Published in final edited form as:
Memory. 2004 September ; 12(5): 586–602. doi:10.1080/09658210344000125.

Illusory recollection of voices

Henry L. Roediger III,
Washington University in St. Louis, MO, USA

Kathleen B. McDermott,
Washington University in St. Louis, MO, USA

David B. Pisoni, and
Indiana University, Bloomington, IN, USA

David A. Gallo
Harvard University, Cambridge, MA, USA

Abstract

We investigated source misattributions in the DRM false memory paradigm (Deese, 1959, Roediger & McDermott, 1995). Subjects studied words in one of two voices, manipulated between-lists (pure-voice lists) or within-list (mixed-voice lists), and were subsequently given a recognition test with voice-attribution judgements. Experiments 1 and 2 used visual tests. With pure-voice lists (Experiment 1), subjects frequently attributed related lures to the corresponding study voice, despite having the option to not respond. Further, these erroneous attributions remained high with mixed-voice lists (Experiment 2). Thus, even when their related lists were not associated with a particular voice, subjects misattributed the lures to one of the voices. Attributions for studied items were fairly accurate in both cases. Experiments 3 and 4 used auditory tests. With pure-voice lists (Experiment 3), subjects frequently attributed related lures and studied items to the corresponding study voice, regardless of the test voice. In contrast, with mixed-voice lists (Experiment 4), subjects frequently attributed related lures and studied items to the corresponding test voice, regardless of the study voice. These findings indicate that source attributions can be sensitive to voice information provided either at study or at test, even though this information is irrelevant for related lures.

In the present study, we investigated subjects' phenomenological experience during both true and false recognition in the DRM paradigm (Deese, 1959; Roediger & McDermott, 1995). In this paradigm, subjects encode lists of related words (e.g., hill, valley, climb), and interest lies in the high tendency to falsely recall and recognise a word (e.g., mountain) that was not studied but that is related to the studied words. Unlike other types of laboratory-induced false memories, these false memories are subjectively compelling, as indexed by "remember/know" judgements and other methodologies (see Roediger, McDermott, & Robinson, 1998, and Lampinen, Neuschatz, & Payne, 1998, for early reviews). To investigate this illusory phenomenology, we examined subjects' source attributions to nonpresented but related lures (e.g., mountain) when the lists were presented using two voices. Previous research on recognition memory has demonstrated that listeners are good at remembering not only the specific words that are presented to them during study but also the physical attributes of the voice of presentation (Fisher & Cuervo, 1983; Palmeri, Goldinger & Pisoni, 1993). This outcome suggests that memory representations for spoken words

Correspondence should be addressed to Henry L. Roediger III, Department of Psychology, Washington University, One Brookings Drive, Campus Box 1125, St. Louis, MO 63130-4899, USA. roediger@artsci.wustl.edu.

contain highly specific details about the event of hearing the word (see also Goldinger, 1992). In the present series of experiments, we were interested in how false recognition can come to be similarly detailed, and in understanding the attribution process underlying memory for such details corresponding to false memories.

Several previous studies have investigated voice attributions in the DRM paradigm. The general conclusion from these studies has been that subjects are often quite willing to claim to remember which voice spoke the nonpresented but related critical word. The first investigation of this phenomenon was reported by Payne, Elie, Blackwell, and Neuschatz (1996). In their Experiment 3, subjects heard eight-item DRM lists presented with two voices. There were two types of voice conditions (manipulated within-subject). In the pure-voice condition, the words from each list (i.e., all words associated to the critical nonpresented word *mountain)* were presented in a single voice; in the mixed-voice conditions, the words from each list were presented in both voices (which were either switched in the middle of the list, or were alternated throughout the list). Following study of all the lists, the subjects recalled the words from the list three times. After the third recall attempt, subjects were asked to indicate beside each recalled word whether it had been presented using a male voice or a female voice, or to indicate that they could not recollect the voice in which the word had originally been presented.

In Payne et al.'s experiment, the probabilities of both veridical and false recall were just over .30. Subjects made voice attributions for 80–90% of the recalled words, both for the studied words and the critical nonstudied words, and rarely used the "don't know" category even for words that had not been presented. Voice attributions occurred somewhat more frequently for list words (.94) than for critical words (.87), but attributions for critical items were still quite frequent. Attributions for list items were quite accurate, and correct attributions were greater in the pure case (.84) than in the mixed case (.71 with alternating voices). In contrast, attributions for critical items in the pure case were close to chance (.53), indicating that subjects were just as likely to claim that the critical item had occurred in either voice.

Using different procedures, Mather, Henkel, and Johnson (1997) also found that subjects were willing to make voice attributions to the critical nonpresented words. In their experiment, subjects heard 10-item DRM lists and on a subsequent recognition test made voice-attribution judgements (as well as other types of judgements). Mather et al. (1997) also included a pure-voice and mixed-voice (alternating voices) manipulation. Like Payne et al. (1996), Mather et al. (1997) found that pure-voice presentation led to greater attributions for list items (.84) than did mixed-voice presentation (.65). In contrast to Payne et al. (1996), the critical word was frequently attributed to the voice of its corresponding list (in the pure-voice condition). In fact, "correct" attributions for critical items (.83) were not different from those given to list items (.84) in their study.

As discussed by Mather et al., one reason for this discrepancy may have been that the format of presentation (pure vs mixed) was manipulated between-subjects in their experiment, whereas this format was manipulated within-subject in Payne et al. As a result, subjects in the pure-voice condition of Mather et al. may have been better able to use their knowledge of the structure of the study lists to infer that the critical lure had been presented in the voice of its corresponding list. Consistent with this claim, using procedures that were more similar to those of Mather et al. (i.e., a pure-list manipulation and a visual recognition test), we have also found that critical lures were often attributed to the source of their corresponding list (Gallo, McDermott, Percer, & Roediger, 2001); in our previous study, however, we compared auditory and visual presentations, which may differ from between-voice comparisons.

One issue that may be important to understanding these erroneous voice attributions is the rate of misattributions to critical items following mixed-voice presentation. If these attributions are driven primarily by inferential processes that rely on knowledge of the list structure, then one might expect them to be greater following pure-voice conditions (i.e., "This word's list was presented in a female voice, so this word, too, must have been presented in that voice"). In contrast, if they are driven by the illusory recollection of perceptual characteristics imagined during study, or by a more automatic familiarity-based attribution process at retrieval, then they may not be affected by a pure/mixed manipulation. Although both Payne et al. and Mather et al. included mixed-voice conditions, they did not separately report the overall level of attributions for critical items following mixed lists.

Hicks and Marsh (2001, Experiment 2b) reported data relevant to this issue. Subjects studied six 15-item DRM lists, presented with alternating male and female voices. At test, subjects made voice-attribution judgements for those items that they identified as "old" (a "don't know" option was not included). In this study, subjects were equally likely to attribute critical lures to the male and female voices (whereas attributions for list items were fairly accurate). Thus, as might be expected, when critical lures could not be logically attributed to a particular voice (because their corresponding list had been presented by both voices), subjects were equally likely to attribute the lure to either voice. More interestingly, the attribution rate for critical lures (mean = .45, averaging across source attributions) was at least as great as that for studied items (.40), demonstrating robust levels of misattribution.

