

Two participants in the 3-week delay condition failed to recall any utterances from the target conversation during memory testing. All other participants ($n = 88$) produced at least one recalled utterance during the free recall task. Regardless of accuracy, participants in the 1-week delay condition recounted significantly more utterances in their free recall reports compared with participants in the 3-week delay condition [$t(88) = 2.85$, $p < 0.01$, $d = 0.60$]. The vast majority of recalled utterances (98%) represented an utterance made by the child during the target conversation. Children in the 1-week delay condition recalled significantly more of their own original utterances than children in the 3-week delay condition [$t(88) = 3.06$, $p = 0.003$, $d = 0.64$], regardless of accuracy. Children in the short delay condition reported significantly more accurately recalled utterances [$t(71.87) = 3.45$, $p = 0.001$, $d = 0.74$] and accurately recalled child-generated utterances

DOI: 10.1002/bsl

[$t(69.92) = 3.67$, $p < 0.001$, $d = 0.79$] than children in the long delay condition (Table 3).

Adult-generated recalled utterances were extremely limited. Only six participants recalled any adult-generated utterances during free recall. Further, only five participants spontaneously recalled question–answer pairs when prompted with open-ended questioning.

Proportionate to the Target Conversation

The above analyses demonstrate that most children reported multiple utterances when asked to recount the conversation. However, to appreciate children's memory for the entire conversation, children's reports were contextualized by the total number of utterances said during the target conversation. Proportion scores were computed for each participant by dividing the number of accurately recalled utterances by the total number of utterances within the child's target conversation. These percentages were utilized to examine the completeness of children's free recall reports. Means and standard deviations are reported in Table 3.

Children in the 1-week delay condition accurately recalled a significantly higher proportion of the overall conversation compared to children in the 3-week delay condition [$t(64.78) = 2.75$, $p < 0.01$, $d = 0.59$]. Children in the 1-week delay condition accurately recalled significantly more of their own utterances stated during the target conversation than children in the 3-week delay condition [$t(59.87) = 3.14$, $p < 0.01$, $d = 0.68$]. As so few participants spontaneously recalled adult-generated utterances and question–answer pairs, further statistical analyses were not conducted (Table 3).

Accuracy of Reports

The overall accuracy of free recall reports was examined. Accuracy proportion scores were created by dividing the number of accurately recalled utterances by the total

Table 3. Means and standard deviations of recalled utterances in free recall reports by utterance type and delay condition

	1-week delay ($n = 49$)	3-week delay ($n = 41$)
Child-generated utterances		
Accurately recalled [a]⋆⋆	6.86 (4.80)	3.81 (3.77)
Total recalled [a] ⋆	10.37 (8.40)	6.16 (6.44)
% of accurate recall to original [a] ⋆	13.92 (8.60)	8.01 (8.36)
Adult-generated utterances		
Accurately recalled	0.08 (0.40)	0.24 (1.09)
Total recalled	0.12 (0.44)	0.24 (1.09)
Total utterances		
Accurately recalled [a]⋆⋆	6.96 (4.77)	4.13 (3.70)
Total recalled [a] ⋆	10.48 (8.43)	6.60 (6.28)
% of accurate recall to original[a] ⋆	6.85 (4.27)	4.46 (4.44)

Note. [a] indicates that the analysis was conducted with logarithmic-transformed data and geometric means are presented. ⋆ indicates statistically significant differences between delay conditions at $p \leq 0.01$, ⋆⋆ represents $p \leq 0.001$. Percentage of accurate recall to original means were computed by averaging the number of accurately recalled utterances divided by the number of utterances said within the target conversation for each individual. Percentage of accurate recall to original is not reported for adult-generated utterances, as free recall of this type was extensively limited.

DOI: 10.1002/bsl

number of recalled utterances. The assumption of normality was not violated in this analysis, and non-transformed percentages are reported. Participants who did not recall any utterances from the conversation were excluded from this analysis. Reports made 1-week after the target conversation were more accurate ($M = 70.73\%$, $SD = 22.74\%$) than reports collected 3-weeks after the conversation ($M = 65.43\%$, $SD = 22.30\%$). However, this difference was not significant [$t(86) = 1.09$, $p = 0.28$]. Free recall reports were 68.38% (SD = 22.60%) accurate, on average. Additionally, the delay condition did not significantly influence the accuracy of recalled child-generated utterances ($M = 70.45\%$, $SD = 23.00\%$ for the 1-week delay; $M = 64.16\%$, $SD = 32.63\%$ for the 3-week delay) [$t(85) = 1.30$, $p = 0.19$]. On average, children were 67.70% (SD = 22.51%) accurate in their reports of child-generated utterances in free recall.

