Memory & Cognition
1999, 27 (6), 937-947

Can Bartlett's repeated reproduction experiments be replicated?

ERIK T. BERGMAN and HENRY L. ROEDIGER III
Washington University, St. Louis, Missouri

Surprisingly, Bartlett's (1932) famous repeated reproduction experiments, in which he found systematically increasing errors in recall from the same people tested over time, have never been successfully replicated. Several studies have attempted partial replications, which were unsuccessful, and their authors concluded that the original observations might not be replicable. We conducted a study modeled closely after Bartlett's procedures: Subjects studied "The War of the Ghosts," took an initial test 15 min later, and then took a delayed test after 1 week. A follow-up test was conducted 6 months later on as many subjects as could be obtained. We did replicate Bartlett's results, in that (1) subjects forgot the story over delays but (2) introduced rationalization and distortion into their accounts of the story, with increases in the proportion of material distorted as retention interval increased. Subjects also imported new propositions at long delays, further confirming Bartlett's empirical observations and conclusions. Bartlett's repeated reproduction results can be replicated.

Bartlett (1932) published *Remembering: A study in Experimental and Social Psychology*, a book that has deservedly received its place among the classic works in all of psychology (Roediger, 1997; Thompson, 1997). The book reported research dating from 1913 on perceiving, remembering, thinking, and social and cultural influences on these processes. The book is famous for Bartlett's theoretical ideas on the constructive nature of remembering and the notion that schemata (organized mental structures) were critical to the comprehension, assimilation, and remembering of information.

Bartlett (1932) used several different techniques to study the phenomena in which he was interested. With regard to memory, the two best known methods were the serial reproduction technique and the repeated reproduction technique. In serial reproduction, a subject is exposed to some material, such as a story, and then tries to recall it. A second subject reads the first subject's account of the original event and tries to recall that. The third subject reads the second subject's account, and so on. (The technique is like that in the party game of rumor or telephone.) Serial reproduction can often lead to dramatic distortions in recall over repeated reconstructions of the event (see Bartlett, 1932, pp. 118–153, for examples). Although rarely used now, this experimental technique was used in later studies with results generally confirming those of Bartlett (e.g., Paul, 1959). Psychologists interested in transmission of rumors used the technique, among others (e.g., Allport & Postman, 1947).

The other primary technique that Bartlett (1932) used to study remembering is repeated reproduction. In the most famous use of this method, Bartlett read his subjects the native American folktale, "The War of the Ghosts." After hearing the story read twice, subjects were given a 15-min delay and then tried to recall the story. After this first recall, subjects were tested again at some later point in time, and often they were tested on more than one occasion. "The War of the Ghosts" is an unusual story for the English college students who seem to have been Bartlett's subjects (although this detail is unclear). The story is rather disjointed and contains supernatural elements. Bartlett was interested in how his subjects would alter the story when remembering it across repeated occasions. One of his primary findings was what he called *rationalization*, by which he meant that "whenever anything appeared incomprehensible . . . it was either omitted or explained" by the addition of material (Bartlett, 1932, p. 68). Rationalization occurred "in practically every reproduction or series of reproductions" (Bartlett, 1932, p. 71). In addition, intrusions would occur in recall, when subjects remembered events that were not in the story (sometimes in the process of rationalization). The distortions in recall of the story tended to grow over repeated reproductions, as in the case of the serial reproduction results. However, the distortions tended to be somewhat less dramatic in repeated reproduction than in serial reproduction.

The basic idea that Bartlett proposed was that remembering was a constructive (or reconstructive) process, one rife with error. People remember the schema or gist of an event but forget the details; later, when they recall the event, they may get the gist correct but create details (and even

This research was supported by NIH Grant MH1156-02 to the first author. The authors thank Kate Pfeifer for her help in conducting and scoring this research and Walter Kintsch, Ulric Neisser, and David Rubin for helpful comments. These data were first reported at the 38th Annual Meeting of the Psychonomic Society, Philadelphia, November 1997. Correspondence should be addressed to H. L. Roediger III, Department of Psychology, Box 1125, Washington University, One Brookings Drive, St. Louis, MO 63130-4899 (e-mail: roediger@artsci. wustl.edu).

new themes and importations) that were not in the original story or event. Often, the shifts could be dramatic across repeated retellings of the same story, especially if the interval between retellings was long.

Bartlett's (1932) conclusions appear in countless textbooks; they are a staple for introductory psychology, cognitive psychology, and learning and memory texts. The statements are typically like those in the previous paragraph. However, the results on which Bartlett based these conclusions were largely anecdotal. His testing of subjects was haphazard rather than systematic; he seemed to test people more or less as he bumped into them. In addition, he never presented aggregate results but rather presented selected protocols from individual subjects. The instructions given to subjects for recall are not specified, and no concerted effort seems to have been made to hold the study and testing conditions constant. In his introduction to the new edition (Bartlett, 1932/1995) of *Remembering*, Kintsch (1995) wrote:

> Bartlett had a very informal way of conducting and reporting his experiments. He was mostly concerned that the experimental conditions be as natural as possible and did not worry much about replicable, stable experimental conditions. There are no statistics, and there is little data aggregation. What we get are selected examples. In my opinion, this is the weakest aspect of the book and something that has limited its historical influence. (p. xiv)

Because Bartlett's experiments were so casual and because the conclusions he drew are regarded as so important and are so widely cited, one might expect that his studies would have been replicated many times over, to confirm his preliminary observations. One would be wrong. All texts cite Bartlett's original observations but do not report replications of them, because there seem to be none. Of course, many other observations confirm aspects of Bartlett's conclusions about remembering (see Roediger, 1996, and Schacter, 1995, for overviews of research on memory distortion), but direct replications of the repeated reproduction experiments are not available.