Lampinen, Neuschatz, and Payne (1999) also used mixed-list presentation. In their Experiment 1, subjects were presented with 10 DRM lists (each of which was 10 words long), using mixed-voice presentation. After a final recognition test, subjects were asked to make voice-attributions for each "old" word (with a "don't know" option included). Unlike Hicks and Marsh (2001), Lampinen et al. found that source attributions were more likely for list items (.83) than for critical items (.74), but this was probably because shorter lists were used in this experiment (and subjects were given a "don't know" option). More importantly, attributions for critical items were still much higher than those to unrelated lures (.52), and subjects were unlikely to change many of these attributions, even after they were told that they had made some mistakes. These results, coupled with those of Hicks and Marsh (2001), demonstrate that voice attributions to critical lures remain quite high even under mixed-voice presentation conditions. However, it is still unclear whether these attributions will be made as frequently as those following the pure-voice manipulation because a direct comparison between these two methodologies has never been reported.

To more directly investigate this and other issues, in the present study we investigated voice attributions in a single series of comparable experiments. All of the experiments used the typical DRM paradigm (with 15-word lists). Half of the study lists were followed immediately by a free recall test and half were followed by math problems (as in Roediger & McDermott, 1995). All lists were tested with a final recognition test that included voice-attribution judgements. Words were spoken by one of two talkers during encoding. In all experiments, subjects were given a "don't know" option, as in Payne et al. (1996). In this way, we can be more confident that any erroneous source attributions were made on the basis of subjectively compelling phenomenology. The first two experiments manipulated voice presentation between lists (pure voice, Experiment 1) or within-lists (mixed-voice, Experiment 2) with other conditions held constant to justify comparison of the results. Experiments 1 and 2 can be considered as a single large experiment, in which presentation (pure or mixed voice) was manipulated. Both experiments used visual recognition tests, as were used in all of the previous experiments reviewed above. Given prior research, we expected that critical lures would often be erroneously attributed to a voice, but it was

unclear if these attributions would be enhanced by pure-voice presentation (relative to mixed-voice presentation).

We had three main goals in these first two experiments. First, under pure-voice conditions, we sought to determine if critical lures would be attributed to the same voice that had spoken the corresponding list (as in Mather et al., 1997), or if these attributions would be distributed evenly across sources (as in Payne et al., 1996). Second, we wanted to directly compare the levels of voice attributions between pure-voice conditions (Experiment 1) and mixed-voice conditions (Experiment 2). If these erroneous voice attributions are driven by knowledge about the list structure, then we would expect them to occur more frequently under pure-voice conditions.

The third motivation for the current experiments was to further investigate one aspect of Roediger and McDermott's (1995, Experiment 2) results that has not been consistently replicated. They showed that prior recall of lists of words generally increased later recognition, both for studied items and for related lures. However, others have not consistently obtained these effects (see Roediger et al., 1998, for discussion). Experiments in other paradigms have reported generally small positive effects of a prior free recall test on later recognition for studied items, but only for items occurring at the end of the list (see Jones & Roediger, 1995; Lockhart, 1975). Therefore, the inconsistencies in the outcome may not be too surprising. However, given that McDermott (1996) has shown that prior recall has a powerful effect on later true and false recall, we suspect that the effects of prior recall on false recognition are real, but simply harder to detect.

Because repeated testing seems to represent a key aspect to the development of false memories (see Roediger, McDermott, & Goff, 1997), the effect of recall on later recognition in this paradigm deserves more careful examination. This is especially true because recognition memory is an ideal testing ground for the measurement of illusory phenomenology, which may also be affected by prior recall. Two previous studies are directly relevant. In the aforementioned Lampinen et al. (1999) experiment, false recognition of critical lures was not affected by prior recall (relative to math), and neither were the corresponding voice attributions (there were no effects on studied items, either). Similarly, in the aforementioned Gallo et al. (2001) study, prior recall did not reliably increase false recognition of critical lures (relative to math). However, in that study, prior recall did enhance "remember" judgements given to falsely recognised critical lures, and it also increased erroneous source attributions given to these items (and an even larger effect was found for list items). Thus, not only does prior recall sometimes boost false recognition, but it can also boost the illusory phenomenology that accompanies these false memories. The present experiments provided a further test of this idea.

In addition to these first two experiments, which serve primarily to clarify several previous findings and provide new comparisons, we conducted two additional experiments. These experiments served as the companions to Experiments 1 and 2, with a pure-voice manipulation used in Experiment 3 and a mixed-voice manipulation used in Experiment 4. The novel manipulation in these experiments was to use auditory presentation at test, so that presenting a test item in the same or different voice as at study was possible. In this way we were able to investigate whether source attributions would be influenced by specific voice information presented during retrieval, thereby providing additional insights into the attribution process that can lead to illusory recollection. We will discuss the theoretical rationale for these latter two experiments following the presentation of the first two experiments.

EXPERIMENT 1

In Experiment 1, each list was presented by one of two voices (pure-voice presentation), and a visual recognition test was used.

Method

Subjects—A total of 34 Indiana University undergraduates participated in a 1 hour session in partial fulfilment of an Introductory Psychology course requirement. All the subjects were native English speakers and reported no speech or hearing impairments at the time of testing. The data from one subject were omitted from the final analysis because he failed to follow instructions and recalled items on all (instead of half) of the lists.

Materials—The 24 word lists developed by Roediger and McDermott (1995) were used in the present experiment. Each list consisted of 15 words that were highly associated to a non-presented word, with the highest associates occurring first. Each item was separately recorded digitally by both a male and female talker, using a sound card running at a 20 kHz sampling rate with 16 bit resolution. The root-mean-square amplitude of all stimulus items was equated using a signal-processing package.

Design—The 24 lists were divided into three sets of eight. These sets were counterbalanced through the study + recall, study + math, and nonstudied conditions, with roughly equal numbers of subjects tested in each counterbalancing condition (because subjects were tested in groups of one to six, exact numbers were not obtained in each condition). All subjects were presented with 16 lists during the study phase of the experiment, with 8 of the lists tested for immediate free recall and 8 lists followed by math problems. Half of the lists in each testing condition were presented using a male voice, and the other half were presented using a female voice. The remaining eight lists were not presented during the study phase and provided unrelated lures that were used on the subsequent recognition test.

The recognition test was composed of 96 randomly arranged items, 48 of which had been studied and 48 of which had not. The 48 studied items were obtained by selecting three items from each of the 16 presented lists (always from the serial positions 1,7,10). The nonstudied items consisted of the 24 critical lures from all 24 lists (16 studied lists, 8 nonstudied) and the 24 items from the nonstudied list (again always those from positions 1, 7, and 10).

Procedure—Subjects were told that they would be participating in a memory experiment in which they would hear lists of spoken words over their headphones. They were told that after the presentation of each list they would hear a tone or knock (with examples given) that would indicate whether they should recall the list or perform some math problems. Subjects were instructed to listen carefully to each list because the signal for the task would not occur until after the list was presented; therefore, subjects were unaware until the end of the list whether they would be required to recall the items. The inter-stimulus interval was 1.5 seconds within lists. Subjects were given 1 minute after presentation of the signal either to recall items or to do math problems. Each of these tasks was performed on a piece of paper supplied by the experimenter. After 1 minute, a tone occurred and subjects were instructed to turn over their response sheets (so they were no longer in view) and to prepare for the next word list.

