Who Said What? An Event-Related Potential Investigation of Source and Item Memory

Ava J. Senkfor and Cyma Van Petten
University of Arizona

Event-related potentials (ERPs) were recorded during recognition tasks for spoken words alone (items) or for both words and the voice of the speaker (sources). Neither performance nor ERP measures suggested that voice information was retrieved automatically during the item-recognition task. In both tasks, correctly recognized old words elicited more positive ERPs than new words, beginning around 400 ms poststimulus onset. In the source task only, old words also elicited a focal prefrontal positivity beginning about 700 ms. The prefrontal task effect did not distinguish trials with accurate and inaccurate voice judgments and is interpreted as reflecting the search for voice information in memory. More posterior recording sites were sensitive to the successful recovery of voice or source information. The results indicate that word and voice information were retrieved hierarchically and distinguish retrieval attempt from retrieval success.

Everyone has had the experience of remembering a fact without being able to recall how it was learned. Remembering the source of one’s knowledge is not always important, but in some cases, it may be critical for one’s subsequent actions. If you own a cellular phone, it is important to remember whether the link between such phones and brain tumors was reported in the New England Journal of Medicine or the National Enquirer. Source memory can be broadly defined as memory for the context of a learning episode, encompassing perceptual features (e.g., the voice or face of a speaker who conveyed some information, visual or auditory modality of a word), spatial location, temporal sequence of events, and whether an event was imagined or actually took place (for a review, see Johnson, Hashtroudi, & Lindsay, 1993).

A few recent neuropsychological studies have suggested that source memory is dissociable from other forms of memory. In a group of patients with amnesia subsequent to brain damage, Schacter, Harbluk, and McLachlan (1984) found that the level of impairment for remembering trivia facts was uncorrelated with performance in remembering which experimenter conveyed the facts that were recalled. Shimamura and Squire (1987, 1991) reported similar findings in both amnesic patients and those with temporary memory impairment following electroconvulsive therapy. Janowsky, Shimamura, and Squire (1989) obtained a particularly clear dissociation between item and source memory in patients who were not generally amnesic. Patients with damage confined to the frontal lobe exhibited normal recall and recognition of trivia facts learned in the laboratory but a disproportionate number of errors of attributing their knowledge to an extraexperimental source. This demonstration that frontal cortex is critical for accurate source memory confirmed and extended a number of previous studies showing that patients with frontal lobe damage are impaired in tasks requiring memory for temporal context, including reconstruction or identification of the sequence of previously presented items, frequency judgments, and recency judgments (Ladavas, Umilta, & Provinciali, 1979; McAndrews & Milner, 1991; Milner, Corsi, & Leonard, 1991; Shimamura, Janowsky, & Squire, 1990; M. L. Smith & Milner, 1988).

Comparisons of item and source memory have also been conducted in neurologically intact individuals, with a wider variety of source tasks, not all of which involve temporal judgments. In particular, many studies of healthy elderly adults have required source judgments for perceptual features, including the voice that initially spoke a word, auditory versus visual word presentation, the color of line drawings, and the spatial location of objects (Kausler & Puckett, 1981; Lehman & Mellinger, 1984; Light, LaVoie, Valencia-Laver, Owens, & Mead, 1992; Park & Puglisi, 1985; Senkfor & Van Petten, 1996; Uttl & Graf, 1993; for general reviews, see Light, 1991; Spencer & Raz, 1995).

These studies have generally shown partial dissociations between item and source memory in that older adults tend to have lower performance than young adults in item-memory tasks and also make disproportionately more source errors. Postmortem studies, as well as more recent in vivo measurements with neuroimaging methods, have indicated that prefrontal cortex is more susceptible to age-related declines in neuron size and cortical volume than are many other brain areas.
areas (Coffey et al., 1992; Haug, 1985; Martin, Friston, Colebatch, & Frackowiak, 1991; Raz, Torres, Spencer, & Acker, 1993; Terry, DeTeresa, & Hansen, 1987). It has been suggested that the neurological and behavioral findings are causally related, such that aging impairs all forms of source
memory through its impact on the frontal lobe (Craik, Morris, Morris, & Loewen, 1990; Spencer & Raz, 1994).

One obvious objection to this conclusion is that source-
memory tasks are typically more difficult than item tasks, so
that the apparently specific deficit in elderly adults may
reflect a general memory failure rather than a dissociable
component of memory. A recent study by Glisky, Polster,
and Routheaux (1995) addressed this issue by harnessing
the variability found within the healthy elderly population
to obtain a double dissociation between item and source
memory. Participants were initially categorized by factor
scores summarizing performance on neuropsychological
tests thought to be sensitive to frontal versus medial
temporal lobe damage. Performance on a source-memory
task of judging whether or not a sentence was presented in
the same voice at study and at recognition differed between
participants with good versus poor scores on the “frontal”
battery. Recognition of the sentences themselves did not
differ between the high and “low frontal” groups. In con-
trast, the ability to discriminate studied versus unstudied
sentences differed between participants with good versus
poor scores on the “temporal” battery, but the temporal
score was unrelated to source-memory performance.

Although explicit source-memory paradigms have pro-
vided strong evidence that memory for voice information
can be dissociated from memory for verbal content, another
set of studies has suggested that voice information is
automatically stored and recovered. In an early study, Cole,
Coltheart, and Allard (1974) asked participants to make a
same–different judgment for pairs of letters presented at
short delays (up to 8 s). Reaction times were faster when the
two letters were presented by the same voice versus two
different voices. The authors concluded that letters presented
in the same voice benefited from a “physical code which
preserved ... the physical characteristics of the auditory
stimulus” (Cole et al., 1974, p. 5), whereas letters presented
in a different voice could not rely on the physical code and
mandated a time-consuming conversion into a “name code.”
Two other studies used continuous recognition paradigms
with longer delays. In both studies, participants were asked
to indicate if each word in a list was old or new, with words
repeated at lags ranging from 1 to 32 (Craik & Kirsner,
1974) or from 1 to 64 intervening words (Palmeri, Gold-
ingner, & Pisoni, 1993). Both studies obtained higher recog-
nition accuracies for words repeated in the same voice than in
different voices, across all lags. Same-voice words also
elicited faster reaction times in both studies. These studies
thus demonstrated persistence of voice information over
several minutes when it was not required by the assigned
recognition task, giving rise to the idea that word and voice
are stored together. Palmeri et al. thus concluded that “voice
is not a contextual aspect of a word; rather ... it is an
integral component of the stored memory representation
itself” (p. 325).

These studies present an interesting contrast to the
source-memory studies reviewed above: If voice informa-
tion is automatically recovered with words, why are judg-
ments about a word-plus-voice linkage always less accurate
than comparable measures for word memory alone (Craik &
Kirsner, 1974; Geiselman & Crawley, 1983; Glisky et al.,
1995; Palmeri et al., 1993)? The circumstances that drive
the encoding and retrieval of voice information have been a
matter of some controversy. Schacter and Church (1992)
have argued that the perceptual attribute of voice influences
primarily perceptually driven tasks such as word-stem
completion and that the evidence that voice manipulations
influence explicit memory for words is “not exactly over-
whelming” (p. 928).

Research to date has been convincing in demonstrating
that source-memory tasks are distinct from other memory
tasks but also has raised a number of questions. Under what
circumstances is voice information automatically stored and
recovered, and how does such automatic retrieval differ
from effortful source-monitoring tasks? If prefrontal cortex
is essential for linking an item with its context, is this part
of the brain involved during encoding, at retrieval, or both?
And finally, is “source” memory different from the standard
definition of “episodic” memory as consisting of “multifea-
ture representations in which numerous different kinds of
information—spatial, temporal, contextual, and so forth—
are bound together with the individual’s awareness of
personal experiences in subjective time” (Schacter & Tulv-
ing, 1994, p. 28)?