In fact, there are several reasons to suspect that Bartlett's (1932) original observations cannot be replicated. Gauld and Stephenson (1967) investigated the possibility that Bartlett's instructions to his subjects may have led them to construct rather than to remember. As noted above, Bartlett (1932) was vague as to his test instructions, and deliberately so: "I thought it best, for the purposes of these experiments, to try to influence the subjects' procedure as little as possible" (p. 78). Gauld and Stephenson argued that if Bartlett's subjects took their task to be one of retelling a story rather than remembering it, then results showing constructive and invented aspects with considerable distortion would be expected. As they noted,

> Most people who retell a story are unlikely to care very much whether the story they retell is the same, detail by detail, as the story they originally heard. In other words, they are most unlikely to take pains that what they come out with is always what they remember rather than what they guess at or even consciously invent. Now if the changes and inventions in the reproduction of stories ... are to serve as the foundation for a theory of remembering, [then it should be established that the subjects] were indeed seriously trying to remember, and were not more or less consciously guessing or romancing in order to fill in gaps in their memories. (Gauld & Stephenson, 1967, p. 40)

Gauld and Stephenson (1967) carried out three experiments varying the type of instructions that subjects were given while they attempted recall. In general, distortions in recall occurred only under the lenient instructions and not when subjects were encouraged to be accurate. This led the authors to conclude that Bartlett (1932) must have used rather loose instructions, ones that encouraged invention and reconstruction, in his original experiments. They were glum about the prospects of Bartlett's observations being replicable under appropriate conditions encouraging accurate remembering. However, one difference between Gauld and Stephenson's experiments and those of Bartlett is that their tests occurred immediately after study with no retention interval between successive tests. Bartlett, in contrast, waited 15 min before first testing his subjects, and the successive tests were often given weeks, months, or even years (in one or two cases) later. This difference may be critical, as our research demonstrates.

Wheeler and Roediger (1992) also conducted research whose outcome also seems opposed to Bartlett's repeated reproduction findings. They noted that the technique of repeatedly testing memory has a long history in experimental psychology and most results run counter to those Bartlett (1932) obtained. For example, Ballard (1913) gave children passages of poetry, among other materials, and asked them to recall the material repeatedly. He reported that the children often recalled lines of poetry on a later test that they could not recall on an earlier test, a phenomenon he called *reminiscence*. Later research confirmed this finding, and some studies also showed absolute improvements in recall over time—that is, more total items were recalled on later tests than on earlier tests, a phenomenon labeled *hypermnesia* (e.g., Erdelyi & Becker, 1974; Roediger & Payne, 1982).

Wheeler and Roediger (1992) designed experiments to reconcile the discrepancies between Ballard's (1913) and Bartlett's (1932) findings from repeated testing experiments. In one experiment, they gave subjects prose passages in which objects were named. Whenever the object was first named in the story, a picture would appear on a slide. Altogether, subjects studied 60 pictures embedded in the story. Three different groups received no test, one test, or three tests immediately after being presented with the story and pictures. A week later, the subjects returned and received three more tests in succession. Wheeler and Roediger found that, in general, recall improved across the tests. When subjects received three tests immediately after hearing the story, they recalled more pictures on each test. Furthermore, recalling the material just after story presentation benefited recall a week later. Subjects who took three immediate tests remembered the story better a week later relative to subjects who took only a single immediate test; both these groups outperformed subjects who did not take an immediate test. Forgetting did occur across the week delay, but it was relatively modest and was not accompanied by a high proportion of errors, as Bartlett had reported.

Thinking that the difference in materials might be critical, Wheeler and Roediger (1992) conducted a second experiment in which subjects were exposed to either "The War of the Ghosts" or another story. They read the story, were distracted for 5 min, and then recalled it for 8.5 min. After another 5-min delay, they were given another test; a final test occurred 1 week later. The results showed improved recall between the two immediate tests (hypermnesia) and then forgetting a week later. However, despite the forgetting, Wheeler and Roediger did not observe a high proportion of errors to indicate rationalization or inference, even in recall of "The War of the Ghosts." Again, though, Bartlett's (1932) procedures were not followed. Perhaps the accurate recall on the immediate tests helped consolidate accurate retention of the story and thereby reduced the possibility of finding distortion on the later test.

A third challenge to Bartlett's (1932) findings comes from recent work by Wynn and Logie (1998), who asked first-year university students to repeatedly reproduce several events from the first few days of orientation over the next few months. The first recall was 2 weeks after the events in question, and the last recall was 6 months later. They found very little forgetting of the events over time and also very few errors. Furthermore, the errors did not systematically increase over time, as might have been expected from Bartlett's (1932) work. Of course, events surrounding the first few days of university life might be quite salient and, hence, more distinctive and memorable than other events. Also, their procedures did not follow Bartlett's very closely.

Nonetheless, these prior results led us to wonder whether Bartlett's (1932) results could be replicated under carefully controlled conditions. In the present research, we examined Bartlett's methods and did our best to replicate what he seemed to have done. Because the instructions he gave his subjects at test were unclear, we used both strict and lenient instructions during the test, following Gauld and Stephenson (1967), to see whether this difference would matter. We also examined whether the initial instructions and their effect on a first recall test would carry over to a later test when all subjects were tested under the same instructional set.

Subjects read "The War of the Ghosts" twice and then performed a distractor task for 15 min before an initial recall test. One group received general instructions and another group received strict instructions during the initial test, with the strict instructions being for subjects to recall only material in which they were confident; a third group of subjects received no immediate test. After 1 week, all subjects received a second test under either general or strict instructions. Finally, as many subjects as could be contacted were tested 6 months later in a final test under general instructions. We examined recall for accurate statements and for distortions, seeking to replicate Bartlett's (1932) anecdotal results under more tightly controlled conditions with aggregate results. Because Bartlett was so casual about his reporting of results, there is even some issue as to what data would constitute a replication of his work; we discuss this issue in the Discussion section.

METHOD

Subjects

Thirty Washington University undergraduates volunteered to participate in an experiment on story comprehension. The subjects were recruited through advertisements posted around the campus and were paid $10 for attending the two initial experimental sessions. Thirteen of these subjects agreed to participate in another session 6 months later for an additional $6 (all of the original subjects had indicated at the first session that they would be willing to be contacted about further research). The subjects were informed that this additional session was a follow-up study to the one in which they had participated in the previous fall semester.

Materials

Bartlett's (1932/1995, p. 65) version of the Native American story "The War of the Ghosts" was typed on one side of a sheet of paper (single spaced in Times 12-point font) and presented to each subject.

Design

The experiment consisted of a period of study followed by repeated reproductions across three delays, with the instructional set manipulated on the first test (strict or lenient criteria for recall). During the first session, all subjects read the story twice. After a 15-min distractor task, 10 subjects recalled it under strict instructions, 10 recalled it under lenient instructions, and 10 did not recall it at all. One week later, the subjects returned, and half of the subjects from each initial recall group were asked to recall the story again under strict instruction conditions and the other half under lenient conditions. (This variable turned out not to matter at all on the second test, so the data on this test were collapsed over instructions.) All subjects who participated in the 6-month recall test did so under general recall instructions. Thus, the experiment was a 3 (type of initial recall: no test, strict instructions, lenient instructions) × 2 (type of recall at 1 week: strict, lenient) between-subjects factorial design, with a subset of 13 subjects returning for a final general recall at 6 months.