The recognition test occurred about 5 minutes after the last test or math period. During this time, subjects were given instructions about the recognition and voice-attribution

judgements. They were told that they would see one item at a time presented on a CRT screen and that they would be required to make one of four responses. If the item had been presented in the previous study phase and they remembered the voice in which it was originally spoken, then they were to press the "male–old" or "female-old" buttons, accordingly. If they thought the item had been presented but did not know the voice, then they were to press a button simply labelled "old". Finally, if they did not think the item had been presented, they were to press "new". The labels on the response boxes were changed for each group of subjects, to balance order of response buttons to type of response across subjects.

Results

All effects that are reported as significant in this series of experiments were significant at $p < .05$.

Recall and recognition—On the immediate free recall tests, true recall of list items was .41 (averaged across all 15 items). False recall of critical items was .29, which was slightly lower than true recall of list items from middle serial positions (6–10, mean = .34). This pattern is similar to that found by Roediger and McDermott (1995) and many others, and demonstrates robust levels of false recall of the critical items.

The recognition results are presented in Table 1. Overall recognition is presented in the leftmost column, which is then decomposed into the three types of responses (male-old, female-old, or don't know). The proportion of the items recognised as "old" that were attributed to a voice (collapsing across male or female) is presented in the rightmost column. Consider first the overall recognition data. Recognition of list items (mean = .72, across conditions) was similar to false recognition of critical items (.71), as is typically the case in this paradigm. There was little evidence for the predicted effect of prior task on subsequent recognition (i.e., recall > math) for either type of item. A 3 (counterbalancing order) × 2 (item type: studied vs critical lure) × 2 (task type: recall vs math) ANOVA confirmed that there were no main effects of order, $F(2,30) = 1.42$, type of item, $F(1,30) < 1$, or task type, $F(1,30) < 1$. There were no significant interactions. False alarms to unrelated lures were quite low (.12), indicating that subjects were not simply guessing "old" during the test.

Source attributions—The primary question investigated in this experiment is the extent to which subjects would be willing to assign a study voice to falsely recognised critical lures. Collapsing across all other variables, the overall levels of voice attributions to critical lures (.38) were about the same as those to list items (.39), demonstrating a powerful misattribution effect. Further, as can be seen from Table 1, both list items and critical lures tended to be attributed to the voice that had presented their corresponding list more often than to the other voice. A 2 (item type: studied vs critical lure) × 2 (task type: recall vs math) × 2 (study voice: male vs female) × 2 (voice attribution: consistent vs inconsistent with study voice) ANOVA revealed only one main effect of attribution $F(1,32) = 10.93$. This effect confirms that subjects were relatively accurate at correctly identifying the voice that spoke the words in these lists (.20 consistent, .08 inconsistent), and that a parallel effect was found for the critical lures (.20 consistent, .07 inconsistent). In contrast, attributions to items from nonstudied lists were much lower and were equally distributed between the two voices. The finding that critical lures were just as likely to have been attributed to the source of their corresponding list as were actually presented items replicates Mather et al. (1997), and suggests that the somewhat different results obtained by Payne et al. (1996) were due to procedures that were unique to their experiment (as discussed in our Introduction).

The failure to find an effect of prior task (recall or math) on source attributions is not surprising, considering that no effect was found in overall recognition. However, note that the overall proportion of list items that were attributed to a voice (collapsing across male and female) tended to be greater following recall than following math (.42 vs .36), and similarly for critical lures (.40 vs .35). Thus, these effects were in the predicted direction, and we revisit them in the subsequent experiments reported here.

EXPERIMENT 2

In Experiment 2, each list was presented by two voices (mixed-voice presentation), and a visual recognition test was again used. Other conditions were held constant.

Method

Subjects—*A* total of 32 Indiana University undergraduates were drawn from the same pool used in Experiment 1.

Design and procedure—All procedures were similar to Experiment 1, except that the items from each list were spoken by alternating male and female voices (the mixed-voice condition). Half of the lists started with a male voice and half with a female voice. As in the previous experiment, lists were rotated through the three study conditions (study + recall, study + math, and nonstudied), with a roughly equal number of subjects assigned to each of the counterbalancing conditions. Again, recognition test items were presented visually, and subjects made “male-old”, “female-old”, “old” (unsure of voice), or “new” judgements.

Results

Recall and recognition—As in the previous experiment, relatively high levels of immediate false recall were obtained. True recall of list items was .57 (averaged across all 15 items), and false recall of critical items was .39, which was slightly lower than recall of items from middle serial positions (6–10, mean = .43). In general, the overall levels of true and false recall were somewhat greater in this experiment (.57 and .39, respectively) than in the last experiment (.41 and .29). Thus, contrary to the idea that varying the voice within a list would enhance perceptual or item-specific processing and thereby reduce false recall, false recall was not diminished in this experiment (see also Hicks & Marsh, 1999).

The recognition results are presented in Table 2, with a similar format as used in the previous experiment. Note that performance for critical lures could not be broken down into those associated with a male or female voice at study, because their corresponding lists had been presented with both voices. As in the previous experiment, levels of false recognition of the critical lures were quite high relative to true recognition. There was also an effect of prior testing, such that recognition of list items and critical lures was greater following recall than following math. This effect was not found in the previous experiment, and may have been due to the greater levels of overall recall in this experiment. A 3 (counterbalancing order) × 2 (item type: studied vs critical lure) × 2 (task type: recall vs math) ANOVA on the overall recognition data confirmed that there was a main effect of item type, $F(1,29) = 5.22$, and task type, $F(1,29) = 5.61$. Critical lures were recognised more often (mean = .80, collapsing across other conditions) than were studied items (.74), and both true and false recognition were greater following recall (mean = .79, collapsing across item type) than math (.74). There was no effect of order, $F(2,29) < 1$, and none of the interactions was significant. The false alarm rate to unrelated lures was .19.

Source attributions—The main question of interest was whether subjects would persist in attributing critical lures to a particular voice even when their corresponding list had been

presented in two different voices. The answer was “yes”. As in the previous experiment, voice attributions occurred quite frequently for both list items and critical lures, and list item attributions were fairly accurate. Attributions for critical lures were evenly distributed between the two sources, as was the case for lures for nonstudied lists (replicating Hicks & Marsh, 2001, Experiment 2B). A 3 (counterbalancing order) × 2 (studied vs critical lure) × 2 (recall vs math) ANOVA on the overall attribution data revealed a significant difference between voice attributions for list items (.48) and critical lures (.40), $F(1,29) = 8.82$. This indicates that subjects were somewhat more likely to make a voice attribution for studied items than for critical lures, although both attribution rates were higher than those made to lures from nonstudied lists (.32). There was also a marginally significant effect of having previously recalled a list, $F(1,29) = 3.84$, $p = .06$. This indicates that subjects were more likely to make a voice attribution (for either list items or critical lures) following recall (mean = .46, collapsing across item type) than following math (.41), consistent with the overall recognition data. There was no effect of order, and there were no significant interactions.

A separate ANOVA was conducted on the variables of task (recall vs math), study voice (male vs female), and voice attribution (consistent vs inconsistent) for correct recognition of the studied items. The analysis revealed a significant effect of task $F(1,30) = 7.30$, again indicating that subjects were more likely to make a voice attribution following recall than math. There was an effect of voice attribution $F(1,30) = 14.07$, indicating that correct attributions (mean = .26) were more likely than incorrect attributions (.09). There was no significant effect of study voice (male or female), $F(1,30) = 2.91$, $p = .10$. The only significant interaction was that of task × attribution $F(1,30) = 14.07$. This interaction suggests that correct voice attribution (i.e., consistent > inconsistent) occurred more frequently for lists associated with prior recall.