No single experiment will resolve these issues, but the
present study touches on each question by exploring whether
item- and source-memory tasks elicit differential brain
activity in participants with neither suspected nor diagnosed
neurological problems. The voice of the speaker was used as
the perceptual source for spoken words, followed by epi-
sodic recognition tasks for the words alone (items) or the
relationship between words and their voices (sources). In
both tasks, some words were presented in the same voice at
study and at recognition whereas other words changed
voice. Reaction times and event-related brain potentials
(ERPs) were recorded during the study and recognition
phases of both tasks. The data were analyzed in terms of
quantitative differences between tasks, which may reflect
the difficulty and time course of mnemonic processes (e.g.,
the amplitude and latency of memory effects), and qualitative
differences between tasks, which may suggest the involve-
mint of different neural circuits (e.g., the spatial distribution
of memory effects across the scalp). Next, we review the
sensitivity of ERPs to mnemonic processes.

**ERPs and Memory**

The neurophysiology of cognitive processing can be
investigated in humans by the noninvasive recording of ERP
activity from the scalp. ERPs are small voltage fluctuations
in the electroencephalogram (EEG) that are time locked to
sensory, motor, or cognitive events; these potentials are the
scalp reflections of summed synaptic activity (for reviews,
see Coles, Gratton, & Fabiani, 1990; Hillyard & Picton,
A number of studies have shown that words remembered during a recognition test elicit a more positive waveform than correctly identified new words or unrecognized old words, beginning at 300-400 ms poststimulus onset (Berman, Friedman, & Cramer, 1991; Friedman, 1990; Rugg & Nagy, 1989; M. E. Smith & Halgren, 1989; Van Petten & Senkowski, 1996). In individuals with unimpaired memory, a similar late positivity is elicited by items that are repeated in a variety of nonmemory tasks such as lexical decision, reading for comprehension, or semantic judgments (Bentin & Peled, 1990; Besson, Kutus, & Van Petten, 1992; Rugg, 1990; Van Petten, Kutus, Kluender, Mitchner, & McIsaac, 1991; Van Petten & Senkowski, 1996). A number of studies indicate the late positive memory effect reflects conscious recollective processes rather than perceptual priming (Paller & Kutas, 1996). The ERP literature on memory for repeated words in nonmemory tasks and old words in nonmemory tasks are unrecalled or unrecognized in subsequent tests (Besson et al., 1995; Paller & Gross, in press; Rugg & Doyle, 1994; Van Petten & Senkowski, 1996). The ERP literature on memory for printed words and pictures is fairly extensive, but only a handful of studies have used spoken words. To date, auditory studies have yielded results similar to the visual modality: Repeated words in nonmemory tasks and old words in recognition tasks elicit a larger late positivity than new words (Ferlazzo, Conte, & Gentilomo, 1993; Rugg, Doyle, & Melan, 1993; Rugg, Doyle, & Wells, 1995).

ERPs recorded during the study phase of study–test experiments are also sensitive to some aspects of initial encoding. In particular, studied words that are later remembered typically elicit a more positive ERP than words that are unrecalled or unrecognized in subsequent tests (Besson & Kutus, 1993; Fabiani, Karis, & Donchin, 1986, 1990; Fox, Michie, & Coltheart, 1990; Neville, Kutus, Chesney, & Schmidt, 1986; Paller, 1990; Paller & Kutus, 1992; Paller, Kutus, & Mayes, 1987; Sandquist, Rohraugh, Syndulko, & Lindsley, 1980; Van Petten & Senkowski, 1996; Wyatt, Allen, Bootzin, & Anthony, 1995). The comparison of study-phase ERPs contingent on subsequent memory is typically referred to as a difference-according-to-memory (Dm) analysis. The magnitude, or even existence, of a Dm effect has been variable across studies depending on the nature of the study-phase task and the mnemonic strategies applied by the participants (Fabiani et al., 1986, 1990; Van Petten & Senkowski, 1996). So, although the Dm effect necessarily bears some relationship to learning, the exact cognitive process it may reflect is an unresolved issue. In this study, we analyzed the study-phase data for signs of differential encoding of words and voices later recognized versus forgotten.

**Experiment 1**

**Method**

**Participants**

Twelve men and 12 women ranging in age from 19 to 30 years ($M = 24$ years) participated as paid volunteers. All were native English speakers with normal hearing. Twenty-two participants were right-handed (Oldfield, 1971); 7 of these reported having a left-handed family member (parent or sibling). Two participants were left-handed, 1 of whom reported having a left-handed family member. Nine additional individuals began the experiment but did not complete it because of excessive eye movement artifacts in the first session, discomfort with the electrode cap, or failure to return for all four sessions.

**Stimuli**

A total of 1,088 concrete nouns were recorded in both a male and a female voice. The mean duration was 621 ms (range = 276–1,326 ms). Mean frequency of usage was 17 per million (range = 0–769; sum of all regularly inflected forms according to Francis & Kucera, 1982). The words were initially recorded onto analog tapes (ElectroVoice RE16 microphone, Sony TC-WR87ES tape deck), then low-pass filtered at 9 KHz (Butterworth 6-pole), and digitized at a sampling rate of 20 KHz by an analog–digital card (DT 2821) under the control of a personal computer. Each digital file was edited to ensure good synchronization between the beginning and the ending of the file and between the acoustic onset and offset of the word. The intensity of all stimuli was scaled to equate the maximum peak-to-peak values across words. During the experiment, the stimuli were played back through the same analog–digital card and filter to an audio monitor (Yamaha MS202). Volume levels were set to a comfortable listening level for each participant.

**Electrophysiological Methods**

EEG's were recorded from tin electrodes secured in a commercially available cap (Electrocap International). Recording sites included midline frontal (Fz), central (Cz), and parietal (Pz) together with lateral occipital (O1, O2) scalp sites as defined by the 10–20 system (Jasper, 1958). Eight nonstandard lateral sites were included and were approximately over prefrontal cortex (PFL, PR), Broca's area and its right-hemisphere homologue (BL, BR), auditory cortex (TL, TR), and Wernicke's area and its right-hemisphere homologue (WL, WR). PFL and PR were placed above the nasion (5% of the nasion-to-inion distance) and 10% of the interaural distance laterally, BL and BR were midway between P7 and T3 and P8 and T4, respectively, in the 10–20 system. TL and TR were located 33% of the interaural distance lateral to Cz. WL and WR were located 30% of the interaural distance lateral to Cz. 12.5% of the nasion-to-inion distance posterior to Cz. An additional active electrode was placed over the right mastoid. Vertical eye movements and blinks were monitored by an electrode placed below the right eye (Le). The scalp sites and vertical electrooculogram electrode were referenced to the left mastoid during recording but were rereferenced off-line to an average of the left and right mastoids. Horizontal eye movements were monitored by a right-to-left bipolar montage at the external canthi of the two eyes. The EEG was amplified by a Grass Model 12 polygraph with half-amplitude cutoffs of 0.01 and 100 Hz, digitized on-line at a sampling rate of 250 Hz, and stored on optical disk along with stimulus codes for subsequent averaging. Trials with artifacts due to eye movement or amplifier saturation were rejected prior to averaging.