Procedure

All subjects attended two sessions spaced 1 week apart, and 13 subjects returned 6 months later (plus or minus 1 week). The subjects were tested individually or in groups of 2. All instructions were read aloud by the experimenter, and responses were handwritten by the subjects.

Upon arrival for the first session, the subjects were instructed that the experiment involved comprehension of verbal materials. They were given a sheet of paper with the story on it (face down) and asked not to turn it over until so indicated. The subjects were instructed to read the story (to themselves) twice at their normal reading speed and to turn the paper face down again when they had finished. When they were told to begin, the experimenter started timing.

When the subjects finished reading the story, the elapsed time was noted, and the paper was collected. The subjects were next given several sheets of paper with a series of moderately difficult math problems (the distractor task) and asked to complete as many as possible. When 14.5 min had elapsed from when the subjects had finished reading the story, the math papers were collected. The subjects then either were dismissed and asked to return for other tasks 1 week later (the subjects were told that a long rest period was required) or were given strict or lenient recall instructions. (An additional delay of 15 sec occurred before the lenient instructions to equate total delay, because these instructions were shorter.) The lenient instructions were as follows:

> Please write down the story you read earlier. Don't worry about being exact; you are not being tested for accuracy. Just tell the story as you remember it. Imagine you are relating it to a friend who has never heard the story before. When you are through, turn the paper over. You have about 8 min, should you need it.

The strict instructions were as follows:

> Please write down the story you read earlier as best you can. Please try to reproduce it exactly. It is very important that you be as precise as you can. Try to use exactly the same words as they appeared in the story as much as possible. Where you cannot remember the exact wording, be sure to at least get the facts and events exactly correct. Do not invent facts to make it a better story; imagine that you are giving a statement to a policeman and accuracy is important. If you cannot remember something, don't guess, just leave it blank. You have about 8 min, should you need it.

The subjects were additionally instructed to turn the paper over when they felt sure they could remember no more, and the elapsed recall time was recorded by the experimenter.

Exactly 1 week later, all subjects returned to the laboratory and were asked to recall the story they had read the previous week under either lenient or strict instructions, and the amount of time they took to recall the story was recorded.

Six months later, these same subjects were recruited again. All who could be contacted agreed to return to the laboratory for another memory experiment that was described as a follow-up experiment to the one in which they had participated in the previous fall semester. The subjects were brought into the laboratory and were given general instructions to recall the story, with a caution against guessing. When the subjects indicated that they had finished, the elapsed time was noted, and they were given a black-ink pen and asked to use it to continue their recall attempt by adding anything else they could remember in the next few minutes. The subjects were told that if they wanted to change or withdraw anything already there, they should not erase it but rather draw a single line through it as appropriate. The subjects were timed to ensure that the recall attempt continued until a total elapsed recall time of 8 min was reached (excluding the time taken for instructions).

Scoring and Measures

Following Mandler and Johnson's (1977) analysis, the story was divided into 42 propositions or idea units. The subjects' individual narratives were transcribed (typed) and divided into propositions in the same manner as the story. The experimental transcripts were assigned a number and randomly ordered and individually scored by the first author, who was blind to the corresponding experimental condition. Each recall transcript was scored by taking each proposition from the actual story in turn and identifying the proposition(s) in the subjects' recall transcript that had the closest correspondence in terms of the information expressed. The goal of the scoring procedure was to identify the propositions in each recall that were precisely correct, those that were correct with some information missing, and those that were distorted to various degrees and in various ways. The overall strategy was to first employ a very detailed scoring procedure and then derive two more global measures from these detailed scores. The rationale behind this strategy was to ensure that the global measures did not simply correspond to some general impression of the material but were rooted in careful analysis of specific elements.

We adopted a strict criterion for analyzing the data.[1] Accordingly, the two measures of interest were derived from our scoring procedure in the following way. Accurate propositions were the propositions that were either exactly correct or essentially correct with omission but no distortion. Propositions designated as distorted were those that either contained some distorted elements or were entirely distorted. Distorted propositions were further scored as propositions with only minor distortion or propositions with major distortion. Thus, minor distortion was intended to reflect only changes in the surface structure of the propositions (a rephrasing such that the proposition was noticeably changed yet still essentially correct), whereas major distortion reflected changes in the meaning of the proposition.

Major distortion was of three possible types: normalization, inference, or importation. For example, remembering that the story events happened during the "day" instead of at "night" would constitute a normalization because the events (such as hunting) would more typically be expected to take place during the day. Another example of normalization would be replacing "canoe" and "hunting seals" with "boat" and "fishing." A proposition was judged to show inference-based distortion if information that was merely implied by the story but nonetheless could be inferred (correctly or incorrectly) was added to the proposition. For example, recalling that the Indian was hit by an "arrow" (the story never specifies what hit him) was judged an inference-based distortion. Importation was defined as a new element being added to a proposition (the source of the element could be another part of the story or of some indeterminable origin). Finally, entirely new propositions that did not correspond to any of the actual story propositions were counted separately as major intrusions. In order to assess reliability for these measures, an undergraduate assistant studied the scoring manual and was instructed to score the transcripts in the same manner as the primary scorer. Totals for each subject on each measure were tabulated, and Pearson product-moment correlations between the primary scorer and the undergraduate scorer were calculated. The correlations (across subjects) were .87 for the accuracy measure, .92 for the distortion measure, and .79 and .80 for major and minor distortion types, respectively. Identification of totally new propositions (intrusions) was reliable at .86. All data reported below are those of the primary scorer.

RESULTS

All analyses were conducted on the measures described above: the numbers of accurate and distorted propositions produced, with the distorted propositions subdivided into those with only minor distortion and those with major distortion. Each measure reflects the total number of propositions produced by each subject who met the specific criteria for that measure and is expressed as a proportion of the total number possible (42), unless stated otherwise. Initial analysis did not reveal any effects of the form of the second recall (strict or lenient instructions) on accuracy or distortion (all numerical differences were less than .01), so the data were collapsed across this variable. Thus the strict and lenient groups designation will refer to the instructions given on the first recall test for the remainder of the paper. All analyses were significant at the $p = .05$ level or better, unless otherwise noted. Analyses of variance (ANOVAs) and planned comparisons (t tests) were used for statistical testing, except where noted.