Seventeen participants did not make any errors in their free recall reports. Of children who made at least one error ($n = 73$), children were most inclined to make elaboration errors (67% of errors), followed by factual errors (24%), inference errors (9%), and speaker-switches (<1%). The type of errors did not differ significantly by delay condition (p-values > 0.05).

Cued Recall Reports

One participant in the 3-week delay condition was not asked the cued recall question during memory testing and was excluded from the following analyses. The majority of participants ($n = 77$) correctly reported that the adult researcher asked questions during the target conversation when specifically asked about the adult's involvement (i.e., "Did the storyteller ask you questions to help you create your story?"). However, reports of adult-generated utterances remained limited. Approximately a third of the sample ($n = 28$) did not recall any adult-generated utterances during cued questioning. Of the children who reported at least one adult-generated utterance ($n = 61$), 2.80 (SD = 1.61) adult-generated utterances were recalled, on average, with 64.20% (SD = 34.84%) accuracy. The amount of adult-generated utterances recalled during cued recall and accuracy of reports did not vary significantly by delay condition [$t(59) = 0.01$, $p = 0.99$; $t(59) = -0.57$, $p = 0.57$, respectively].

Additionally, recalled question–answer pairs were limited in cued recall reports from children in both delay conditions. Over a third of the sample ($n = 32$) did not report any question–answer pairs from the target conversation when specifically prompted during cued recall. Of participants who recalled at least one question–answer pair ($n = 57$), children reported 2.51 (SD = 1.48) question–answer pairs and were 57.61% (SD = 39.03%) accurate, on average. The quantity and quality of reports regarding question–answer pairs did not significantly differ between delay condition [$t(54.90) = 0.44$, $p = 0.66$; $t(55) = -0.40$, $p = 0.69$, respectively].

Recognition Test Performance

Recognition tests were constructed to include 30 questions. In some instances, adult-generated statements used to create the individualized recognition test items were not said during the target conversation due to researcher error. When this error occurred, the recognition item was not included in the recognition test. Fifteen children had a recognition test composed of 29 items, one child was asked 28 questions, and one child was asked 26 questions.

DOI: 10.1002/bsl

The following analyses were conducted on raw data as parametric assumptions were not violated. Children in the 1-week delay condition ($M = 74.28\%$ correct, $SD = 6.02\%$) significantly outperformed children in the 3-week delay condition ($M = 68.08\%$ correct, $SD = 7.50\%$) on overall recognition test scores [$t(88) = 4.35$, $p < 0.001$, $d = 0.91$]. To further explore recognition test performance by question type and delay condition, a 2 (delay condition: 1-week, 3-week) × 3 (question type: child-generated, adult-generated, question–answer pairs) mixed-factorial ANOVA with repeated measures on the last factor was conducted. There was a significant main effect for question type [$F(2, 176) = 237.38$, $p < 0.001$, $\eta_p^2 = 0.73$] and delay condition [$F(1, 88) = 18.94$, $p < 0.001$, $\eta_p^2 = 0.18$]. These main effects were qualified by a significant interaction [$F(2, 176) = 4.47$, $p = 0.01$, $\eta_p^2 = 0.05$]. Bonferroni adjusted t-tests showed that children in the 1-week delay condition performed significantly better on the child-generated recognition items [$t(69.93) = 3.13$, $p < 0.01$, $d = 0.67$] and the adult-generated recognition items [$t(88) = 4.11$, $p < 0.001$, $d = 0.86$] compared with children in the 3-week delay condition. However, performance on the question–answer pair recognition items did not significantly differ between delay conditions [$t(88) = 0.25$, $p = 0.80$]. Percentages are reported in Table 4 for ease of interpretation.