---

1 The names of electrode sites are descriptive to convey a rough idea of their location on the scalp. No electrode name should be taken to imply that it detects electrical activity only in the immediately underlying cortex.
Procedure

The experiment was conducted across four sessions lasting about 2.5 hr each. The first two sessions consisted of item study and recognition tests; the second two sessions consisted of source study and recognition tests. The item-memory sessions were always conducted first, before participants were informed about the source-memory aspects of the study. During the item sessions, participants studied a set of 136 concrete nouns, half in the male voice and half in the female voice (randomly intermixed), presented at a rate of 1 every 3.5 s. During the study phase, participants were instructed to form a mental image of each object and decide if it was larger or smaller than a 10-in. × 10-in. (25.4 cm × 25.4 cm) square located in front of them. Keypresses with the right and left index fingers were used to indicate “larger” and “smaller,” respectively; the mapping between hand and response was counterbalanced between sessions. The recognition test included all the studied words (136), half in the same voice as study and half in the other voice. Presentation rate was 1 word every 5 s. These were intermixed with an equal number of new words equally divided between the two voices. Participants made keypresses with the right and left index fingers to indicate whether the words were old or new, respectively. Participants were instructed to balance speed and accuracy. Response hands were counterbalanced between sessions. Before each item study list, participants received a practice set of 25 words for large-small judgments and were informed about the upcoming recognition test for words, but they received no instructions concerning the speakers’ voices. No mention was made of the voice-change manipulation for old words. A few participants mentioned the voice change from study to recognition and were told it was not relevant to the recognition task.

The third and fourth sessions used the same study task as the first two sessions, except that participants were informed that the speakers’ voices would be included in the recognition test. The recognition phase was also similar in that each list included all of the old words (half in the same voice as study and half in the other voice) mixed with an equal number of new words. Participants used the index and second fingers of one hand to indicate “old, same voice” and “old, different voice” and the index finger of the other hand to indicate “new.” Pilot work indicated that the source-recognition task was much more difficult than word recognition alone. Study lists of 17 words were used in the source-memory sessions to ensure reasonably high levels of voice-recognition accuracy. A practice study and recognition test was administered at the beginning of each session. Eight study-recognition blocks took place in each session, so that the total number of words was the same as in the item sessions (136 studied words and 136 new words).

Across all four sessions, there were eight stimulus categories of 136 old (studied) words formed by crossing the task (item or source recognition) and the voice-change manipulation (same or different voice as study for old words, and male or female voice for old and new words). Sixteen different stimulus lists were constructed to counterbalance task, old–new words, male–female voice, and same–different voice at recognition. Across 24 participants, each list was used once; the final 8 participants received lists fully counterbalanced for task, old–new words, and same–different voice but incompletely for male versus female voice.

Results

The results are presented in six sections describing (a) behavioral performance in the item- and source-recognition tasks, (b) ERPs elicited by studied versus unstudied words, (c) comparisons between the two recognition tasks, (d) analyses of session effects, (e) ERPs contingent on the success or failure of the source judgment, and (f) ERPs elicited during the study phase.

Recognition Performance

Accuracies and reaction times for both recognition tasks are shown in Table 1. Overall accuracy (old and new words, hits plus correct rejections) for the item-recognition task averaged 89%. For old words, neither hit rate nor reaction time was influenced by the voice-change manipulation (Fs < 1).

In the source-recognition task, responses to old words were classified according to two levels of accuracy depending on whether both the word and the voice were correctly categorized or if only the word was recognized. Hit–hit responses consisted of correct “old, same voice” and “old, different voice” judgments. Hit–miss trials included old words judged as old but incorrectly assigned to the same- or different-voice categories. The sum of the hit–hit and hit–miss categories provides a measure of word recognition during the source task when combined with correct responses to new words. This averaged 96%, a higher proportion than in the item-recognition task using longer study lists, F(1, 23) = 65.70, p < .0001, MSE = 16.8. For old words, there was a small (1.5%) accuracy advantage for those presented in the same voice during study and recogni-

Table 1

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Accuracy (%)</th>
<th>Reaction time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SE</td>
</tr>
<tr>
<td>Item-recognition task</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td>86</td>
<td>0.9</td>
</tr>
<tr>
<td>Same</td>
<td>87</td>
<td>1.8</td>
</tr>
<tr>
<td>Different</td>
<td>86</td>
<td>1.4</td>
</tr>
<tr>
<td>New</td>
<td>92</td>
<td>1.1</td>
</tr>
<tr>
<td>Source-recognition task</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old (hit-hit + hit-miss)</td>
<td>96</td>
<td>0.8</td>
</tr>
<tr>
<td>Same</td>
<td>96</td>
<td>0.9</td>
</tr>
<tr>
<td>Different</td>
<td>95</td>
<td>0.9</td>
</tr>
<tr>
<td>Voice</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old (conditional)</td>
<td>79</td>
<td>2.6</td>
</tr>
<tr>
<td>Same</td>
<td>86</td>
<td>1.9</td>
</tr>
<tr>
<td>Different</td>
<td>72</td>
<td>3.7</td>
</tr>
<tr>
<td>New</td>
<td>99</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Note. Hit-hit = both word and voice judgments were correct. Hit-miss = an old word was recognized as old, but the same–different judgment for voice was incorrect. Conditional accuracy measure for voice judgments on old words = hit–hit divided by (hit–hit + hit–miss).
tion, $F(1, 23) = 4.31, p = .05, \text{MSE} = 3.77$, but reaction times were equivalent for same- and different-voice words, $F(1, 23) = 1.77$.

Only when a word was classified as old could participants proceed to evaluate whether the speaker’s voice had changed $F(1, 23) = 1.77$. Source-memory accuracy was thus higher for same-voice than for different-voice words, $F(1, 23) = 26.10, p < .0001, \text{MSE} = 89.7$. This effect may represent a true memory advantage for same-voice words or a bias to respond “same.” There were too few false alarms to new words to evaluate same versus different response bias. Hit–hit responses to same-voice words were also faster than those to different-voice words, $F(1, 23) = 16.20, p < .001, \text{MSE} = 7,155$.

Analyses of the reaction time data supported the intuition that the source-recognition task was performed hierarchically, with an initial (covert) memory judgment about each word preceding a decision about the speaker’s voice. Responses to remembered words (hit–hit and hit–miss trials) in the source-recognition task were more than 500 ms slower than the corresponding hit responses in the item-recognition task, $F(1, 23) = 113.90, p < .0001, \text{MSE} = 31,064$. In contrast, correct responses to new words (which never called for a voice judgment) were more than 100 ms faster in the source task than in the item task, $F(1, 23) = 19.90, p < .0005, \text{MSE} = 6,631$.

**ERPs to Studied Versus Unstudied Words**

*Item-recognition task.* The ERPs elicited during the item-recognition task are shown in Figure 1. In every ERP figure, negative voltage is plotted upward. Both old and new words elicited early peaks typical of the response to acoustic stimuli: a small amplitude positive wave peaking at about 20 ms poststimulus onset, followed by larger negative and positive peaks at about 100 ms and 200 ms (N1 and P2). The later portion of the waveforms was characterized by a broad negative wave at all scalp locations, followed by a positive-going wave most evident at parietal and occipital sites.

The difference between correct old and new words (hits vs. correct rejections) consisted of more positive ERPs for old words, beginning around 400 ms poststimulus onset and extending beyond 1,200 ms at most electrode sites. This old–new difference, or word retrieval effect, was evident at all scalp sites. The word recognition effect was quantified as the mean voltage within two time windows, 400–800 ms and 800–1,200 ms, both relative to a 100-ms prestimulus baseline. The initial analysis of variance (ANOVA) using variables of old versus new words, time window (2 levels), and scalp site (13 levels) revealed a main effect of old versus new words, $F(1, 23) = 63.90, p < .0001, \text{MSE} = 5.19$. A second analysis using only the lateral scalp sites was conducted to describe the time course and the scalp distribution of the word recognition effect in the item task.$^3$ This ANOVA used variables of old versus new words, time window, and laterality (left vs. right) and a variable reflecting the anterior–posterior location of the electrodes (5 levels). Like the initial analysis including the midline sites, the greater positivity elicited by old words was reflected in a main effect of old versus new, $F(1, 23) = 61.90, p < .0001, \text{MSE} = 14.80$. This recognition effect was of equivalent amplitude in the two time windows and varied little in amplitude across the scalp sites (Old–New Words × Time Window, Old–New Words × Laterality, and Old–New Words × Anterior–Posterior Location: $Fs < 2$).