First and Second Recall Sessions

Table 1 presents the proportions of propositions recalled accurately or in a distorted manner from the first

Table 1
Mean Proportions of Propositions Recalled (out of 42 Possible), and Standard Deviations, in Either an Accurate or a Distorted Manner in the First and Second Recall Sessions as a Function of the Initial Instructional Set

	Recall Session			
Initial Recall	First (15 min)		Second (1 week)	
Instructions	*M*	*SD*	*M*	*SD*
Strict				
Accurate	.26	.12	.12	.09
Distorted	.33	.09	.37	.14
Proportion of errors	.57	.12	.75	.19
Lenient				
Accurate	.17	.10	.13	.08
Distorted	.38	.10	.36	.10
Proportion of errors	.69	.14	.75	.13
Control				
Accurate			.04	.03
Distorted			.18	.14
Proportion of errors			.81	.06

Note—Proportions of errors are presented beneath their corresponding distortion scores. Proportions of errors were calculated by dividing the number of distorted proportions by the total number of propositions recalled.

and second recall sessions for the three conditions. Again, note that the type of recall instruction only refers to the instructions given in the first test session. In addition, the proportions of errors (the number of distorted propositions divided by the total number of propositions recalled) are also shown.

Accuracy. Examination of Table 1 reveals that, during the first test session, strict recall instructions led to a greater proportion of accurate propositions being recalled than did lenient recall instructions (.26 vs. .17) [$t(18) = 1.87$, *SEM* = 0.05, one-tailed]. However, the subjects given strict recall instructions showed greater forgetting over the week such that, by the second test session, both instructional groups recalled nearly the same proportion of accurate propositions (.12 for strict, .13 for lenient). This pattern was confirmed by a main effect of delay [$F(1,18) = 30.9$, $MS_e = 0.003$] and an interaction between delay and instructional set [$F(1,18) = 8.77$]. This lack of an instructional effect in the second test session cannot be attributed simply to the first test session having entirely lost influence after a week delay: both instructional groups (combined) recalled more during the second session than did the control group who did not recall during the first session ($M = .12$, $SD = .08$ vs. $M = .04$, $SD = .03$) [$t(28) = 3.17$, *SEM* = 0.03].

Distortion. The proportions of propositions recalled in a distorted manner are presented in Table 1 and are expressed as proportions of the total possible. The only reliable effect for this distortion measure was that the subjects who recalled during the first session produced a greater number of distorted propositions during the second test session than did the control subjects: Both instructional groups (combined) showed more distortion than did the control group who did not recall during the first session ($M = .36$, $SD = .12$ vs. $M = .18$, $SD = .14$) [$t(18) = 3.78$, *SEM* = 0.05]. The amount of distortion remained unchanged between the first and second sessions [$F(1,18) = 0.29$, $MS_e = 0.04$] and did not differ by instructional set [$F(1,18) = 0.17$, $MS_e = 0.02$]. Although slightly fewer errors occurred in the first test session with strict instructions than with lenient instructions (.33 vs. .38), a planned comparison between strict and lenient instructions on the first test detected no reliable difference [$t(18) = 1.1$, *SEM* = 0.04].

Another way to measure the amount of distortion is to calculate it as a proportion of the total amount recalled (both accurate and distorted), thus obtaining a proportion of errors, or what is sometimes called an *error rate* (Gauld & Stephenson, 1967). The proportions of errors (the number of distorted propositions produced divided by the total number of propositions produced, either accurate or distorted) for each condition are presented in Table 1. On the first test, strict recall instructions led to a reduced proportion of errors relative to lenient instructions (.57 vs. .69) [$t(18) = 2.11$, *SEM* = 0.06]. This result replicates Gauld and Stephenson's (1967) findings. Importantly, this difference was transitory: Proportions of errors were higher but equivalent on the second recall test (.75 for both groups). This observation was verified by a main effect of delay [$F(1,18) = 13.47$, $MS_e = 0.01$], no effect of instructional set [$F(1,18) = 1.18$, $MS_e = 0.03$], and a marginally significant interaction between delay and instructional set [$F(1,18) = 3.50$, $p = .08$]. It seems that the form of the first test (i.e., whether strict or lenient instructions were used) did not affect the proportion of errors on the second test. Thus, distortion (as indexed by the proportion of errors) increased in the second test session regardless of recall instructions given during the first test session. This is a confirmation of Bartlett's (1932) observations that distortion increases over time and a disconfirmation of Gauld and Stephenson's hypothesis that Bartlett's findings were due to the form of the recall instructions.

One interpretation of this pattern of error proportions is that the distorted material is simply better retained than the accurate material. This differential forgetting could be a result of the distorted material being largely schema based and, so, easy to remember, whereas the accurate material might be nonschema based and, so, more susceptible to forgetting. In essence, the effect observed here would be an effect of the type of material (or perhaps whether the materials were generated by the subject rather than read in the text) and not an effect of some reconstructive process per se. This interpretation does not seem to hold for the present experiment. An item analysis was performed in which the fate of individual items (across subjects) was tracked between the first and second recalls. Of the items that were correct on the first test, a mean proportion of .26 ($SD = .44$) of them was forgotten on the second test; of the items that were distorted on the first test, a mean proportion of .24 ($SD = .43$) was for-

Table 2
Mean Proportions of Propositions (out of 42 Possible), and Standard Deviations, Exhibiting Major Distortion and Those Exhibiting Only Minor Distortion on the First and Second Recall Tests

	Recall Session			
	First (15 min)		Second (1 week)	
Distortion Type	*M*	*SD*	*M*	*SD*
Experimental Condition*				
Major	.15	.06	.20	.09
Proportion of errors	.27	.11	.40	.12
Minor	.21	.09	.17	.07
Proportion of errors	.36	.13	.35	.12
Control Condition				
Major			.13	.10
Proportion of errors			.39	.28
Minor			.06	.08
Proportion of errors			.18	.13

Note—Proportions of errors are presented beneath their corresponding distortion scores. Proportions of errors were calculated by dividing the number of distorted proportions in each category by the total number of propositions recalled. *The results combine the two groups (strict or lenient instructions) on the first test.

gotten on the second test. These proportions did not differ statistically [$t(477) = 0.5$]. Thus, there was no evidence for differential forgetting rates, and, so, the increase in the error proportion across tests was not due to the material difference alone. Rather, the subjects introduced new errors into their accounts.