DISCUSSION

A growing body of research demonstrates that child witnesses often are asked to recall abuse-related conversations during forensic proceedings (Lyon & Stolzenberg, 2014; Malloy et al., 2013; Stolzenberg & Lyon, 2014). Children's ability to provide accurate conversational testimony hinges on their ability to remember the conversation in question. The current study is among the first to systematically evaluate children's memory for dyadic conversation.

This research yielded a number of novel findings. First, all children recalled the conversation took place, and most were able to recall a general topic of conversation. This finding was not surprising as the conversational event in this study was a special activity for children, and the topic of the conversation was somewhat novel and unique. These results suggest that children are able to recall the occurrence of unique and novel conversations and can accurately report the topic of discussion up to 3-weeks following the conversation.

All but two children in the current study were able to report at least one utterance from the target conversation during free recall. Most children provided several utterances from the conversation in their free recall reports. The vast majority of children's recall pertained to statements they generated. Children rarely included adult-generated utterances or question–answer pairs in their free recall reports of the conversation. This finding is concerning from a forensic standpoint as investigators often want to know both what the child said and how the child's statement was elicited in allegation-related

Table 4. Means and standard deviations of recognition test performance by question type and delay condition in percentages

Question type	1-week delay ($n = 49$)	3-week delay ($n = 41$)	Total ($n = 90$)
Child-generated	84.30 (10.42)	75.70 (14.83)	80.38 (1.33)
Adult-generated	87.76 (10.06)	78.30 (11.81)	83.44 (11.82)
Question–answer	50.78 (10.40)	50.24 (9.61)	50.54 (10.00)

DOI: 10.1002/bsl

conversations (Bruck et al., 1999). The current findings suggest that children have a strong tendency to provide their own statements in free recall reports of conversations and the structure of the conversation (i.e., question–answer pairs) is lost, even after a delay as short as 1-week.

Although children were able to recall some specific content from the conversation during free recall, their reports were not completely accurate. In the current study, reports were judged by gist accuracy, and verbatim recollection was not necessary. Regardless of delay condition, free recall reports were approximately 68% accurate. Our hypothesis that children would produce highly accurate free recall reports was partially supported. Children were mostly accurate in their reports, but approximately one in every three recalled utterances was inaccurate. This result was surprising considering the large body of research demonstrating that children's event reports are highly accurate when prompted with open-ended questions (similar to the free recall prompt used in the current study) (for review see London & Kulkofsky, 2009).

There are two potential explanations for this finding. One possibility is that the accuracy of children's conversational reports was an artifact of the conversation in this study. The child and the adult researcher verbally created a fictitious story about a monkey and a baseball during a single dyadic interaction. While recalling the conversation, children might have reported inaccurate information in efforts to tell a good story rather than to recall the target conversation (see Kulkofsky & Klemfuss, 2008). This explanation might also account for why most errors made by children were classified as elaborations errors, where children recalled utterances that went above and beyond the content of the target conversation. However, the unique conversational content was intentionally devised in efforts to test children's memory for a novel conversation. We were interested in testing memory for the conversation while attempting to minimize the role of scripted information the child might interpolate into their recollection. A second possibility to account for the sizeable proportion of inaccurate statements observed in our study is that children's free recall reports of previous conversations are truly less accurate than children's reports for activities that occurred during a physical event (Duke et al., 2007). Future research is needed to explore this possibility as the current study did not compare children's memory for conversation versus children's memory for an event.

Children's free recall reports of the conversation were far from complete. Children in the 1-week delay condition accurately recalled significantly more of the conversation (7%) than children in the 3-week delay condition (4%). This finding was expected, as memory typically experiences a sharp decline before stabilizing across time (Ebbinghaus, 1885/1964). However, conversational recall was low for all children regardless of delay, even when judged for gist accuracy (versus verbatim). This result suggests that, similar to adult's conversational memory, children's ability to recall conversations in which they have engaged will be a limited reflection of the actual conversation.