Figure 2 shows that responses to correctly recognized old words were not influenced by the voice-change manipulation. An ANOVA with the variables same versus different voice, scalp site (13 levels), and time window (0–400 ms, 400–800 ms, and 800–1,200 ms) revealed no main effect or interactions involving the voice-change variable ($Fs < 1.20$).

*Source-recognition task.* We initially examined responses in the source-recognition task independent of accuracy in recognizing the voice of the speaker. The hit–hit and hit–miss trials were thus collapsed to form a category of recognized words that was equivalent to the hit category in the item-recognition task. Figure 3 shows that the difference between old and new words was generally similar to that observed in the item task. Compared with unstudied words, correctly recognized old words elicited greater positivity across all scalp sites, beginning at about 400 ms poststimulus. However, Figure 3 also shows that the difference between old and new words was characterized by a late positivity at prefrontal scalp sites that was much less evident in the item-recognition task. The recognition effect during the source task was initially evaluated in the same manner as for the item task; the two recognition tasks are explicitly compared in the next section of this article.

An ANOVA including data from all scalp sites in the source-recognition task yielded a significant main effect of old versus new words, $F(1, 23) = 43.70, p < .0001, \text{MSE} = 14.40$. An analysis of the data from the lateral scalp sites was performed to assess the time course and the scalp distribution of the retrieval effect, using variables of old versus new words, time window (400–800 ms vs. 800–1,200 ms), laterality, and anterior versus posterior location. This analysis yielded a number of interactions indicating that the difference between old and new words varied both across time and across scalp sites. The large amplitude positivity at prefrontal sites led to an interaction between old versus new words and anterior versus posterior scalp sites, $P(4, 92) =$.

---

$^3$ The midline electrodes were excluded from analyses where scalp distribution was of interest because not all of the lateral pairs were accompanied by a midline site. There is, for instance, no “temporal midline” to match the left and right temporal sites. It is also typical for ERP effects to be largest at midline sites (perhaps because these combine activity from the right and left hemispheres), so that the asymmetry of experimental effects is best assessed in analyses including only lateral sites.
11.40, p < .0001, $\eta^2 = .47$, $MSE = 1.53$. The late onset of the prefrontal effect led to an interaction between old versus new words, time window, and anterior versus posterior scalp sites, $F(4, 92) = 51.40, p < .0001, \eta^2 = .49$, $MSE = 0.41$. Finally, this analysis indicated that the laterality of the old–new difference varied across time, Old–New Words × Time Window × Laterality; $F(1, 23) = 5.30, p < .05$, $MSE = 0.35$. Follow-up analyses showed that the old–new difference was larger over the left than the right scalp in the earlier but not the later latency window, 400–800 ms: $F(1, 23) = 10.70, p < .005, MSE = 0.41$; 800–1,200 ms: $F < 1$.

4 The Huynh–Feldt epsilon correction for inhomogeneity of variance was used for all $F$ values with more than one degree of freedom in the numerator.
Figure 2. Grand average event-related potentials (ERPs) from 24 participants during the item-recognition task of Experiment 1. ERPs elicited by correctly recognized old words presented in the same voice during study and test are contrasted with correctly recognized old words presented in a different voice at test. Left scalp sites are displayed in the left column, midline sites in the middle column, and right scalp sites in the right column, from most anterior (top) to most posterior (bottom). Pf corresponds to prefrontal, B to inferior prefrontal (Broca’s), T to temporal, W to parietotemporal (Wernicke’s), and O to occipital. Le denotes an electrode site below the right eye. Fz, Cz, and Pz are located at the frontal, central, and parietal midline, respectively.

Figure 4 suggests that the early asymmetry of the old—new difference was present in both recognition tasks, although it was statistically significant only in the source task.5

Comparisons between the two recognition tasks. The observation of more positive ERPs for studied than unstudied words in both recognition tasks was expected given similar results in a number of previous studies reviewed above, although most of those studies were conducted in the visual modality. Of greater interest here is whether the requirement to make a source-recognition judgment elicits an ERP signature distinct from the basic item-recognition signature.

5 A left-larger-than-right asymmetry for the difference between old and new items has been observed in other ERP memory studies using printed rather than spoken words (Allan, Doyle, & Rugg, 1996; Allan & Rugg, 1997; Van Petten et al., 1991). The influence of handedness on this memory-related asymmetry has not been investigated. However, it is clear that scalp asymmetries for
Figure 3. Grand average event-related potentials from 24 participants during the source-recognition test of Experiment 1. “Hit” includes all trials with successfully recognized words, independent of the accuracy of the voice judgment (e.g., hit-hit plus hit-miss). Left scalp sites are displayed in the left column, midline sites in the middle column, and right scalp sites in the right column, from most anterior (top) to most posterior (bottom). Pf corresponds to prefrontal, B to inferior prefrontal (Broca’s), T to temporal, W to parietotemporal (Wernicke’s), and O to occipital. Le denotes an electrode site below the right eye. Fz, Cz, and Pz are located at the frontal, central, and parietal midline, respectively.

different component of the ERP (the N400) are most pronounced for right-handed individuals without left-handed parents or siblings (Kutas, Van Petten, & Besson, 1988). Analyses of the present data revealed a similar pattern of results. The leftward asymmetry of the old-new effect observed in the 400–800-ms time window was significant for the 15 right-handed participants from exclusively right-handed families, F(1, 14) = 5.24, p < .05, but not for the 9 participants who were left-handed or had left-handed family members (F < 1).

positivity. A comparison of Figures 1 and 3 suggests that the most prominent difference between the two tasks was observed at the prefrontal scalp sites. In both tasks, old words elicited a more positive ERP than new words, beginning 300–400 ms poststimulus onset. This word retrieval effect was fairly similar across tasks and across electrode sites until about 700 ms poststimulus. The later portion of the recording epoch was marked by a large prefrontal positivity elicited by old words in the source-
recognition task. The prefrontal old-new difference reached maximum amplitude at about 1,000 ms, substantially later than the basic recognition effect observed in both tasks at other electrode sites but well before the average reaction time in the source-recognition task. The focal scalp distribution and the late onset of the prefrontal effect are illustrated in Figure 5.

The impact of recognition task was evaluated by an ANOVA using task (item vs. source), old versus new words, and time window (400-800 ms vs. 800-1,200 ms) as variables, together with scalp site variables of laterality and anterior versus posterior. Much of this analysis replicated those reported above; therefore, only those outcomes involving the task variable or clarifying the previous analyses are noted. The difference between studied and unstudied words was slightly larger overall in the source-recognition task than in the item-recognition task, Task × Old-New Words: \( F(1, 23) = 4.48, p < .05, \text{MSE} = 4.19. \) The initial phase of the old-new difference (400-800-ms time window) was larger over the left than the right, whereas the late phase was symmetric, Old-New Words × Time Window × Laterality: \( F(1, 23) = 6.70, p < .02, \text{MSE} = 0.35. \) Although the independent analyses of the item and source tasks reported above included a significant leftward asymmetry in only the source task, the present combined analysis revealed no interactions involving Task × Laterality (see Figure 4). The late prefrontal positivity elicited by old words in the source-recognition task produced significant interactions of Task × Old-New Words × Anterior-Posterior Location, \( F(4, 92) = 14.10, p < .0001, e = .53, \text{MSE} = 0.77, \) and Task × Old-New Words × Time Window × Anterior-Posterior Location, \( F(4, 92) = 29.30, p < .0001, e = .44, \text{MSE} = 0.26. \)

**Session effects.** The two item-recognition sessions were conducted before the two source-recognition sessions to avoid possible contamination of the item-memory measures by carryover effects from the source task. The fixed order of the item- and source-recognition sessions raises the possibility that the source sessions were subject to either a beneficial effect of practice or a detrimental effect of interference from previously studied items. Some aspects of the experiment suggest that both possibilities are slight: (a) Each session began with a practice study-recognition list; (b) the sessions were conducted on different days; and (c) for a given participant, no words appeared in more than one session. However, the nonspecific impact of time-on-task can be empirically evaluated by assessing whether any of the performance or ERP memory measures changed between Session 1 and Session 2 or between Session 3 and Session 4. For item-recognition performance, neither hit nor correct rejection rates changed from Session 1 to Session 2 (main effect of session, \( F_s < 1 \)). In the two source-recognition sessions, neither item accuracy (hit-hit plus hit-miss rate, correct rejection rate) nor source accuracy levels changed between Session 3 and Session 4 (\( F_s < 1 \)).