Interestingly, recalling during the first test session (with either strict or lenient instructions) did not affect the proportion of errors in the second test session: The overall proportion of errors for the experimental group test did not differ reliably from that of the control group ($M = .75$, $SD = .16$ vs. $M = .81$, $SD = .02$) [$t(18) = 1.11$, $SEM = 0.06$]. Thus, recalling during the first test session served to increase the absolute number of distortions produced (as indexed by the overall distortion measure reported above), but not the relative amount of distortion (the proportion of errors).

Distortion type. Table 2 presents a breakdown of the distortion data into major and minor distortion types (as proportions of the 42 possible).[2] There was no effect of instructional set, so the data in Table 2 are collapsed across this variable. The results show an interaction between delay and distortion type [$F(1,18) = 9.10$, $MS_e = 0.004$]. The proportion of major distortions increased from the first to the second test (.15 to .20) [$t(19) = 2.67$, $SEM = 0.02$], a result we take to further confirm Bartlett's (1932) observations. In contrast, the proportion of minor distortions dropped slightly (.21 to .17), but this drop was only marginally significant [$t(19) = 1.99$, $SEM = 0.02$, $p = .06$]. In the second test session, the subjects who had recalled during the first test session produced more major distortions [$t(28) = 2.16$, $SEM = 0.03$] and more minor distortions [$t(28) = 4.52$, $SEM = 0.03$] than did the subjects in the control condition.

As with overall distortion, proportions of errors for the major and minor distortion categories were calculated and are presented in Table 2. A pattern of results similar to that observed with the absolute measures was observed with the proportion of errors data: The proportion of major errors increased from the first to the second test (.27 to .40) [$t(19) = 3.76$, $SEM = 0.03$], whereas the proportion of minor errors dropped a small amount (.36 to .35), and this drop was not statistically significant [$t(28) = .30$, $SEM = 0.04$]. Like the above absolute measures, an initial recall during the first test session increased the proportion of minor errors in the second test session relative to control (.35 vs. .18) [$t(28) = 3.47$, $SEM = 0.05$]. However, unlike the absolute measures, an initial recall did not effect the proportion of major distortion errors relative to control (.40 vs. .39) [$t(28) = 0.08$, $SEM = 0.07$].

Intrusions. The mean number of entirely new propositions (intrusions) that the subjects produced on each test was not reliably affected by instructional set or by delay: Means across these conditions were similar [for strict recall, 1.4 ($SD = 1.3$) in the first test session and 1.5 ($SD = 1.4$) in the second; for lenient recall, 0.80 ($SD = 0.92$) in the first test session and 1.2 ($SD = 1.5$) in the second]. However, the initial test did have an effect here: the subjects who participated in the first test had fewer intrusions of new propositions in the second test session irrespective of instructional set [0.14 ($SD = 1.4$) vs. 3.3 ($SD = 2.3$)] [$t(28) = 2.9$]. Thus, in contrast to the proportions of distorted propositions produced, the number of entirely fictitious propositions was reduced by an initial recall test.

Time spent recalling. The amount of time spent in attempting recall can increase the total amount recalled, even over prolonged periods (e.g., Roediger & Thorpe, 1978). Perhaps the subjects given the strict recall instructions remembered more propositions accurately during the first test session simply because they spent more time trying to recall. Indeed, the subjects receiving strict instructions spent a mean of 453 sec ($SD = 99$) writing the story, whereas the subjects receiving lenient instructions spent a mean of 387 sec ($SD = 95$) writing the story. However, because the variability between conditions was quite high, this difference was not reliable [$t(18) = 1.51$, $SEM = 43.3$, $p = .15$]. Moreover, Pearson product-moment correlations between the time spent in recall and the amount recalled (either accurately or inaccurately) on both tests were low and nonsignificant. The only measure that correlated with the amount of time spent on the first test was the amount of time spent on the second test ($r = .65$, $p < .01$), suggesting that the variance in the amount of time spent in recall may be a function of writing speed or temperament, not recall effort or accuracy.

Third Recall Session

For the third recall session, the number of subjects who participated in each condition was limited to those who could be contacted, but all who were contacted did agree to participate. Nevertheless, the possibility that sub-

Table 3
Mean Proportions of Propositions Recalled (out of 42 Possible), and Standard Deviations, on Recall Sessions 1–3 in Either an Accurate or a Distorted Manner by the Subjects Who Participated in the Third Recall Session

	Recall Session					
	First (15 min)		Second (1 week)		Third (6 months)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Experimental Condition (*n* = 8)						
Accurate	.19	.10	.09	.07	.04	.03
Distorted	.36	.09	.35	.13	.23	.10
Proportion of errors	.67	.13	.79	.20	.81	.15
Control Condition (*n* = 5)						
Accurate			.05	.04	.00	.00
Distorted			.18	.11	.07	.06
Proportion of errors			.77	.05	1.00	.00

Note—Proportions of errors are presented beneath their corresponding distortion scores. Proportions of errors were calculated by dividing the number of distorted proportions in each category by the total number of propositions recalled.

ject selection could affect these results is a concern. To verify that the subjects participating in the third recall were representative of the entire group tested earlier, their data for performance on the first and second tests were examined. All measures fell within 1 standard deviation of the mean for all subjects and conformed to the same pattern of results observed for all subjects. The proportions of propositions recalled on the first, second, and third tests by subjects who participated in the third recall session are presented in Table 3. The data are given for accurate and distorted proposition types. There was no effect of instructional set, so the data in Table 3 are collapsed across this variable.

Accuracy and Distortion. Examination of Table 3 reveals that, across the three tests, the subjects consistently recalled more propositions in a distorted manner than in an accurate manner [$F(1,6) = 34.67$, $MS_e = 0.016$]. However, fewer propositions of either type were recalled across the successive tests: The proportions of both accurate and distorted propositions declined [$F(2,12) = 31.44$, $MS_e = 0.022$]. Participating in the initial test served to increase (overall) the proportion of accurate and distorted propositions produced on the second and third tests relative to the control condition [$F(1,11) = 10.74$, $MS_e = 0.012$]. Furthermore, across the second and third test sessions, the initial recall during the first test session had a greater effect on the proportion of distorted propositions produced than on the proportion of accurate propositions, as evidenced by an interaction [$F(1,11) = 5.16$, $MS_e = 0.010$].

In contrast to the measure of absolute distortion, the relative distortion or proportion of errors increased over successive tests [$F(1,10) = 14.12$, $MS_e = 0.006$]. This observation, coupled with the above-noted general decline in the absolute proportions recalled, indicates that the number of accurate propositions declined more quickly across tests than the number of distorted propositions did. The initial recall during the first test session had no effect on the overall proportion of errors in the second test session [$t(10) = 0.100$, $SEM = 0.102$] but lowered the proportion of errors for the third test [$t(11) = 2.79$, $SEM = 0.067$].