Pure vs mixed presentation—We now directly compare the attribution results of Experiment 1 and Experiment 2, which differed in the fact that the voice manipulation was between-lists in Experiment 1 (pure-voice) and within-list in Experiment 2 (mixed-voice). To make the comparison easier, we constructed Table 3 by collapsing over the recall/arithmetic variable for the two experiments. Statistics are not needed to make the points that can be gleaned by eye from Table 3, which tell an interesting story. Consider first the studied items at the top of the table. The pure-mixed manipulation can be seen to have little effect on overall recognition, as was the case with recall: Correct recognition was .72 for pure lists and .74 for mixed lists. Similarly, this variable did not much affect correct voice attributions. We had expected that when the lists were presented in one voice, subjects might be more accurate in attributing voice on the final recognition test than when the lists were mixed (as in Payne et al., 1996, and Mather et al., 1997). Actually, if anything, the difference between conditions was in the opposite direction from that expected: Subjects made a correct voice attribution for .26 items in the mixed-voice lists in Experiment 2, but only .20 for the pure-voice lists in Experiment 1. The fact that these voice attributions were not diminished by mixed-voice presentation indicates that, at least with our procedures, subjects were relying on item-specific recollection to make these correct attributions (which should not have been affected by the list structure manipulation).

An even more striking finding was that a similar pattern of results was found for critical lures. On overall recognition, critical lures were falsely recognised at least as often following mixed-voice presentation (.80) as following pure-voice presentation (.71). As was the case with the false recall results, these data disconfirm the notion that providing multiple voices at study would promote item-specific processing (and hence reduce false remembering). Turning to the attribution data, we thought that critical lures would surely be attributed to a source more often when their corresponding list had been associated with a

single source. In reality, subjects were just as likely to attribute the critical lures to a source under mixed-list conditions (.40) as under pure-list conditions (.38), and if anything the effect was in the opposite direction.

This pattern sheds new light onto the attribution processes underlying the DRM memory illusion. When lists were presented in a single voice (Experiment 1), subjects attributed the critical lure to the voice of their corresponding list. This suggests that subjects had mistakenly inferred that the critical lure had occurred in the voice that they knew had spoken the corresponding list. However, the findings of Experiment 2 suggest that such inferences are not necessary for such misattributions: Even when lists were presented in mixed voices, so that critical lures could not have been associated with a voice via inferential processes, source misattributions remained just as high.

EXPERIMENT 3

In the previous two experiments, critical lures were equally likely to have been attributed to a voice following the pure-voice and mixed-voice conditions. In the former case, they were attributed to the voice of their corresponding list, and in the latter case, they were equally attributed to either voice. Experiments 3 and 4 were designed as the companions to Experiments 1 and 2, except that auditory tests were used instead of visual tests. Previous research has demonstrated that matching presentation modalities (i.e., auditory and visual) at study and test can influence source-monitoring processes (e.g., Gallo et al., 2001). In the present two experiments, the main question was how voice information provided at both study and test would influence source attributions. In Experiment 3, pure-voice presentation was used at study, and in Experiment 4, mixed-voice presentation was used.

The primary question in the present experiment (pure-voice lists) was whether attributions for studied items and critical lures would correspond to the voice of their corresponding list (as in the pure-list conditions of Experiment 1). For studied items, providing voice information at test was predicted to facilitate accurate source attributions, because the processing of items that are presented in the same voice at study and test should be particularly fluent (e.g., more familiar or easier to remember). Of greater interest was whether voice information at test would influence source judgements for critical lures. Because these items were never studied, voice information provided at test should be irrelevant to source judgements. Instead, subjects might simply infer that the critical lure had been presented in the voice of their corresponding list (as in the pure-list conditions of Experiment 1). However, if these judgements are based on a relatively automatic attribution process that is sensitive to perceptual fluency (cf. Jacoby, Kelly, & Dywan, 1989), then presenting test items in a particular voice might influence such an attribution process and increase false recognition when the test item is presented in the voice in which the list was presented.

Method

Subjects—A total of 32 Indiana University undergraduates were drawn from the same pool used in the previous experiments. The data from two of the subjects were omitted from the final analysis because they mistakenly recalled all of the lists (instead of half), yielding 10 subjects in each counterbalancing condition.

Design and procedure—The study phase was identical to that of Experiment 1, with all of the items from each list presented in a male or female voice (the pure-voice condition). The novel manipulation in this experiment was that test items were also presented auditorily (over headphones). Test items were counterbalanced so that half were presented in the male voice and half in the female voice. For the test items that had been studied, half were in the

same voice as was used at study and half were in the other voice. Similarly, for the critical lures, half were presented in the same voice as their corresponding study list and half in the other voice.

The test phase was similar to that of the previous experiments, with only slight modifications to accommodate auditory presentation at test. Subjects were again instructed to determine whether an item had been presented during the initial study phase, and if so, whether they could remember the voice that had early presented the item. If the subject remembered that the study item had been presented in the same voice as the test item, then they were to push the button labelled “old-same”. If the test item had been in a different voice from the study item, then they were to push a button labelled “old-different”. If the subject remembered that the item had been presented during the initial study phase but did not know the voice in which the item had been presented, then they were to push a button simply labelled “old”. Finally, if the subjects determined that an item had not been presented during the initial study phase, they were required to press a button labelled “new”.

Results

Recall and recognition—The immediate recall results were similar to those of the previous two experiments. True recall of list items was .46 (averaged across all 15 items). False recall of critical items was .37, which was about the same as recall of items from middle serial positions (6–10, mean = .36).

The recognition results are presented in Table 4. Even though different voices were used at test, which might have been thought to elicit greater item-specific or perceptual processing (and thereby reduce false recognition), robust levels of false recognition were obtained. For instance, false recognition in this experiment (.76) was at least as great as that in Experiment 1 (.71), which used similar procedures with the exception of a visual test and slightly different response format. As in the previous experiment, a 3 (counterbalancing order) × 2 (item type: studied vs critical lure) × 2 (task type: recall vs math) ANOVA was conducted on the overall recognition data. There was no effect of item type ($F(1,27) = 2.26$, indicating that the false alarm rate for the critical lures (.76) approximated the hit rates for the studied items (.72). As in the previous experiment, a main effect of task type was observed $F(1,27) = 9.70$, indicating that hit and critical false alarm rates were higher following recall of lists (mean = .77, collapsed across item types) than following math (.71). The false alarm rate to unrelated lures was .10.

In general, the effects of study-to-test voice changes on true recognition were in the predicted direction (same > different), but were quite small. Recognition of list items presented in the same voice was .73, whereas recognition of list items in the different voice was .70. In more extensive investigations of voice effects on recognition memory, we have found similar results (Pilotti, Bergman, Gallo, Sommers, & Roediger, 2000a; Pilotti, Gallo, & Roediger, 2000b; see also Palmeri et al., 1993). There was also little effect of study-to-test voice changes (with respect to the voice of their corresponding study list) on false recognition of critical lures (means = .75 and .76, respectively).