The influence of session on the ERP measures was analyzed by repeating the most central analyses presented above with session as a variable. The item-recognition
Scalp distribution of the recognition effects

**Figure 5.** Mean amplitude of the difference between hits and correct rejections measured in two time windows, 400–800 ms and 800–1,200 ms poststimulus onset, relative to a 100-ms prestimulus baseline. Left and right electrode sites are averaged to show scalp distribution in the anterior–posterior dimension. Vertical lines depict standard errors of the means. Pf, B, T, W, and O are located at the prefrontal, inferior prefrontal (Broca's), temporal, parietotemporal (Wernicke’s), and occipital scalp sites, respectively.

measures were analyzed with a repeated measures ANOVA on the lateral electrode sites, using session (1 vs. 2), old–new (hit vs. correct rejection), time window (400–800 ms vs. 800–1,200 ms), laterality (left vs. right), and anterior–posterior (5 levels) as variables. The overall amplitude of the recognition effect did not change across sessions (Session × Old–New, F < 0.60). Changes in the timing or scalp distribution of the recognition effect across sessions would be reflected in interactions involving Session × Old–New and one of the other variables in this analysis. The majority of the three-way and four-way interactions, as well as the five-way interaction, were nonsignificant (Fs < 1.35). The sole exception was a small shift in the laterality of the old–new effect at the more posterior electrode sites from the first to the second session, as seen in Table 2, Session × Old–New × Laterality × Anterior–Posterior: F(4, 84) = 5.34, p < .01, ε = .53, MSE = 0.52. A similar analysis of the two source-recognition sessions yielded no significant interactions involving Session × Old–New (Fs < 1.5). Because these analyses did not suggest that the recognition effect became either later in time or more focused at the prefrontal electrode sites across the multiple sessions, they did not qualify the differences between item and source recognition described above.

Retrieving voice information: The success or failure of the source judgment. Thus far, we have considered only successful recognition of old versus new words. However, participants made enough errors of source recognition to examine the ERPs contingent on the success or failure of this discrimination. Figure 6 shows that the hit–hit trials elicited more positive ERPs than hit–miss trials during the source-recognition task. This voice retrieval effect was of late onset, beginning about 800 ms poststimulus. The voice retrieval effect was largest at midline and temporal sites but essentially absent at prefrontal (PfL, PfR) sites.

An initial ANOVA taking data from all the scalp sites revealed a significant effect of voice-retrieval accuracy, hit–hit versus hit–miss: F(1, 23) = 5.57, p < .05, MSE = 12.10. A more detailed analysis of the lateral scalp sites used variables of voice retrieval (hit–hit vs. hit–miss), time window (400–800 ms vs. 800–1,200 ms), anterior–posterior, and laterality. The main effect of voice retrieval was significant, F(1, 23) = 5.00, p < .05, MSE = 5.85. The late

---

6 There were an average of 57 trials in the hit–miss category for each participant. This was the smallest category of interest in the experimental design and did not allow a subdivision by session without an unacceptable decrease in the signal-to-noise ratio.
Mean Amplitudes (in Microvolts) of the Old-New Difference in Event-Related Potentials Elicited During the Two Item-Recognition Sessions in a Latency Window of 400–1,200 ms Poststimulus Onset

<table>
<thead>
<tr>
<th>Electrode site</th>
<th>Session 1</th>
<th>Session 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PfL</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>PfR</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>BL</td>
<td>0.4</td>
<td>0.7</td>
</tr>
<tr>
<td>BR</td>
<td>1.0</td>
<td>0.7</td>
</tr>
<tr>
<td>TL</td>
<td>0.9</td>
<td>1.1</td>
</tr>
<tr>
<td>TR</td>
<td>1.3</td>
<td>0.8</td>
</tr>
<tr>
<td>WL</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>WR</td>
<td>1.1</td>
<td>0.6</td>
</tr>
<tr>
<td>OL</td>
<td>0.1</td>
<td>1.1</td>
</tr>
<tr>
<td>OR</td>
<td>0.6</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Note. PfL = left prefrontal; PfR = right prefrontal; BL = left inferior prefrontal; BR = right inferior prefrontal; TL = left temporal; TR = right temporal; WL = left parietotemporal; WR = right parietotemporal; OL = left occipital; OR = right occipital.

The latencies of the ERP memory effects provided additional information about the time course of retrieving word and voice information and suggest that the hierarchical retrieval model provides a better description than a horse race account.
SOURCE RECOGNITION TASK

Figure 6. Grand average event-related potentials from 24 participants during the source-recognition task of Experiment 1. “Hit-hit” trials are those in which the behavioral response was correct as to both the item and the source-voice. “Hit-miss” trials are those for which the word was correctly judged as old but the source judgment was inaccurate (e.g., judgments of “old, same” for words that changed voice from study to test and judgments of “old, different” for words presented in the same voice at study and at test). Left scalp sites are displayed in the left column, midline sites in the middle column, and right scalp sites in the right column, from most anterior (top) to most posterior (bottom). Pf corresponds to prefrontal, B to inferior prefrontal (Broca’s), T to temporal, W to parietotemporal (Wernicke’s), and O to occipital. Le denotes an electrode site below the right eye. Fz, Cz, and Pz are located at the frontal, central, and parietal midline, respectively.

race model. The ERPs elicited by correctly categorized old and new words began to diverge at about the same time in the item- and source-memory tasks. Both tasks showed a word retrieval effect as early as 400 ms poststimulus onset (on average, some 200 ms before the offset of the spoken words and 700 ms before the behavioral response). The ERPs recorded during the source-recognition task suggest that retrieval of the contextual voice information not only took additional time but engaged a brain circuit that was not strongly activated during item recognition. In addition to the
Table 3

<table>
<thead>
<tr>
<th>Electrode site</th>
<th>$F(1, 23)$</th>
<th>$p&lt;$</th>
<th>$\text{MSE}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prefrontal (PFL, PFR)</td>
<td>1.51</td>
<td>ns</td>
<td>0.55</td>
</tr>
<tr>
<td>Inferior frontal (BL, BR)</td>
<td>8.23</td>
<td>.01</td>
<td>0.36</td>
</tr>
<tr>
<td>Temporal (TL, TR)</td>
<td>19.70</td>
<td>.0005</td>
<td>0.34</td>
</tr>
<tr>
<td>Parietotemporal (WL, WR)</td>
<td>17.00</td>
<td>.0005</td>
<td>0.31</td>
</tr>
<tr>
<td>Occipital (O1, O2)</td>
<td>5.85</td>
<td>.05</td>
<td>0.55</td>
</tr>
</tbody>
</table>

Basic word retrieval effect observed in both tasks at all scalp sites, old words in the source-recognition task elicited a later positive potential restricted to the prefrontal scalp sites. We interpret this late prefrontal positivity as a voice search effect that occurred only after identifying a word as a studied item requiring a voice judgment. The similarity of the prefrontal ERPs for hit-hit and hit-miss trials indicates that it was the attempt to recover voice information that triggered the prefrontal positivity rather than the success or failure of the attempt. The ERPs also differentiated successful (hit-hit) from unsuccessful (hit-miss) recognition of voice information, but only at more posterior scalp sites and not until some 700–800 ms poststimulus onset. Because the prefrontal effect is related to effort rather than the outcome of the source-memory judgment, its latency is germane to the question of when participants first attempted to recover the linkage of word-plus-voice. The delayed onset of the prefrontal voice search effect relative to the word retrieval effect suggests that source retrieval processes began only after old and new words had been successfully distinguished.