Approximately one-third of children in the current study did not report any adult-generated and/or question–answer pairs, even when specifically prompted with cued recall questioning. Children who did recall adult-generated and questions–answer pairs accurately reported approximately two of each in cued recall. Although the number of reported adult-generated and question–answer pairs increased with cued recall questions (compared with free recall), reports were still limited. During the target conversation, the adult's role was to ask the children questions about the monkey and the baseball rather than to provide information. Children were told to, "Tell me your story

DOI: 10.1002/bsl

exactly the same way it happened from the very beginning to the end" as the free recall prompt. The phrasing of this question might explain why children showed such a strong tendency to report only their own utterances from the conversation rather than the adult's utterances or question–answer pairs in free recall reports. Children's increased ability to report adult-generated utterances during cued recall suggests that children retained at least some of this information from the conversation. However, children still recalled little of the adult-generated utterances and question–answer pairs even when specifically asked to provide this information in cued recall. These findings suggest that children's free recall and cued recall reports of their conversational partner's utterances are quite limited.

Although children recalled a small minority of the target conversation during free recall testing, their performance on the recognition memory test indicates that they did remember some of the conversation's content. Children performed moderately well in both the child-generated and adult-generated sections of the recognition test. Although children in the 1-week delay condition outperformed children who experienced a 3-week delay on these sections, children in both conditions performed fairly well in identifying true from false items concerning both their own and their conversational partner's statements. However, children performed poorly in recognizing the specific types of questions that elicited information from them during the conversation. Similar to adults' memory for conversation (Bruck et al., 1999; Lamb et al., 2000), our results indicate that specific content and details about conversational turn-taking fades quickly from children's memory.

Limitations and Future Directions

The current study was among the first to examine children's memory for a dyadic conversation, and there were limitations to it. First, the current study did not examine developmental trends in children's conversational memory. Developmental differences in children's reports of prior conversations are likely to occur given children's developing understanding of conversations and cognitive capacities (see Bloom & Tinker, 2001; Bohannon & Bonvillian, 2001).

Second, while beyond the scope of the current study, evaluating children's memory for dyadic conversations after an extended delay is forensically relevant. Children often delay telling others about abuse (London, Bruck, Ceci, & Shuman, 2005), and forensic investigations often take a long time. In the current study, children's memory for a single target conversation was assessed either 1-week or 3-weeks after the target conversation occurred. Future studies need to systematically evaluate children's memory for dyadic conversations after an extended delay.

Third, the finding that children had a strong preference in reporting their own utterances from the target conversation as opposed to their conversational partner's utterances should be generalized within the confines of the dyadic conversation that occurred. In the current study, children participated in a semi-structured dyadic conversation where their conversational partner was provided scripted lines to guide the flow of the conversation. The use of a semi-structured conversation was a strength of the research design, as it allowed for more control and internal validity across conversations. The storyteller's role in the conversation was to ask questions rather than to contribute substantive information to the conversation. However, one possibility is that

DOI: 10.1002/bsl

children did not recall the storyteller's utterances (and the question–answer pairs) because the researcher did not provide direct information about the conversational topic.

Future studies should examine children's memory for dyadic conversations within varying contexts. The target conversation in this study did not contain contextual components that may occur in abuse investigations (e.g., suggestibility, inclusion of sexual subject matter). Children's memory for conversations could vary according to the content of the conversation, the context in which the conversation occurred, and the context in which children are asked to recall the conversation.

Additionally, future studies need to evaluate children's memory for conversations with various conversational partners. In the present study, the storyteller and the child were strangers. However, child witnesses are often asked to recount their disclosure conversations as well as conversations between themselves and the perpetrator (Stolzenberg & Lyon, 2014). Child sexual abuse victims usually know the perpetrator (U.S. Department of Health and Human Services, 2013), and they typically disclose abuse to their mother, peers, and/or other family members (London et al., 2005; Malloy et al., 2013). Thus, evaluating children's memory for conversations that occur with a familiar conversational partner is of forensic importance.

Another future area of direction is exploring children's source monitoring abilities as a function of their memory for conversations. In source monitoring tasks, children are asked to identify the source of their information as something that actually occurred during an event or something they only heard (Johnson, Hashtroudi, & Lindsay, 1993). Children often perform poorly on source monitoring tasks and have the tendency to say that non-experienced but heard details of an event actually happened (Principe & Schindewolf, 2012). Source monitoring tasks inherently assume that children remember the target conversation and are able to utilize it as a source of their knowledge. The current study suggests that children are able to remember the general topic of a conversation, but have difficulty reporting what was actually said during the conversation. Future research is warranted to explore the relationship between source monitoring performance and children's memory for the conversation.