Source attributions—In general, list items and critical lures were often attributed to a voice, and these attributions did not appear to be consistently affected by prior recall. A 2 (item type: studied vs critical lure) × 2 (task type: recall vs math) ANOVA on the overall attribution data revealed no significant differences. This confirms that subjects were just as likely to make a voice attribution to critical lures (.48) as to studied items (.52). A similar pattern was observed in Experiment 1, which also used pure-voice presentation at study. The failure to find an effect of task type indicates that attributions following recall (.51) were

similar to those attributions following math (.49). It should be noted, though, that the effect was in the predicted direction (recall > math) for list items (.55 vs 49).

Of greater interest is the effect of study-to-test voice changes on source attributions. A 2 (item type: studied vs critical lure) × 2 (test voice: same vs different) × 2 (voice attribution: consistent vs inconsistent) ANOVA revealed a main effect of test voice $F(1,29) = 6.80$, indicating that subjects were more likely to attribute a voice to an item if it was presented in the same voice at study and test (mean = .39, collapsed across item type and voice attribution) than if it occurred in a different voice (.34). There was also a main effect of voice attribution $F(1,29) = 31.18$, indicating that consistent attributions (mean = .24, collapsed across item type and test voice) were more likely than inconsistent attributions (.12). That is, when people made voice attributions, they tended to attribute the test items (studied and critical) as having been presented in the voice that spoke the corresponding list. Finally, the item type × attribution interaction was marginally significant, $F(1,29) = 4.06$, $p = .053$. This indicates that the attribution effect (consistent > inconsistent) was somewhat greater for list items than for critical lures. Nevertheless, as can be seen from the table, the expected pattern (i.e., consistent attributions > inconsistent attributions) was seen for both list items and critical lures in almost all of the conditions, whereas attributions for unrelated lures were low and equally distributed across the two voices. The general point to take from these results is that, with a pure-list manipulation, if subjects made a voice attribution they were very likely to attribute both list items and critical lures to the source of their corresponding list, regardless of the test voice.

EXPERIMENT 4

Experiment 4 was similar to Experiment 3, except that mixed-voice conditions were used at study. We expected that such a manipulation should not much affect correct voice attributions to studied items because these attributions were quite accurate in Experiment 2, which also used mixed-voice conditions. In contrast, using mixed presentation at study might affect voice attributions to critical lures. In Experiment 2, attributions for critical lures were evenly distributed across voices. Because each list was associated with both voices, the corresponding critical lure was just as likely to have been attributed to each voice. However, a visual test was used in that experiment, whereas an auditory test was used in the present experiment. If voice attributions for critical lures are based solely on information provided during study, then we would expect a similar outcome in the present experiment (critical lure attributions would be equally distributed across voices). In contrast, if voice attributions can also be influenced by voice information provided at test, then we would expect that attributions for critical lures would be biased towards the voice that had presented the item during the test.

Method

Subjects—*A* total of 44 Indiana University undergraduates were drawn from the same pool used in the previous experiments. The data from four subjects were omitted from the final analysis because they failed to follow instructions by recalling items on every list. As in the previous experiments, a roughly equal number of subjects was assigned to each of the three counterbalancing conditions.

Design and procedure—The study phase conditions were the same as those of Experiment 2, in which the items from each list were spoken by an alternating male and female voice (the mixed-voice condition). The test phase conditions were the same as those of Experiment 3, in which an auditory test was used (the auditory test condition). As in that experiment, subjects were asked to make old/new and voice judgements (same, different, or don't know) by pressing keys labelled "old-same", "old-different", "old", or "new".

Results

Recall and recognition—Recall levels were quite similar to those of the previous experiment. True recall of list items was .49 (averaged across all 15 items), and false recall of critical items was about the same as recall of items from the five middle serial positions (.38 and .37, respectively). That false recall probabilities were roughly the same in this experiment (.38) as in the previous experiment (.37) replicates a similar pattern found across Experiments 1 and 2: Immediate false recall was not diminished by presenting each list with two voices (as opposed to a single voice). True recall was also quite similar between this experiment (.49) and the last (.46).

The recognition results are presented in Table 5. These results resemble those of the previous experiment, in that false recognition was very strong and seemed undiminished by having an auditory test. For instance, overall false recognition in this experiment (.78) approximated that in Experiment 2 (.80), which was similar except that a visual test was used. A 3 (counterbalancing order) × 2 (item type: studied vs critical lure) × 2 (task type: recall vs math) ANOVA on the present data revealed that there was no effect of item type, as false recognition of critical lures was equal or even greater to recognition of list items (means = .78 and .70, respectively). There was a significant effect of initial task, $F(1,37) = 5.80$, indicating that hits and critical false alarms were greater following recall (mean = .78, collapsing across item type) than following math (.73). This effect replicates that of the previous two experiments, and further suggests that the failure to find such an effect in Experiment 1 was probably due to the relatively lower levels of recall in that experiment. There was no effect of order $F(2,37) < 1$. The false alarm rate to unrelated lures was .13.

Source attributions—As in the previous experiment, we first analyse overall levels of voice attributions, and then turn to the effects of study and test voices on these attributions. The voice attribution data were analysed using a 3 (counterbalancing order) × 2 (item type: studied vs critical lure) × 2 (task type: recall vs math) ANOVA. We observed a significant difference in the rates of voice attribution between the studied items (.47) and critical lures (.40), $F(1,37) = 4.26$. This difference replicates that found in Experiment 2, which also used mixed-voice presentation at study, and these findings stand in contrast to those of Experiments 1 and 3 (each of which used pure-voice presentation and showed similar levels of attributions for list items and critical lures). Of course, subjects still made frequent voice attributions to critical lures (relative to unrelated lures) even under mixed-voice conditions. We did not observe an overall effect of task type, indicating that subjects made similar levels of voice attributions whether or not the corresponding list had been recalled. There was no significant effect of order $F(2,37) < 1$.

We turn next to the effect of study-to-test voice changes on voice attributions. In the previous experiment (pure-voice study), subjects attributed list items and critical items to the voice of their respective list, regardless of test voice. A quite different pattern was observed in this experiment, which used mixed voices at study. As can be seen in Table 5, list items and critical lures were attributed to the voice used at test, regardless of the voice used at study. To confirm this observation, a 2 × 2 ANOVA was conducted on the variables of item type (studied vs critical lure) and voice attribution (same vs different). This analysis revealed an overall effect of voice attribution, $F(1,39) = 35.36$, indicating that if subjects did attribute a voice to the recognised item, they were likely to say the item had been presented in the same voice at study and test, even if the subject had been presented with a different voice at study and test.

This response bias suggests that, under these conditions, voice attributions were the result of some general feeling of familiarity arising from the test item and were not necessarily the result of true recollection of an event. This outcome makes sense for the critical lures, which

were not associated with a study voice (because their corresponding lists were presented in both voices). In the present experiment, providing voice information at test tipped the scales in the direction of that voice. We view this outcome as an illusion of perceptual fluency (Jacoby et al., 1989): these lures were so familiar that they were readily processed at test, leading to the erroneous belief that they had been presented in that same voice during study.