In the study phases of the experiment, we observed little sign of an ERP difference between successfully and unsuccessfully encoded item or source information. This finding was surprising given the number of previous studies reporting larger late positive components for subsequently recognized words (Dm effect). The large majority of the previous studies were conducted in the visual modality, but one preliminary report described a similar phenomenon with spoken words (Wyatt et al., 1995), so that the null effect here is unlikely to reflect a fundamental difference between printed and spoken words. Instead, the absence of a typical Dm effect here can probably be attributed to the study-phase task of size judgment. We have previously noted that study-phase trials accompanied by affirmative judgments (e.g., "living" or "edible") elicit much larger Dm effects than those accompanied by negative judgments (e.g., "nonliving" or "inedible"); Paller et al., 1987; Van Petten & Senkfor, 1996). In the present experiment, the study-phase judgments of "large" or "small" did not include an obvious affirmative-negative polarity. Similarly, one previous study using a size judgment task yielded a small and statistically nonsignificant difference between subsequently recognized and unrecognized words (Paller & Kutas, 1992). Further research will be required to understand the functional significance of ERP effects correlated with successful encoding.

Before elaborating on how the recognition-phase results might be related to other studies of source memory, we report a second experiment designed to address two possible confounds in the design of the first experiment. Shorter study lists were used in the source-memory task in an attempt to equate performance across the two recognition tasks. No previous research has suggested that the scalp distribution of ERP memory effects is dependent on list length. But it remains a possibility that list length contributed to the prefrontal positivity, perhaps by increased confidence in recognition judgments for the short source lists of 17 words as compared with the long item lists of 136 words. The item- and source-recognition tasks also differed in their response requirements. During item recognition, participants used only two keys to indicate "old" or "new." During source recognition, they used three different response keys to indicate "old, same"; "old, different"; or "new." The results indicated that word retrieval was initiated before voice retrieval, but it is possible that some aspect of the response requirements encouraged participants to perform the task in two stages. The design of Experiment 2 eliminated both of these possible confounds. Lists of 17 words were used in the study.
Procedure

Recognition Performance

In the source-recognition phase, no new words were presented, and participants were instructed to respond only "same" or "different" on the basis of the speaker's voice. The central goal of Experiment 2 was to more securely link the late prefrontal positivity with source monitoring by ruling out extraneous task differences.

Experiment 2

Method

Participants

Eight men and 8 women ranging in age from 19 to 30 years (M = 25 years) participated as paid volunteers. All participants were native English speakers with normal hearing. Fifteen participants were right-handed; 5 of these participants reported having a left-handed family member (parent or sibling). One participant was left-handed and reported having a left-handed family member. The data from two additional participants were not analyzed because of an excessive number of eye movement artifacts.

Stimuli

Four hundred and eight of the stimuli from Experiment 1 were used. Mean duration and frequency of usage were equivalent to those in Experiment 1.

Procedure

The experiment was conducted in one session lasting about 3.5 hr. The first half of the session consisted of item study and recognition; the second half of the session consisted of source study and recognition tests. Eight study–recognition cycles were used in each phase of the experiment. All aspects of stimulus presentation and tasks were identical to those in Experiment 1, with the exception of shorter word lists in the item phase and the absence of those in Experiment 1.

For both the item and source phases, the study lists contained 17 words each (half male and half female voices). Item-recognition tests consisted of 34 words: Half were new words, one quarter were old words in the same voice as study, and one quarter were old words in the other voice. As in Experiment 1, participants responded "old" or "new" during item recognition. The source-recognition tests contained only old words, half presented in the same voice as study and half in the other voice. Participants were forewarned that all of the words in the source-recognition task were old and were requested to respond only "same" or "different voice."

Eight different stimulus lists were constructed to counterbalance task, old–new words, and same–different voice at recognition. Across 16 participants, each list was used twice. Electrophysiological methods were identical to those used in Experiment 1.

Results

Recognition Performance

Accuracies and reaction times for both recognition tasks are shown in Table 4. Overall accuracy (hits plus correct rejections) in the item-recognition task averaged 95%, higher than that in Experiment 1 with longer lists but comparable to word-recognition accuracy in the source task of that experiment (96%). For old words, hit rate was not influenced by the voice-change manipulation (F < 1.5).

No new words were included in the source-recognition task, so that raw accuracy rates for responding "same" or "different" voice indexed source accuracy. Source accuracy averaged 77%, quite similar to that in Experiment 1 (79%). As in Experiment 1, source accuracy was higher for same-voice trials than for different-voice trials, F(1, 15) = 4.58, p < .05, MSE = 184.

All aspects of the reaction time results were also similar to those of the previous experiment. In Experiment 1, correct responses to old words in the source task were 540 ms slower than in the item task. In Experiment 2, same–different judgments about the voice of old words were 537 ms slower than recognition judgments for old words in the item task, F(1, 15) = 90.5, p < .0001, MSE = 25,449. In the item task, reaction times were not influenced by the voice-change manipulation (F < 1.5). In the source task, "same" judgments were faster than "different" judgments, F(1, 15) = 14.3, p < .002, MSE = 10,061.

ERPs

Item-recognition task. During item recognition, ERPs elicited by same-voice trials were indistinguishable from different-voice trials. The ERPs did not vary significantly in any time window; for an ANOVA with variables of same–different voice, time window (0–400 ms, 400–800 ms, and 800–1,200 ms), and electrode site, all main effects and interactions involving same–different voice yielded Fs < 1. Figure 8 (left column) shows a word recognition effect much like that in Experiment 1 but beginning somewhat earlier, around 300 ms poststimulus onset. In both the 400–800-ms and 800–1,200-ms time windows, hits elicited more positive ERPs than correct rejections, but the effect was larger in the early time window. Old–New Words × Time Window × Electrode Site: F(12, 180) = 24.4, p < .0001, e = .23, MSE = 0.23. A follow-up ANOVA with only the prefrontal sites showed that the difference between hits and correct rejections did not significantly change from the early to later time window at these sites, F(1, 15) = 2.25, p > .15. The similarity of these results to those of Experiment 1 con-

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Accuracy (%)</th>
<th>Reaction time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SE</td>
</tr>
<tr>
<td>Item-recognition task</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hit</td>
<td>95</td>
<td>0.8</td>
</tr>
<tr>
<td>Old, same</td>
<td>97</td>
<td>0.8</td>
</tr>
<tr>
<td>Old, different</td>
<td>95</td>
<td>0.7</td>
</tr>
<tr>
<td>New</td>
<td>95</td>
<td>1.5</td>
</tr>
<tr>
<td>Source-recognition task</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hit</td>
<td>77</td>
<td>2.9</td>
</tr>
<tr>
<td>Same</td>
<td>83</td>
<td>1.9</td>
</tr>
<tr>
<td>Different</td>
<td>73</td>
<td>4.9</td>
</tr>
</tbody>
</table>
firmed our hypothesis that list length has little impact on the timing or scalp distribution of the word retrieval effect.