Distortion type. Table 4 shows the proportions of distorted propositions recalled, divided into major and minor distortion types. Examining first the data for the subjects in the experimental condition, the number of major distortions remained nearly the same across all three recalls,[3] whereas the number of minor distortions declined. This observation was confirmed by an interaction between delay and distortion type [$F(2,12) = 4.32$, $MS_e = 0.005$], coupled with a follow-up analysis showing no effect of delay on major distortions [$F(2,14) = 0.664$, $MS_e = 0.003$] but an effect of delay on minor distortions [$F(2,14) = 11.07$, $MS_e = 0.004$]. Recalling in the first test session led the experimental group to later recall (in both the second and the third test sessions) a greater number of propositions distorted in both ways [$F(1,11) = 8.64$, $MS_e = 0.010$].

The proportion of errors data (in Table 4), like the absolute measure, exhibited an interaction between distortion type and delay [$F(2,14) = 12.13$, $MS_e = 0.017$]. However, the interaction observed with proportion of errors was stronger: The relative number of major distortions increased across successive tests [$F(2,14) = 16.03$, $MS_e = 0.011$], whereas the relative number of minor distortions decreased [$F(2,14) = 4.79$, $MS_e = 0.013$]. Recalling in the first test session reduced the proportion of errors for major distortion on the third test relative to the control condition [$t(11) = 4.81$, $SEM = 0.072$] but had no effect on the proportion of errors for minor distortion [$t(11) = 2.14$, $SEM = 0.074$].

Table 4
Mean Proportions of Propositions Recalled (out of 42 Possible), and Standard Deviations, on Tests 1–3 Exhibiting Major Distortion and Those Exhibiting Only Minor Distortion

	Recall Session					
	First (15 min)		Second (1 week)		Third (6 months)	
Distortion Type	*M*	*SD*	*M*	*SD*	*M*	*SD*
Experimental Condition						
Major	.15	.06	.18	.08	.16	.08
Proportion of errors	.29	.13	.39	.14	.58	.14
Minor	.21	.12	.18	.07	.07	.04
Proportion of errors	.37	.15	.40	.14	.23	.15
Control Condition						
Major			.12	.08	.06	.05
Proportion of errors			.41	.23	.93	.10
Minor			.06	.04	.01	.01
Proportion of errors			.21	.13	.07	.10

Note—Proportions of errors are presented beneath their corresponding distortion scores. Proportions of errors were calculated by dividing the number of distorted proportions in each category by the total number of propositions recalled.

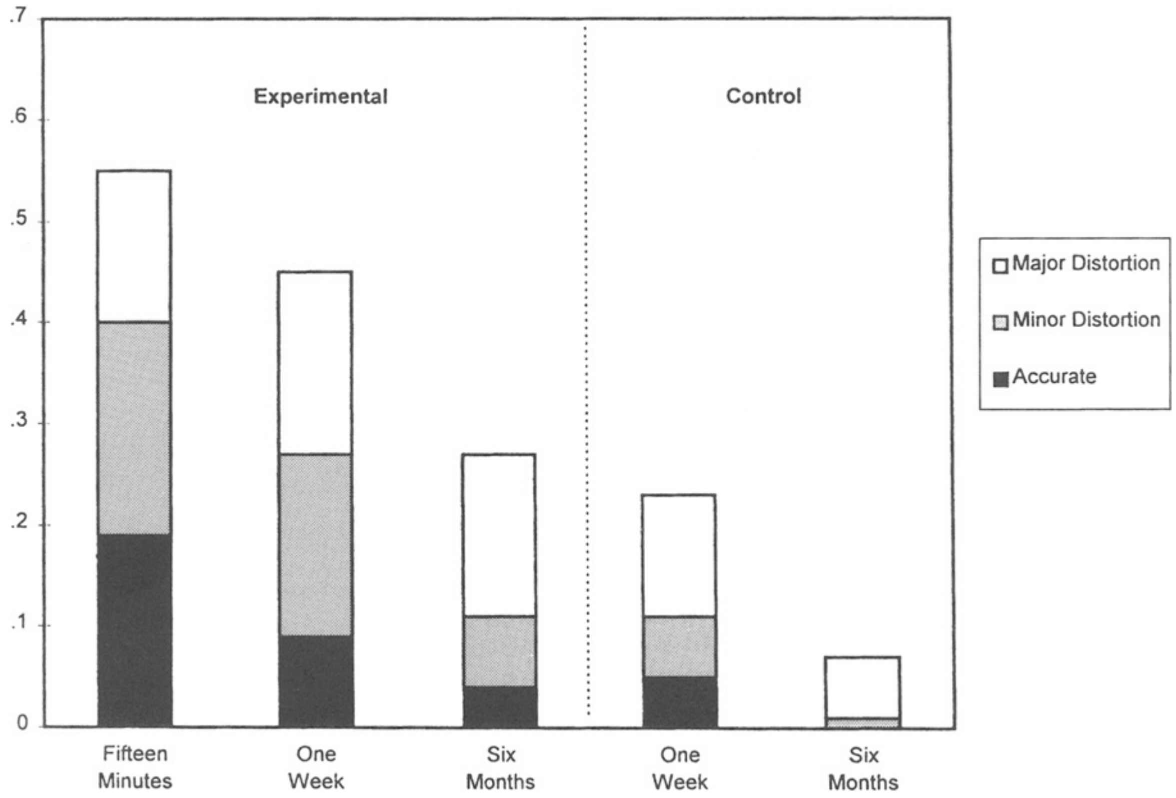

Figure 1. Proportions of propositions recalled (out of 42 possible), divided into propositions that were accurate or distorted in either a major or a minor fashion, for the first, second, and third recall sessions. Data from the subjects who participated in the third recall session are included in this figure. The two control conditions represent performance on the second and third recalls for the subjects who had no initial recall.

Intrusions. The mean number of entirely new propositions intruded in the third test session was 2.5 (*SD* = 2.1) for the subjects who had received two prior tests. This figure did not reliably differ from that for the subjects in the control condition given only one test (*M* = 4.0, *SD* = 4.2) [*t*(11) = 0.856, *SEM* = 1.75], although the power was quite low to detect an effect. The trend here was in the same direction as observed on the second test, with more prior tests reducing the likelihood of intrusions. It is also interesting that the subjects intruded entirely new ideas in the final test.