Perhaps more surprising is the finding that list items also followed this same pattern. In Experiment 2, which also used mixed voices at study but had a visual test, subjects were fairly accurate at remembering the voice used to present list items during study. We had expected that the procedures used in the present experiment (using an auditory test that permitted matching of voice at study and test) would elicit more accurate voice judgements than had been observed in that experiment. In contrast, studied items were prone to the same illusion of perceptual fluency as were the critical lures. Apparently, even though subjects should have been able to recollect voice information for the studied items, the use of a different voice at test overrode this tendency. It is unclear why this pattern should have occurred for studied items, but one explanation is that using mixed voices at study and different voices at test simply made the task of recollecting the appropriate study voice too confusing or difficult for subjects. Instead, their responses were driven by the more automatic perceptual fluency that was caused by auditory test presentation (cf. Jacoby et al., 1989).

Pure vs mixed presentation—We now consider the comparison between pure-voice study conditions (Experiment 3) and mixed-voice conditions (Experiment 4) with auditory tests. These data are presented in Table 6, using a similar format as Table 3. There are two main points to be taken from these data. On the overall recognition data, it can be seen that studying the lists under pure-voice or mixed-voice conditions had little effect on hits (.71 and .73, respectively) or on critical false alarms (.75 and .78, respectively). As was the case when Experiments 1 and 2 were compared, presenting the study lists in mixed voices did not reduce overall levels of false recognition compared to pure-voice presentation. Coupled with similar findings in recall (reported here and by Hicks & Marsh, 1999), these data bolster the claim that using mixed voices does not appear to encourage item-specific processing that could reduce false remembering.

The second point to take from these data is that overall attributions for list items and critical lures tended to occur more often following pure-voice presentation (.52 and .48, respectively) than following mixed-voice presentation (.47 and .41, respectively). This pattern is different from that of Experiments 1 and 2, in which attributions were equally likely to have been made following pure-and mixed-voice conditions. This discrepancy may be due to our hypothesis that subjects in Experiment 4 (mixed voices, auditory test) were more likely to make their voice attributions (for both list items and critical lures) on the basis of test fluency. As previously discussed, the pure-voice conditions in Experiment 3 apparently overrode the tendency for test voice to influence attributions. If pure-voice information from study was indeed weighted more heavily in the attribution process than was test information, then one might also expect that attributions following pure-voice conditions would occur more frequently. Of course, this is only a tentative hypothesis, and further work is needed. The more central point to be taken from these results is that voice information provided at test did influence the source attribution process, even though such information was irrelevant (at least for critical lures).

GENERAL DISCUSSION

The primary goal of these experiments was to investigate erroneous voice attributions given to falsely recognised critical lures. These results can be readily summarised. On a visual test

that provided no voice-specific information, subjects attributed the critical lure to the voice information associated with their list (Experiment 1, pure voices at study) or, if such information was not available (Experiment 2, mixed voices at study), distributed their attributions equally across voices. Importantly, voice attributions were just as high under mixed-voice conditions as under pure-voice conditions for false alarms as well as for hits. This indicates that these misattributions are not wholly dependent on knowledge of the list structure (i.e., imply inferring that the critical lure had been studied in the voice of its corresponding list), although such inferences may play a large role (as discussed below). Instead, the misattributions appear to involve the illusory recollection of perceptual characteristics, such as those that may have been imagined during study, and/or attributed to the lure at retrieval.

Voice information provided at test did not influence attributions for critical lures when each list was associated with a particular voice (Experiment 3, pure voices). In contrast, when voice information was provided only at test, this information drove attributions for critical lures (Experiment 4, mixed voices). Subjects showed a bias to say that both list items and critical lures had occurred in the voice in which the test item was presented. This pattern suggests that when study information was available (the pure-voice condition), it took precedence over test information in driving source misattributions. Apparently, subjects were able to recollect the voice of the appropriate list, and (reasonably) infer that the critical item had also been presented in that voice. In contrast, when no such study information is available, attribution processes are driven by voice information provided at test (in the auditory test conditions of Experiment 4), and in the absence of such information, attributions are distributed equally across voices (in the visual test conditions of Experiment 2). Overall, these data suggest that the attribution process is a function of the subject's knowledge of the structure of the memory task, which in turn determines which of these sources (study or test voice) will contribute more heavily to the attribution.

A second issue examined in the present experiments was whether recall (of studied and nonstudied items) affected later recognition. In general, the answer is yes, confirming the results of Roediger and McDermott (1995) and others. This was the case in all four experiments for studied items and in three experiments for critical lures. Averaging over all four experiments, hits for list items were .76 after recall of the relevant list and .70 after math problems. The comparable false alarm rates for critical lures were equivalent at .76, because of the large reversal in Experiment 1 (see Table 1). It should also be noted that overall recall in Experiment 1 (where no positive effect was observed) was lower compared to recall in the other three experiments. This low level of recall might account for the lack of an effect on recognition. Some effects of prior recall on voice attributions (for both studied items and critical lures) were also apparent, although these were only significant (i.e., attributions following recall > attributions following math) in Experiment 2. In summary, prior recall of a list seems to have a relatively consistent effect on later recognition, although in a review of the evidence Roediger et al. (1998) noted that the effect is (surprisingly) more consistent in recognition of studied items than in recognition of critical lures, a pattern observed in these experiments, too. The level of recall of both list items and critical lures probably modulates this effect; if conditions were arranged so that either list recall or false recall were higher, then the effect on the later recognition test may well be larger.

Many experiments have shown that false recognition in the DRM paradigm with 15-item lists is approximately equivalent to verifical recognition, at least when the recognition test is given after the presentation of many lists. The four experiments here showed the same pattern: Collapsing across the recall/math variable, the hit rate for list items was .73 and the false alarm rate for critical items was .76. An analysis across the four experiments reveals that this difference, though small, was statistically significant, given our great power to

detect an effect, $F(1,536) = 4.80$. Under those conditions developed by Roediger and McDermott (1995), illusory recognition is very robust, with false recognition of the critical lure about equal to, or even greater than, veridical recognition of items in the list. Of course, it is not the case that illusory recall or recognition probabilities are equivalent to recall or recognition of the critical items when they are presented. McDermott (1997) showed that recall of the critical item was more probable when it had been presented than when it had not been presented, and McDermott and Roediger (1998) showed the same outcome in recognition (see also Miller & Wolford, 1999).

The present experiments were specifically designed to provide us with further insight into phenomenological experience during false recognition. Overall, we found that listeners were quite likely to attribute a voice to an illusory memory of the critical items and this likelihood approximated that for studied items. Over all four experiments, subjects made a voice attribution to 49% of all studied items called “old” and to 42% of all critical items called “old”. Further, this pattern held up over manipulations at encoding (presenting lists purely in one voice or mixed voices within a list) and over manipulations at retrieval (testing either with visual or auditory recognition tests and, in the latter case, with the same or different voice used on the test). The fact that these changes in encoding and retrieval factors did not much affect the levels of false recognition or voice attribution attests to the power of the DRM memory illusion. A reasonable expectation—and one we entertained before conducting these experiments—is that subjects could be able to use differences in voice information (either when whole lists are presented in one or the other voice, or when test items are presented in the same or a different voice rather than being visually presented) as powerful retrieval cues to aid recognition. If so, we should have seen increases in hit rates for studied items when a matching of voice characteristics occurred between study and test, and we should have seen a decrease in false alarm rates when these qualities mismatched. However, subjects apparently could not use this information to reduce levels of false recognition in our experiments (nor to increase hit rates). Apparently, the manipulation of different sources needs to be drastic (i.e., the sources need to be more discriminable) in order to reduce false remembering (e.g., pictures vs words, Schacter, Israel, & Racine, 1999; read vs generate, Hicks & Marsh, 2001).