**Item versus source task.** The second purpose of this experiment was to determine if the source task would elicit a late prefrontal positivity given a dichotomous response requirement equivalent to that used in the item-recognition task. If the prefrontal positivity reflects a search for voice information, it should be elicited by all words in the source task relative to the item task. The observed latency of any difference between the two tasks is relevant to a second issue. If participants must first recognize a word before attempting to retrieve voice information, the task effect should be delayed relative to the basic word retrieval effect observed in the item-recognition task. Alternatively, it may be possible to immediately engage voice search-retrieval processes given that all the words in the source-recognition task were old. Two statistical comparisons are relevant to both the scalp distribution and latency issues.

The first comparison is between old words (all) in the source task and new words (correct rejections) in the item task. The middle column of Figure 8 shows that the old–new difference began about 300 ms poststimulus onset but was marked by a late prefrontal positivity not evident in the item task alone. The early phase (400–800 ms) of the old–new difference did not vary across scalp sites, main effect of old–new: $F(1, 15) = 27.1, p < .0001, MSE = 7.20$; Old–New $\times$ Anterior–Posterior: $F < 1$. The late phase (800–1,200 ms) was frontally distributed and largest at the prefrontal sites, old–new: $F(1, 15) = 11.8, p < .005, MSE = 5.90$; Old–New $\times$ Anterior–Posterior: $F(4, 60) = 19.6, p < .0001, e = .51, MSE = 2.15$. An ANOVA including both time windows thus yielded an Old–New $\times$ Time Window $\times$ Anterior–Posterior interaction, $F(4, 60) = 29.3, p < .0001, e = .43, MSE = 0.70$.

The second comparison—old words in the item task versus old words in the source task—provides a direct
contrast between making item and source judgments about previously studied words. The right column of Figure 8 shows that the ERPs elicited by old words in the two tasks were very similar up to about 600 ms. Responses to old words in the source task became more positive after about 600 ms, particularly at frontal sites. An ANOVA using task, time window (400–800 ms and 800–1,200 ms), anterior–posterior, and laterality as variables thus included a significant Task × Time Window × Anterior–Posterior interaction, $F(4, 60) = 4.07, p < .05, e = .55, MSE = 0.44$, without a main effect of task or Task × Time Window interaction ($Fs < 1$).

**Discussion**

Overall, the results of Experiment 2 replicated the critical findings of Experiment 1. Decreasing the number of studied words in the item-recognition task from 136 to 17 yielded a higher accuracy rate equivalent to item-recognition accuracy in the short-list source task of Experiment 1. The change in list length did not, however, alter the scalp distribution of the old–new effect in the item-recognition task. The second procedural difference between the experiments was the deletion of new words from the source-recognition task, so that participants had only two response options of “same” and “different” rather than three alternatives of “same,” “different,” and “new.” This alteration also had remarkably little impact. Source accuracy rates were nearly unchanged, and responses to old words in the source task were still more than half a second slower than in the item-recognition task. Finally, the source-recognition task elicited a large prefrontal positivity not observed in the item-recognition task. As in the first experiment, the onset of the prefrontal task difference was delayed well beyond the onset of the basic old–new difference, or word retrieval effect. The delayed onset of the prefrontal task difference suggests that whether or not an old–new judgement is explicitly required, participants postpone source-memory processes until an item has been recognized.

**General Discussion**

Across the two experiments, two central findings emerged: (a) Words and their associated voices were not retrieved simultaneously but in sequence, and (b) the attempt to recover source information engaged a neural circuit that was distinct from that required for word retrieval alone. These two findings are discussed in turn.

**When Is Voice Information Recovered?**

Neither experiment suggested that voice information was recovered automatically, without an explicit task requirement. In the item-recognition tasks, words that changed voice from study to test did not differ from same-voice words in accuracy, reaction time, or ERPs elicited. These results are unlike those of continuous recognition paradigms, showing greater accuracy and faster reaction times for same-voice words (Craik & Kirsner, 1974; Palmeri et al., 1993). Neither the number of voices nor the lag time between study and test provides obvious explanations of this discrepancy. Palmeri et al. obtained similar results with 2 or 20 voices and with lags of up to 64 intervening items (comparable to Experiment 2 using study lists of 17 and recognition lists of 34, for an average lag of 25 items). One difference among studies is that continuous recognition paradigms do not have a separate encoding phase, unlike a study–recognition paradigm. It is possible that in the absence of a specific encoding task, perceptual attributes such as voice become a central focus during the encoding of each episodic event. In contrast, our size judgment study task focused attention on the semantic aspects of each word rather than the voice. Glisky et al. (1995) reported a similar null effect of a voice-change manipulation on recognition speed and accuracy in an experiment using a semantic study task.

A recent study by Goldinger (1996) supports the idea that voice-change effects in word recognition are best observed when initial encoding emphasizes auditory or phonetic aspects of the stimuli. In a study–test paradigm, participants who performed gender or phoneme classifications during the study phase were faster and more accurate (by about 11%) in recognizing same- than in recognizing different-voice words. In contrast, participants who performed a syntactic classification task at study showed no reaction time facilitation and a smaller accuracy advantage of about 4% for same-voice words (Goldinger, 1996). Schacter and Church have also compared recognition accuracy for words studied with auditory (pitch judgment) versus semantic encoding tasks and observed larger voice-change effects after shallow study tasks. However, in this series of experiments, none of the voice-change effects on recognition accuracy (ranging from 0.8% to 6.4%) reached statistical significance under any encoding condition (Church & Schacter, 1994; Schacter & Church, 1992; Schacter, Church, & Treadwell, 1994). Because reported accuracy advantages for same-voice words hover around 10% under optimal conditions (Goldinger, 1996; Palmeri et al., 1993), quantifying the influence of encoding manipulations is a difficult endeavor requiring large numbers of trials, participants, or both. The present observations that changing voice from study to test did not influence explicit word recognition we redrawn from a large stimulus set (272 and 136 words studied in the item tasks of Experiments 1 and 2, respectively) and 40 participants. The results suggest that voice information is not relied on during the recognition of words that have been semantically encoded.

Although voice retrieval was not apparent or required in the item-recognition task, word and voice retrieval was required for successful performance in the source task. Comparisons across the two recognition tasks suggest a hierarchical retrieval process. The response times for correctly remembered old words were slower in the source task as compared with the item task, although new words were somewhat faster. This pattern of results suggests that participants did not assess whether a particular linkage of word and voice was previously encountered until they had established that the word itself was a studied item. This hierarchical
SOURCE MEMORY

regardless of the modality judgment: The hit minus false-retrieval view is further supported by ERP latency measures. In both recognition tasks, the difference between old and new words began around 400 ms at all scalp sites. However, when voice information was asked for, this basic old–new effect was followed by a prefrontal positivity beginning around 700 ms. Because the late prefrontal positivity was equivalent for hit–hit and hit–miss trials, we interpreted it as a voice search effect. At the more posterior sites, successful voice retrieval was observed in the difference between hit–hit and hit–miss trials, which did not begin until around 800 ms. Moreover, the temporal sequence of item and source retrieval processes was not altered even when the participants were given only old items in the source task. In the source task of the second experiment, participants were informed that all the words were old and that they merely needed to judge the voice. The ERPs were similar to those in Experiment 1 in that a word recognition effect began substantially before the prefrontal voice search effect.