Additional recall time. The effects of encouraging the subjects to continue their attempt at recall beyond the point of self-termination were simple: Additional retrieval time only resulted in more distortion. There was no effect of adding additional time on the production of accurate statements. Providing additional time led to an overall average of 1 more distorted proposition being recalled [F(1,12) = 15.60, MS_e = 0.42]. Additional time also caused an average of 2.4 more intrusions to be produced [F(1,12) = 10.43, MS_e = 3.55], which represents the only manipulation to affect the number of intrusions.

DISCUSSION

The aim of our work was to replicate Bartlett's (1932) repeated reproduction experiments. However, as noted in the introduction, because Bartlett was so casual in his reporting of results, determining what data would constitute a replication is not necessarily obvious. We believe the results reported in the preceding section can be considered a replication of Bartlett's (1932) essential findings and that our results do support his claims regarding repeated reproduction.

Figure 1, which shows graphically the data summarized in Tables 3 and 4, reveals a strong case for successful replication. Shown on the left side of the figure are proportions of propositions recalled for the subjects who received all three tests, with each column divided into accurate recall (darkest shading), minor errors (gray shading), and major distortions (the white section). As can be seen, accurate recall drops over the three tests, indicating forgetting. This outcome is hardly a surprise, but it was shown in Barlett's (1932) protocols. More interestingly, the number of major distortions recalled did not decrease over repeated tests, and the number of minor distortions decreased only between the second and third tests. Considering only the proportion of errors for major distortions, across the three tests, it rose from .29 after 15 min to .39 after 1 week to .58 after a 6-month delay. In short, over time, accurate recall of the prose passage declined and distortion as proportion of total recall increased dramatically. After 6 months, most of what people recalled of the story was distorted in some form. These aggregate data confirm Bartlett's observations of his subjects' protocols. In addition, the number of intrusions (totally new propositions) also tended to increase over time, also consistent with Bartlett's (1932) observations.

We included a control condition that Bartlett (1932) did not have, but it provided informative results. On the right side of Figure 1 are data from the subjects who did not have an initial recall test but did receive the test after a 1-week delay and after 6 months. Interestingly, the absolute levels of both accurate and distorted recall are less for these subjects than for the subjects receiving all three tests. As noted by Roediger, McDermott, and Goff (1997), the act of taking a test can enhance both accurate and false recall, a point to which we return below.

In the remainder of this section, we discuss several implications of these results for the historical issue of Bartlett's (1932) results and conclusions and for current issues in memory research.

Relation to Prior Work

As described in the introduction, research by Gauld and Stephenson (1967) and by Wheeler and Roediger (1992; see too Roediger, Wheeler, & Rajaram, 1993) cast doubt on the replicability of Bartlett's (1932) findings. The present research not only establishes the replicability of Bartlett's findings but provides a plausible reconciliation of the various findings. Gauld and Stephenson showed that errors in prose recall were more likely under lenient instructions than under strict instructions on tests given shortly after study. We replicated this effect but nonetheless (on long-delayed tests) were able to replicate Bartlett's results. We suspect that Gauld and Stephenson's use of tests after relatively short retention intervals prevented them from replicating Bartlett's work.

In the same vein, Wheeler and Roediger (1992) tested subjects repeatedly in recall of prose passages, including "The War of the Ghosts," and actually found improvement over repeated tests when there were short intervals between tests. Again, however, longer retention intervals are probably necessary to get Bartlett's (1932) effects. When Wheeler and Roediger used somewhat longer intervals, these were only 1 week and, in addition, subjects had been tested twice soon after study, and the first two tests may have had the function of "freezing" (Howe, 1970; Kay, 1955) the account of the story. In short, to replicate Bartlett's twin findings of dramatic increases in forgetting and in distortion in the repeated reproduction procedure, researchers probably need to have only a single test shortly after study (to prevent the beneficial effects of repeated testing on later tests) and long retention intervals (to permit forgetting of the veridical information and, arguably, to lead the rememberer to a more constructive mode of recollection later).

While conducting this research, we discovered a report by Johnson (1962), which can be construed as containing a replication of Bartlett's (1932) research. Johnson used "The War of the Ghosts" and the repeated reproduction technique, but he was interested in determining whether schematically distorted material is retained better than nondistorted material, an idea known as the *reorganization hypothesis*. He found no evidence for this hypothesis (and neither did we). However, in the course of his research he showed, as we have too, both forgetting over time and an increase in distortions. Although Johnson did not calculate an overall proportion of errors as we did, it can be calculated from the data he reported. Interestingly, the proportions of errors we derived from his data are similar to the proportions of errors we obtained at the same retention intervals: At 15 min and at 1 week (the two retention intervals comparable to ours), his data show proportions of errors of .24 and .39, respectively (we obtained .29 and .39).

Does It Matter?

A critic of the present research could argue that it is all beside the point. Regardless of whether repeated reproduction experiments produce results like those Bartlett (1932) claimed to find, we know that the general ideas he proposed—remembering being a constructive process guided by schemas, with distortion and error frequently occurring—are true. After all, much research on prose retention over the last 30 years has shown that meaning is a critical dimension of encoding and that meaning-based errors are quite likely (e.g., Bransford & Franks, 1971; Brewer, 1977; Owens, Bower, & Black, 1979; Spiro, 1980; Sulin & Dooling, 1974; see Alba & Hasher, 1983, for a review). This work provides general support for Bartlett's theorizing.

All this is true, but the paradigms typically used involve only a single test. Bartlett's (1932) procedure of repeated reproduction captures a significant aspect of remembering that other techniques miss. Many of the salient events of our lives are repeatedly remembered; we may recount many times a vivid childhood experience, the events surrounding a traffic accident, or favorite college memories. Therefore, it is critical to ask whether changes occur systematically across repeated recountings of an event. In agreement with Bartlett's results, we found forgetting of the actual detail of the story and an increase in major distortions across repeated tests. It remains to be seen as to how general this result is, because (as noted in the introduction) other researchers have found little change across repeated tests of autobiographical events

(Wynn & Logie, 1998) or even improvements across repeated tests (Ballard, 1913; Erdelyi & Becker, 1974). We suspect that the type of material, the type of test, intervals between tests, and probably other factors will be critical determinants of when repeated testing reveals improvement or distortion across time.