Children's ability to recall conversations has long been overlooked in the developmental literature. The current study was among the first study to systematically evaluate children's memory for a dyadic conversation and, specifically, children's memory for a dyadic conversation after delay. Generally, children remembered that the conversation occurred and accurately reported its theme, but recalling what was said during the conversation was difficult even after a short delay of 1-week. Further evaluation of children's memory for conversation is a fruitful area for future research.

ACKNOWLEDGEMENTS

This research was support in part by an American Psychology-Law Society Student Grant-in-Aid to M.L.

REFERENCES

Anderson, J., Ellefson, J., Lashley, J., Miller, A. L., Olinger, S., Russell, A., … Weigman, J. (2010). The CornerHouse forensic interview protocol: RATAC®. *Thomas M. Cooley Journal of Practical and Clinical*

DOI: 10.1002/bsl

Law, *12*(2), 193–331. Available from: http://ncptc.nonprofitoffice.com/vertical/sites/%7B8634A6E1-FAD2-4381-9C0D-5DC7E93C9410%7D/uploads/%7B0CB5FDDE-6496-40B7-8D70-E4EE6498903D%7D.PDF

APSAC Task Force on Investigative Interviews in Cases of Alleged Child Abuse (2002). *Practice guidelines: Investigative interviewing in cases of alleged child abuse*. Elmhurst, IL: American Professional Society on the Abuse of Children.

Bloom, L., & Tinker, E. (2001). The intentionality model and language acquisition: Engagement, effort, and the essential tension in development. *Society for Research in Child Development*, *66* (4, Serial No. 267).

Bohannon, J. N., & Bonvillian, J. D. (2001). Theoretical approaches to language acquisition. In J. B. Gleason (Ed.), *The development of language* (pp. 254–314). New York: Allyn & Bacon.

Brainerd, C. J., & Reyna, V. F. (2004). Fuzzy-trace theory and memory development. *Developmental Review*, *24*(4), 396–439. doi:10.1016/j.dr.2004.08.005.

Bruck, M., Ceci, S. J., & Francoeur, E. (1999). The accuracy of mothers' memories of conversations with their preschool children. *Journal of Experimental Psychology: Applied*, *5*(1), 89–106. doi:10.1037/1076-898X.5.1.89.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. doi:10.1037/1040-3590.6.4.284.

Davis, D., & Friedman, R. D. (2007). Memory for conversation: The orphan child of witness memory researchers. In M. P. Toglia, J. D. Read, D. F. Ross, & R. C. L. Lindsay (Eds.), *Handbook of eyewitness psychology (Vol. 1): Memory for events* (pp. 3–52). Mahwah, NJ: Lawrence Erlbaum Associates.

Davis, D., Kemmelmeier, M., & Follette, W. C. (2005). Conversational memory on trial. In Y. I. Noy & W. Karwowski (Eds.), *Handbook of human factors in litigation* (pp. 12-1–12-29). Boca Raton, FL: CRC Press.

Duke, S. B., Lee, A. S., & Pager, C. K. W. (2007). A picture's worth a thousand words: Conversational versus eyewitness testimony in criminal convictions. *American Criminal Law Review*, *44*(1), 1–52. Available from: http://digitalcommons.law.yale.edu/fss_papers/829

Ebbinghaus, H. (1885/1964). *Memory* (H. A. Ruger & C. E. Bussenius, Trans.). New York, NY: Dover. (Original work published 1885).

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Thousand Oaks, CA: Sage Publications Inc.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–24. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/

Hedelin, L., & Hjelmquist, E. (1998). Preschool children's mastery of the form/content distinction in communicative tasks. *Journal of Psycholinguistic Research*, *27*(3), 421–452. doi:10.1023/A:1023263906846.

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Wadsworth.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3–28. doi:10.1037/0033-2909.114.1.3.

Kulkofsky, S., & Klemfuss, J. Z. (2008). What the stories children tell can tell about their memory: Narrative skill and young children's suggestibility. *Developmental Psychology*, *44*(5), 1442–1456. doi:10.1037/a0012849.

Lamb, M. E., Orbach, Y., Hershkowitz, I., Esplin, P. W., & Horowitz, D. (2007). A structured forensic interview protocol improves the quality and informativeness of investigative interviews with children: A review of research using the NICHD Investigative Interview Protocol. *Child Abuse & Neglect*, *31*(11–12), 1201–1231. doi:10.1016/j.chiabu.2007.03.021.