A recent study using a different source manipulation also supports the view that retrieval of source information occurs after item recognition. Johnson, Kounios, and Reeder (1994) evaluated the time course of reality monitoring through a deadline methodology: Participants were required to make recognition judgments at various time points ranging from 300 to 1,500 ms poststimulus onset. Probe words in the recognition test referred to pictures that had been studied (old, perceptual), mental images formed during study (old, Imagine), or were new. Old–new discriminations exceeded chance levels beginning at the 300-ms response deadline, but source discriminations between perceived and imagined pictures did not exceed chance levels until about 500 ms. These results indicate that “items” and “sources” can be dissociable components of an episodic event.

One previous ERP study examined memory for words versus modality in which they were studied (Wilding, Doyle, & Rugg, 1995). A mixed set of spoken and written words was studied, followed by either a visual recognition test (Experiment 1) or an auditory recognition test (Experiment 2). Each old–new recognition judgment was followed by a separate study–modality discrimination. Although the experiments were designed to distinguish perceptual fluency from conscious recollection, they can also be viewed and discussed as a source-memory test. Wilding et al.’s study did not include the prefrontal recording sites used here, but the success or failure of the modality judgment can be compared with the voice judgment of the present study. Both experiments yielded a general old–new effect much like that observed here. In Wilding et al.’s first experiment, both word and modality judgments were more accurate for words presented visually at study and test, so that the within-modality condition contained too few hit–miss trials to form reliable ERPs. In the cross-modal condition, hit–miss trials did not differ from new items at any point in the recording epoch. This result is distinct from the present observation of a late (800–1,200 ms) three-way difference between the categories of hit–hit, hit–miss, and correct rejection. However, the participants in Wilding et al.’s first experiment showed generally weak memory for the cross-modal words regardless of the modality judgment: The hit minus false-alarm rate was 33% as compared with 78% for different-voice words in Experiment 1 of the present study. The similarity of the ERPs elicited by hit–miss and correct rejection trials may only indicate that hit–miss trials were accompanied by a very weak memory trace. Memory performance was better in Wilding et al.’s second experiment, and these results were similar to our observations: Hit–hit and hit–miss trials were similar in the 400–800-ms range (both distinct from new words), but hit–hits elicited more positive ERPs in a latency range of 800–1,300 ms poststimulus. Overall, the results of Wilding et al.’s study are consistent with the present results using a speaker’s voice as a source manipulation.

Frontal Cortex and Source Memory

In both of the present experiments, a late positivity at the prefrontal scalp sites was observed in the source task and not in the item task. Because we observed a focal prefrontal positivity during source recognition and not in the preceding study phase, the present data suggest that prefrontal cortex is preferentially involved in linking items and their sources during retrieval. The focal scalp distribution of this task effect is consistent with the type of memory deficits observed in patients with frontal damage. Recognition is typically intact in patients with frontal lobe damage while source judgments are impaired (Janowsky et al., 1989; Wheeler, Stuss, & Tulving, 1995). However, it is not possible to derive precise anatomical localization from scalp distribution information alone because the spatial distribution of scalp potentials is determined by both the location and geometric orientation of active neurons in a convoluted cortex (Dale & Sereno, 1993; Nunez, 1981, 1990). There are no neuropsychological studies using voice as a source manipulation so that it would be informative to record ERPs from patients with frontal lobe damage in the current paradigm. If the search for voice information is conducted by prefrontal cortex, then the late prefrontal positivity seen in normal participants should be greatly reduced or absent in patients with frontal lobe damage, and their source performance should be impaired. Direct neuropsychological evidence of this sort will be necessary to securely link the voice search effect at prefrontal scalp sites with prefrontal cortex. A combination of neuropsychological and electrophysiological approaches will be particularly fruitful in the study of source memory because ERP methods can differentiate encoding from retrieval processes more easily than studies with patients. Source-memory results from both patients and neurologically intact participants will be of particular importance when attempting to make generalizations across these two literatures.

Functional neuroimaging techniques such as positron emission tomography (PET) are able to identify brain regions differentially active across conditions with reasonable spatial resolution. To date, no PET studies have compared item- and source-memory tasks, but several studies support our conclusion that prefrontal cortex is linked to retrieval effort. A recent interest in different aspects of the retrieval process—attempt versus success—has given
rise to the hypothesis that prefrontal cortex is preferentially involved in the attempt, whereas medial temporal regions are activated by successful retrieval. Numerous PET studies have shown both prefrontal and medial temporal activations during recognition tests of memory (Cabeza et al., 1997; Fletcher et al., 1995; Grady et al., 1995; Kapur et al., 1995; Moscovitch, Kapur, Kohler, & Houle, 1995; Nyberg et al., 1995; Roland, Kawashima, Gulyas, & O'Sullivan, 1995; Schachter, Albert, Savage, Rauch, & Albert, 1996; Tulving et al., 1994). A priori, these activated brain regions may reflect the retrieval attempt, the success of this attempt, or both. One limitation of the PET methodology is the inability to sort trials contingent on participants' memory performance; in the present ERP study, the comparison between hit–hit and hit–miss trials was critical for disentangling retrieval attempt from retrieval success. Three less direct approaches have been used to distinguish between retrieval attempts and retrieval successes in PET studies: varying the proportion of old and new items during the scan, correlating recognition accuracy with blood flow across individual participants, and manipulating the hit rate by multiple presentations of the study item prior to the recognition test. These approaches have suggested that prefrontal cortex is associated with retrieval attempt, and medial temporal regions with retrieval success. Kapur et al. varied the proportion of old–new items and found that lists with 85% old items elicited greater blood flow in medial temporal regions than lists with 15% old items. Blood flow in prefrontal regions did not differ between the lists. Similar designs have produced consistent results, although this finding has also been subject to debate (Fletcher, Frith, & Rugg, 1997; Tulving, Markowitsch, Craik, Habib, & Houle, 1996). The correlation approach has yielded a positive relationship between hit rate and medial temporal blood flow, lending further support to the “successful retrieval” interpretation of medial temporal activity (Nyberg, Cabeza, & Tulving, 1996). Finally, Schacter et al. (1996) observed differences in both prefrontal and medial temporal blood flow between two tests of word–stem completion. When participants were allowed to study words only once prior to the scan, fewer word stems were completed with studied items than after four presentations of the study list. The low recall condition was associated with lesser blood flow in the medial temporal lobe, attributed to the smaller number of trials with successful recall. However, this low recall condition was also associated with greater blood flow in prefrontal regions. This result was attributed to greater effort expended in the attempts to recall studied items after only a single prior presentation.

The PET studies are in general agreement with the present ERP data in suggesting that prefrontal cortex is involved in the attempt to retrieve information but does not reflect the success or failure of these attempts. We concluded this after examining trials sorted on recognition accuracy: Prefrontal ERPs were identical for trials with correct and incorrect voice judgments, although more posterior recording sites did reflect the accuracy of this judgment. This analysis is not currently possible with PET, but the superior spatial resolution of functional imaging techniques complements the ERP technology. However, the PET studies described above investigated only memory for words, much like our item task. During the item-recognition task, we observed a spatially diffuse word retrieval effect (largest temporally), whereas the source-recognition task elicited a focal prefrontal positivity. It is possible that the item-recognition task also engaged prefrontal cortex to some extent but to a much lesser degree than the source-recognition task. Further research will be necessary to reconcile observations about human memory and brain function provided by electrophysiological and blood flow measures.

The present study demonstrates that item and source memory differ in both time course and neural circuitry in normal young participants. The present data support the proposal that linkage of words and voices, source memory, requires retrieval of one attribute before another. The strength of the attribute linkage is reflected in the success or failure to retrieve the attributes presented in conjunction with each other. Some important questions remain for further research. We hope to find that other source manipulations (e.g., temporal order, spatial location) yield similar dissociations in time course and scalp distribution.

References


Received September 18, 1996
Revision received April 3, 1997
Accepted July 7, 1997