The most remarkable aspect of our data is the similarity in judgements attributed to critical lures in relation to list items. In all four experiments, subjects generally provided voice judgements for critical lures that matched the voice judgements given for list items. The same general pattern has been found in many other studies with judgements of confidence and remember/know judgements (Roediger & McDermott, 1995) and for judgements of modality (Gallo et al., 2001). These findings are inconsistent with some theories of false recognition. For example, Whittlesea and Williams’ (1998) discrepancy-attribution account of feelings of familiarity argues that “false claims of recognition can only be based on guessing and feelings of familiarity, whereas true claims can also be based on recall of the event itself” (p. 148). However, our data and those of others clearly show that false recognition can be accompanied by retrieval of specific information (even if illusory) as much as is true recognition. False recognition is not driven merely by vague feelings of familiarity but seemingly by specific recollections. Attributional theories of veridical and false recognition, which we have generally endorsed (Roediger & Gallo, in press; Roediger & McDermott, 1995), must provide a mechanism by which such specific attributions are made.

The same problem arises for fuzzy trace theory, at least as it was originally conceived. Reyna and Brainerd (1995) proposed that events are encoded in two forms, as specific (“verbatim”) traces and in terms of meaning as gist traces. In the original formulations, false recollections were driven by gist traces. In the case of critical items in the DRM paradigm,

the meaning extracted from the critical item would match the gist that had been encoded from the list, leading to high levels of false recognition. However, as in Whittlesea's attributional theory, the experience of false recognition was considered a nonspecific feeling of familiarity driven by gist (which by definition did not contain specific, verbatim information). However, false recognition in the DRM paradigm is accompanied by highly specific recollections, as all four of the current experiments show with regard to voice information. More recently, Brainerd, Wright, Reyna, and Mojardin (2001) have proposed the concept of phantom recollection to explain the specific attributions made to critical items in the DRM paradigm. The basic proposal is that when gist traces become quite strong, they take on characteristics of verbatim traces and thus permit specific attributions. This solution can then provide for specific attributions made during retrieval, although the idea of strong gist traces becoming like verbatim traces would seem to undercut the fundamental distinction on which fuzzy trace theory is based.

We have employed an activation/monitoring framework to explain illusory recollection in the DRM paradigm (Gallo et al., 2001; Gallo & Roediger, 2002; McDermott & Watson, 2001; Roediger et al., 2001). One critical assumption is that during encoding associations and inferences are made (either consciously or unconsciously) and those associations are encoded along with representations of presented material. DRM lists that are high in backward associative strength (that is, ones in which list items are most likely to produce the critical item on free association tests) lead to highest levels of false recall and false recognition (Deese, 1959; Roediger et al., 2001). We assume that in such lists many list items spark associations to the critical items; this convergence of associations leads to encoding of the critical word, even though it was not literally presented. Further, we assume that this encoding of the critical word leads to the specific attributions that are made during the test. Gallo and Roediger (2002) showed that backward associative strength was correlated with false *remember* judgements even when these judgements were conditionalised on false recognition. In the current studies we used lists that were generally high in backward associative strength, and therefore assume that the critical items were encoded along with the list. If so, then during the recognition test and the voice-attribution tests, subjects must monitor their recollections for distinguishing characteristics. Due to encoding of the illusory critical items during list presentation, this challenge is difficult to meet. Consequently, subjects frequently recognise critical items as having actually occurred and also attribute specific voice characteristics to these items.

The activation/monitoring framework therefore provides a reasonable general account of the phenomena under study, although many details remain to be worked out. One straightforward prediction is that if voice recognition experiments were conducted like the current ones, but with lists that varied in backward associative strength, voice attributions would vary directly with backward associative strength (just as Gallo & Roediger, 2002, found for remember judgements). This prediction and others await further research.

Acknowledgments

This research was supported by NIDCD Research Grant DC-00111-18 to David B. Pisoni at Indiana University. We thank Helena M. Saldaña for her experimental assistance in testing subjects and for help in analysing the results. Steve Lindsay provided helpful comments on the manuscript.

References

Brainerd CJ, Wright R, Reyna VF, Mojardin AH. Conjoint recognition and phantom recollection. Journal of Experimental Psychology: Learning, Memory, & Cognition. 2001; 27:307–327.

Deese J. On the prediction of occurrence of particular verbal intrusions in immediate recall. Journal of Experimental Psychology. 1959; 58:17–22. [PubMed: 13664879]

Fisher RP, Cuervo A. Memory for physical features of discourse as a function of their relevance. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1983; 9:130–138.

Gallo DA, McDermott KB, Percer JM, Roediger HL. Modality effects in false recall and false recognition. Journal of Experimental Psychology: Learning, Memory, & Cognition. 2001; 27:339–353.

Gallo DA, Roediger HL. Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. Journal of Memory and Language. 2002; 47:469–497.

Goldinger, SD. Research in Speech Perception Technical Report No 7. Bloomington, IN: Indiana University; 1992. Words and voices: Implicit and explicit memory for spoken words.

Hicks JL, Marsh RL. Attempts to reduce the incidence of false recall with source monitoring. Journal of Experimental Psychology, Learning, Memory, and Cognition. 1999; 25:1195–1209.

Hicks JL, Marsh RL. False recognition occurs more frequently during source identification than during old-new recognition. Journal of Experimental Psychology, Learning, Memory, and Cognition. 2001; 27:375–383.

Jacoby, LL.; Kelley, CM.; Dywan, J. Memory attributions. In: Roediger, HL., III; Craik, FIM., editors. Varieties of memory and consciousness: Essays in honour of Endel Tulving. Hillsdale, NJ: Lawrence Erlbaum Associates Inc; 1989. p. 391-422.

Jones TC, Roediger HL. The experiential basis of serial position effects. European Journal of Cognitive Psychology. 1995; 7:65–80.

Lampinen JM, Neuschatz JS, Payne DG. Memory illusions and consciousness: Examining the phenomenology of true and false memories. Current Psychology: Developmental, Learning, Personality, Social. 1998; 16:181–224.

Lampinen JM, Neuschatz JS, Payne DG. Source attributions and false memories: A test of the demand characteristics account. Psychonomic Bulletin & Review. 1999; 6:130–135. [PubMed: 12199307]

Lockhart R. The facilitation of recognition by recall. Journal of Verbal Learning and Verbal Behavior. 1975; 14:253–258.

Mather M, Henkel LA, Johnson MK. Evaluating characteristics of false memories: Remember/know judgments and Memory Characteristics Questionnaire compared. Memory & Cognition. 1997; 25:826–837.

McDermott KB. The persistence of false memories in list recall. Journal of Memory and Language. 1996; 35:212–230.

McDermott KB. Priming on perceptual implicit memory tests can be achieved through presentation of associates. Psychonomic Bulletin & Review. 1997; 4:582–586.

McDermott KB, Roediger HL. Robust false recognition of associates persists under conditions of explicit warnings and immediate testing. Journal of Memory and Language. 1998; 39:508–520.

McDermott KB, Watson JM. The rise and fall of false recall. Journal of Memory and Language. 2001; 45:160–176.

Miller MB, Wolford GL. The role of criterion shift in false memory. Psychological Review. 1999; 106:398–405.

Palmeri TJ, Goldinger SD, Pisoni DB. Episodic encoding of voice attributes and recognition memory for spoken words. Journal of Experimental Psychology: Learning, Memory & Cognition. 1993; 19:309–328.