The present results agree with many others in showing that the act of recall is not a neutral event that leaves memory for an event unchanged. The act of recall can both increase and decrease distortions in memory. In the present results, the subjects who took a test relatively soon after study produced fewer major intrusions 1 week or 6 months later, relative to the subjects in a condition with no immediate test. Retrieval during an immediate test serves as a memory modifier (Bjork, 1975), enhancing later recall. However, if subjects make errors on an immediate test, retention of these distortions is also more likely on later tests. Our results showing this outcome agree with those of many others (e.g., McDermott, 1996; Roediger, Jacoby, & McDermott, 1996; Roediger & McDermott, 1995; Schooler, Foster, & Loftus, 1988). Repeated recollection can therefore function as a double-edged sword, increasing accurate recall and limiting major distortions but also—if conditions are right to produce errors in immediate recall—increasing the later recollection of those errors as fact. Bartlett (1932) correctly called our attention to the importance of repeated testing in an understanding of how memories change over time. His repeated reproduction results, shown only in sample protocols, replicate under better controlled conditions with more rigorous scoring procedures.

REFERENCES

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, **93**, 203-231.

Allport, G. W., & Postman, L. (1947). *The psychology of rumor*. New York: Holt.

Ballard, P. B. (1913). Oblivescence and reminiscence. *British Journal of Psychology Monograph Supplements*, **1**, 1-82.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Bartlett, F. C. (1995). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press. (Original work published 1932)

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Erlbaum.

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology*, **2**, 331-350.

Brewer, W. F. (1977). Memory for the pragmatic implication of sentences. *Memory & Cognition*, **5**, 673-678.

Erdelyi, M. H., & Becker, J. (1974). Hypermnesia for pictures: Incremental memory for pictures but not for words in multiple recall trials. *Cognitive Psychology*, **6**, 159-171.

Gauld, A., & Stephenson, G. M. (1967). Some experiments related to Bartlett's theory of remembering. *British Journal of Psychology*, **58**, 39-49.

Howe, M. J. A. (1970). Repeated presentation and recall of meaningful prose. *Journal of Educational Psychology*, **61**, 214-215.

Johnson, R. E. (1962). The retention of qualitative changes in learning. *Journal of Verbal Learning & Verbal Behavior*, **1**, 218-223.

Kay, H. (1955). Learning and retaining verbal material. *British Journal of Psychology*, **46**, 81-100.

Kintsch, W. (1995). Introduction. In F. C. Bartlett, *Remembering: A study in experimental and social psychology* (pp. xi-xv). Cambridge: Cambridge University Press.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, **9**, 111-151.

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory & Language*, **35**, 212-230.

Owens, J., Bower, G. H., & Black, J. B. (1979). The "soap opera" effect in story recall. *Memory & Cognition*, **7**, 185-191.

Paul, I. H. (1959). Studies in remembering: The reproduction of connected and extended verbal material [Monograph]. *Psychological Issues*, **1** (whole No. 2).

Roediger, H. L., III (1996). Memory illusions. *Journal of Memory & Language*, **35**, 76-100.

Roediger, H. L., III (1997). Remembering. *Contemporary Psychology*, **42**, 488-492.

Roediger, H. L., III, Jacoby, D., & McDermott, K. B. (1996). Misinformation effects in recall: Creating false memories through repeated retrieval. *Journal of Memory & Language*, **35**, 300-318.

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 803-814.

Roediger, H. L., III, McDermott, K. B., & Goff, L. M. (1997). Recovery of true and false memories: Paradoxical effects of repeated testing. In M. A. Conway (Ed.), *Recovered memories and false memories* (pp. 118-149). Oxford: Oxford University Press.

Roediger, H. L., III, & Payne, D. G. (1982). Hypermnesia: The effects of repeated testing. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **8**, 66-72.

Roediger, H. L., III, & Thorpe, L. A. (1978). The role of recall time in producing hypermnesia. *Memory & Cognition*, **6**, 296-305.

Roediger, H. L., III, Wheeler, M. A., & Rajaram, S. (1993). Remembering, knowing and reconstructing the past. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 30, pp. 97-134). New York: Academic Press.

Schacter, D. L. (1995). Memory distortion: History and current status. In D. L. Schacter, J. T. Coyle, G. D. Fischbach, M. M. Mesulam, & L. E. Sullivan (Eds.), *Memory distortion* (pp. 1-43). Cambridge, MA: Harvard University Press.

Schooler, J. W., Foster, R. A., & Loftus, E. F. (1988). Some deleterious consequences of the act of recollection. *Memory & Cognition*, **16**, 243-251.

Spiro, R. J. (1980). Accommodative reconstruction in prose recall. *Journal of Verbal Learning & Verbal Behavior*, **19**, 84-95.

Sulin, R. A., & Dooling, D. J. (1974). Intrusion of a thematic idea in retention of prose. *Journal of Experimental Psychology*, **103**, 255-262.

Thompson, C. P. (1997). Schematic and social influences on memory. *Contemporary Psychology*, **42**, 492-493.

Wheeler, M. A., & Roediger, H. L., III (1992). Disparate results of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, **3**, 240-245.

Wynn, V. E., & Logie, R. H. (1998). The veracity of long-term memories: Did Bartlett get it right? *Applied Cognitive Psychology*, **12**, 1-20.

NOTES

1. The idea embodied here is that we are scoring relative to verbatim recall. Obviously, such verbatim recall is unlikely, but it provides a standard by which to judge the amount of divergence. The complete scoring manual used here, including greater detail and several examples, is available on line at http://www.artsci.wustl.edu/~ebergman/bartlett.htm

2. Note that these submeasures can overlap. The minor distortion measure reflects propositions with minor distortion only. However, some propositions containing major distortion contained minor distortion as well. In the first recall, of the propositions showing major dis-

tortion, a mean proportion (out of 42) of .06 (*SD* = .06) propositions also showed minor distortion. In the second recall, .09 (*SD* = .05) propositions showing major distortion also showed minor distortion. For the control condition, the overlap was *M* = .06 (*SD* = .06). There was no reliable difference in the proportion of propositions showing overlap in the first test session and the proportion in the second test session, nor was there a difference between the experimental group and the control condition in the second test session.

3. Note that, with the full data set, the number of propositions with major distortion increased from the first recall to the second recall. Indeed, the partial data set for the subjects who participated in the third recall session shows a similar (but slightly smaller) increase for the major distortion category. However, this difference was not confirmed by an overall ANOVA with the partial data set.

(Manuscript received August 6, 1998;
revision accepted for publication February 9, 1999.)