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. doi:10.1023/A:1005556404636.

London, K., Bruck, M., Ceci, S. J., & Shuman, D. (2005). Children's disclosure of sexual abuse: What does the research tell us about the ways that children tell? *Psychology, Public Policy, & the Law*, *11*(1), 194–226.

London, K., & Kulkofsky, S. (2009). Factors affecting the reliability of children's reports. In G. M. Davies, & D. B. Wright (Eds.), *New frontiers in applied memory* (pp. 119–141). New York, NY: Psychology Press.

Lyon, T. D., & Ahern, E. C. (2010). Disclosure of child sexual abuse. In J. Myers (Ed.), *The APSAC handbook on child maltreatment* (3rd ed.), pp. 233–252). Newbury Park, CA: Sage.

Lyon, T. D., & Stolzenberg, S. N. (2014). Children's memory for conversations about sexual abuse: Legal and psychological implication. *Roger Williams University Law Review*, *19*(2), 411–450. Available at: http://docs.rwu.edu/rwu_LR/vol19/iss2/5

Malloy, L. C., Brubacher, S. P., & Lamb, M. E. (2013). "Because she's one who listens": Children discuss disclosure recipients in forensic interviews. *Child Maltreatment*, *18*(4), 245–251. doi:10.1177/1077559513497250.

Miller, J. B., deWinstanely, P., & Carey, P. (1996). Memory for conversation. *Memory*, *4*(6), 615–632. doi:10.1080/741940999.

Olson, D. R., & Hildyard, A. (1981). Assent and compliance in children's language. In P. Dickinon (Ed.). *Children's oral communication skills* (pp. 313–335). New York, NY: Academic Press.

DOI: 10.1002/bsl

Osborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment, Research & Evaluation*, *8*(6). Retrieved April 28, 2015 from http://PAREonline.net/getvn.asp?v=8&n=6

Principe, G. F., & Schindewolf, E. (2012). Natural conversations as a source of false memories in children: Implications for the testimony of young witnesses. *Developmental Review*, *32*(3), 205–223.

Reyna, V. F., & Kiernan, B. (1994). Development of gist versus verbatim memory in sentence recognition: Effects of lexical familiarity, semantic content, encoding instructions, and retention interval. *Developmental Psychology*, *30*(2), 178–191. doi:10.1037/0012-1649.30.2.178.

Schaeffer, P., Leventhal, J. M., & Asnes, A. G. (2011). Children's disclosure of sexual abuse: Learning from direct inquiry. *Child Abuse & Neglect*, *35*(5), 343–352. doi:10.1016/j.chiabu.2011.01.014.

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. doi:10.1111/j.1468-2958.1987.tb00127.x.

Stafford, L., & Daly, J. A. (1984). Conversational memory: The effects of recall mode and memory expectancies on remembrances of natural conversation. *Human Communication Research*, *10*(3), 379–402. doi:10.1111/j.1468-2958.1984.tb00024.x.

State of Michigan Governor's Task Force on Children's Justice and Department of Human Services. (2004). State of Michigan Governor's Task Force on Children's Justice and Department of Human Services forensic interviewing protocol. Retrieved from http://www.mi.gov/documents/dhs/DHS-PUB-0779_211637_7.pdf

Stolzenberg, S. N., & Lyon, T. D. (2014). How attorneys question children about the dynamics of sexual abuse and disclosure in criminal trials. *Psychology, Public Policy, and Law*, *20*(1), 19–30. doi:10.1037/a0035000.

Stolzenberg, S. N. & Lyon, T. D. (2015). The consistency and accuracy of children's reports of an adult transgression across multiple interviews. American Psychology-Law Society Conference, San Diego, CA.

U.S. Department of Health and Human Services (2013). *Child Maltreatment 2012*. Washington, DC: U.S. Government Printing Office.

Warren, A. R., & Woodall, C. E. (1999). The reliability of hearsay testimony: How well do interviewers recall their interviews with children? *Psychology, Public Policy, and Law*, *5*(2), 355–371. doi:10.1037/1076-8971.5.2.355.

DOI: 10.1002/bsl