Payne DG, Elie CJ, Blackwell JM, Neuschatz JS. Memory illusions: Recalling, recognizing and recollecting events that never occurred. Journal of Memory and Language. 1996; 35:261–285.

Pilotti M, Bergman ET, Gallo DA, Sommers M, Roediger HL III. Direct comparison of auditory implicit memory tests. Psychonomic Bulletin & Review. 2000a; 7:347–353. [PubMed: 10909144]

Pilotti M, Gallo DA, Roediger HL III. Effects of hearing words, imaging words, and reading on auditory implicit and explicit memory tests. Memory & Cognition. 2000b; 28:1406–1418.

Reyna VF, Brainerd CJ. Fuzzy-trace theory: An interim synthesis. Learning and Individual Differences. 1995; 7:1–75.

Roediger, HL.; Balota, DA.; Watson, JM. Spreading activation and the arousal of false memories. In: Roediger, HL.; Nairne, JS.; Neath, I.; Suprenant, AM., editors. The nature of remembering: Essays

in honor of Robert G Crowder. Washington, DC: American Psychological Association Press; 2001. p. 95-115.

Roediger, HL.; Gallo, DA. Associative memory illusions. In: Pohl, RF., editor. Cognitive illusions: Fallacies and biases in thinking, judgment and memory. Oxford: Oxford University Press; (in press)

Roediger HL, McDermott KB. Creating false memories: Remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1995; 21:803–814.

Roediger, HL.; McDermott, KB.; Goff, L. The paradoxical effects of repeated testing. In: Conway, MA., editor. False memories and recovered memories. Oxford: Oxford University Press; 1997. p. 118-149.

Roediger, HL.; McDermott, KB.; Robinson, KJ. The role of associative processes in creating false memories. In: Conway, MA.; Gathercole, SE.; Cornoldi, C., editors. Theories of memory II. Hove, UK: Psychology Press; 1998. p. 187-245.

Schacter DL, Israel L, Racine CA. Suppressing false recognition in younger and older adults: The distinctiveness heuristic. Journal of Memory & Language. 1999; 40:1–24.

Whittlesea B, Williams LD. Why do strangers feel familiar, but friends don't? A discrepancy-attribution account of feelings of familiarity. Acta Psychologica. 1998; 98:141–165. [PubMed: 9621828]

TABLE 1

Recognition and voice attribution data for list items and critical lures in Experiment 1 (pure-voice/visual test)

	Proportion of "old" responses				
Item type & condition	**Overall**	**Male**	**Female**	**Don't know**	**Proportion voice attribution**
List items					
Study + recall male	.73	.22	.09	.42	.42
Study + recall female	.75	.10	.22	.43	.42
Mean	.74			.43	.42
Study + math male	.70	.18	.07	.45	.36
Study + math female	.70	.06	.19	.45	.36
Mean	.70			.45	.36
Nonstudied lists	.16	.02	.02	.12	.25
Critical lures					
Study + recall male	.67	.20	.07	.40	.40
Study + recall female	.67	.09	.18	.40	.39
Mean	.67			.40	.40
Study + math male	.73	.20	.08	.45	.38
Study + math female	.78	.05	.21	.52	.33
Mean	.75			.49	.35
Nonstudied lists	.08	.01	.01	.06	.25

TABLE 2

Recognition and voice-attribution data for list items and critical lures in Experiment 2 (mixed-voice/visual test)

	Proportion of "old" responses				
Item type & condition	**Overall**	**Male**	**Female**	**Don't know**	**Proportion voice attribution**
List items					
Study + recall male	.82	.34	.08	.40	.51
Study + recall female	.72	.08	.28	.36	.50
Mean	.77			.38	.50
Study + math male	.78	.24	.10	.44	.44
Study + math female	.62	.10	.18	.34	.45
Mean	.70			.39	.44
Nonstudied lists	.19	.04	.03	.12	.37
Critical lures					
Study + recall	.81	.16	.18	.47	.42
Study + math	.78	.14	.15	.49	.37
Mean	.80	.15	.17	.48	.40
Nonstudied lists	.19	.02	.03	.14	.26

TABLE 3

Comparison of results from Experiments 1 (pure-voice study, visual test) and 2 (mixed-voice study, visual test), with data collapsed over the math/recall variable

	Proportion of "old" responses				
Item type& condition	**Overall**	**Male**	**Female**	**Don't know**	**Proportion voice attribution**
List items					
Pure					
Male	.72	.20	.08	.44	.39
Female	.72	.08	.20	.44	.39
Mixed					
Male	.80	.29	.09	.42	.47
Female	.67	.09	.23	.35	.48
Critical lures					
Pure					
Male	.70	.20	.08	.43	.39
Female	.72	.07	.20	.46	.36
Mixed	.80	.15	.17	.48	.40

Memory. Author manuscript; available in PMC 2012 September 06.

TABLE 4

Recognition and voice-attribution data for list items and critical lures in Experiment 3 (pure-voice/auditory test)

	Proportion of "old" responses				
Item type & condition	**Overall**	**Same**	**Different**	**Don't know**	**Proportion voice attribution**
List items					
Study + recall same	.76	.32	.10	.34	.55
Study + recall different	.73	.12	.28	.33	.55
Mean	.75			.34	.55
Study + math same	.69	.30	.07	.32	.54
Study + math different	.67	.14	.15	.38	.43
Mean	.68			.35	.49
Nonstudied lists	.11	.02	.02	.07	.36
Critical lures					
Study + recall same	.75	.17	.22	.36	.52
Study + recall different	.80	.05	.27	.48	.40
Mean	.78			.42	.46
Study + math same	.74	.21	.17	.36	.51
Study + math different	.72	.09	.25	.38	.47
Mean	.73			.37	.49
Nonstudied lists	.09	.01	.01	.07	.22

TABLE 5

Recognition and voice-attribution data for list items an critical lures in Experiment 4 (mixed-voice/auditory test)

	Proportion of "old" responses				
Item type & condition	**Overall**	**Same**	**Different**	**Don't know**	**Proportion voice attribution**
List items					
Study + recall same	.74	.27	.09	.38	.49
Study + recall different	.77	.23	.14	.40	.47
Mean	.76			.39	.48
Study + math same	.72	.24	.08	.40	.45
Study + math different	.69	.24	.08	.37	.46
Mean	.70			.38	.46
Nonstudied lists	.15	.03	.02	.10	.33
Critical lures					
Study + recall	.79	.21	.09	.49	.38
Study + math	.76	.23	.10	.43	.44
Mean	.78			.46	.41
Nonstudied lists	.10	.00	.02	.08	.20

TABLE 6

Comparison of results from Experiments 3 (pure-voice study, auditory test) and 4 (mixed-voice study, auditory test), with data collapsed over the math/recall variable

	Proportion of "old" responses				
Item type & condition	**Overall recognition**	**Same**	**Different**	**Don't know**	**Proportion voice attribution**
List items					
Pure					
Same	.72	.31	.08	.33	.54
Different	.70	.13	.22	.36	.49
Mixed					
Same	.73	.26	.08	.39	.47
Different	.73	.24	.11	.38	.46
Critical lures					
Pure					
Same	.74	.19	.20	.36	.52
Different	.76	.07	.26	.43	.44
Mixed	.78	.22	.10	.46	.41