RICE UNIVERSITY

Separate Short-Term Memory Buffers for Input and Output Phonology

by

Denise Hsien Wu

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

[Signatures]

Randi Martin, Elma Schneider Professor
Chair, Department of Psychology

Michael Watkins, Professor
Department of Psychology

Geoff Potts, Assistant Professor
Department of Psychology

Nancy Niedzelski, Assistant Professor
Department of Linguistics

HOUSTON, TEXAS
DECEMBER, 2002
DECEMBER, 2002

ABSTRACT

Separate Short-Term Memory Buffers for Input and Output Phonology

by

Denise Hsien Wu

Evidence from brain-damaged patients suggests that there are separate buffers for input and output phonological retention in verbal short-term memory (STM). This possible distinction was investigated with college students (Experiment 1 to 3) and deaf signers of American Sign Language (ASL) (Experiment 4) using different verbal materials in a serial probed recall paradigm. It is reasoned that natural linguistic input (speech for hearing people and ASL for deaf people) would be stored in an input phonological buffer whereas internally generated phonology derived from reading, naming pictured objects, or lip-reading would be stored in an output phonological buffer. In this study, participants were presented with memory lists in which presentation modality (spoken vs. lip-read word, written vs. lip-read word, etc.) was changed after every second item. A probe item from the list was repeated at the end of the list and participants were instructed to either recall the item in the list that has immediately followed the probe or recall the first item after the probe that is in the same modality. Some of these same-modality items were temporally distant, that is, having two
intervening items of a different modality. It is predicted that the temporally distant probe in the same modality with the target results in higher memory performance than the temporally adjacent probe in a different modality only if the switch in modalities is between input and output phonological forms. The results from Experiment 1 demonstrated that spoken words and written words were stored in the input and output phonological buffers, respectively. The results from Experiment 2 and 3 further supported the hypothesis in showing that written words were retained in the same buffer with lip-read words and with nameable pictures, while spoken words were retained in a different buffer from these materials. The findings from lists consisting of words in ASL and nameable pictures in Experiment 4 were not conclusive. However, preliminary data suggested that there might also be a separation between signed words and nameable pictures. Overall, the findings from this study conformed to the predictions from the hypothesis of separate input and output phonological retention.
Acknowledgments

I would like to express my sincere appreciation to all those who have helped me during the completion of this dissertation. I would especially like to thank my mentor, Dr. Randi Martin, for her wisdom and guidance. Her clear vision and expertise have taught me so much throughout my graduate study, and her continuous feedback is invaluable to me. I would also like to thank Dr. Geoff Potts, Dr. Nancy Niedzielski, and Dr. Michael Watkins who served as members on the dissertation committee, as well as Dr. David Lane who provided counsel regarding statistical analyses. Finally, I would like to thank my parents and Erik Chang for their love and support.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Figures</td>
<td>ix</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>29</td>
</tr>
<tr>
<td>Experiment 1A</td>
<td>30</td>
</tr>
<tr>
<td>Method</td>
<td>30</td>
</tr>
<tr>
<td>Participants</td>
<td>30</td>
</tr>
<tr>
<td>Materials and Design</td>
<td>30</td>
</tr>
<tr>
<td>Procedure</td>
<td>31</td>
</tr>
<tr>
<td>Results and Discussion</td>
<td>32</td>
</tr>
<tr>
<td>Experiment 1B</td>
<td>35</td>
</tr>
<tr>
<td>Method</td>
<td>35</td>
</tr>
<tr>
<td>Participants</td>
<td>36</td>
</tr>
<tr>
<td>Materials and Design</td>
<td>36</td>
</tr>
<tr>
<td>Procedure</td>
<td>37</td>
</tr>
<tr>
<td>Results</td>
<td>38</td>
</tr>
</tbody>
</table>
List of Figures

Figure 1  A language based model of verbal short-term memory (adopted from Martin et al., 1999).

Figure 2  Mean accuracy on serial probed recall (in percentage) of written letters for three levels of stimulus similarity across serial positions 2 to 8 from Experiment 1A.

Figure 3  Mean accuracy (in percentage) on serial probed recall of spoken and written letters for four probe types across serial position 4 to 10 from Experiment 1B.

Figure 4  Mean accuracy (in percentage) on serial probed recall of spoken and written digits for four probe types across serial position 4 to 10 from Experiment 1C.

Figure 5  Expected and observed accuracy (in percentage) on serial probed recall of spoken and written digits for the same-mode and different-mode conditions across serial position 5 to 10 from Experiment 1C.

Figure 6  Mean accuracy (in percentage) on serial probed recall of spoken and written letters for four probe types across serial position 4 to 10 from Experiment 1D.

Figure 7  Expected and observed accuracy (in percentage) on serial probed recall of spoken and written letters for the same-mode and different-mode conditions across serial position 5 to 10 from Experiment 1D.

Figure 8  Mean accuracy (in percentage) on serial probed recall of spoken and lip-read words for four probe types across serial position 4 to 8 from Experiment 2A.

Figure 9  Mean accuracy (in percentage) on serial probed recall of written and lip-read words for four probe types across serial position 4 to 8 from Experiment 2A.
Figure 10  Expected and observed accuracy (in percentage) on serial probed recall of spoken and lip-read words for the same-mode and different-mode conditions across serial position 5 to 8 from Experiment 2A.

Figure 11  Expected and observed accuracy (in percentage) on serial probed recall of written and lip-read words for the same-mode and different-mode conditions across serial position 5 to 8 from Experiment 2A.

Figure 12  Mean accuracy (in percentage) on serial probed recall of spoken and lip-read letters for four probe types across serial position 4 to 8 from Experiment 2B.

Figure 13  Mean accuracy (in percentage) on serial probed recall of written and lip-read letters for four probe types across serial position 4 to 8 from Experiment 2B.

Figure 14  Expected and observed accuracy (in percentage) on serial probed recall of spoken and lip-read letters for the same-mode and different-mode conditions across serial position 5 to 8 from Experiment 2B.

Figure 15  Expected and observed accuracy (in percentage) on serial probed recall of written and lip-read letters for the same-mode and different-mode conditions across serial position 5 to 8 from Experiment 2B.

Figure 16  Mean accuracy on serial probed recall (in percentage) of nameable pictures for three levels of stimulus similarity across serial positions 2 to 8 at 1 second per item from Experiment 3A.

Figure 17  Mean accuracy on serial probed recall (in percentage) of nameable pictures for three levels of stimulus similarity across serial positions 2 to 8 at 0.5 second per item from Experiment 3A.

Figure 18  Mean accuracy (in percentage) on serial probed recall of spoken words and nameable pictures for four probe types across serial position 4 to 8 from Experiment 3B.
Figure 19  Mean accuracy (in percentage) on serial probed recall of written words and nameable pictures for four probe types across serial position 4 to 8 from Experiment 3B.

Figure 20  Expected and observed accuracy (in percentage) on serial probed recall of spoken words and nameable pictures for the same-mode and different-mode conditions across serial position 5 to 8 from Experiment 3B.

Figure 21  Expected and observed accuracy (in percentage) on serial probed recall of written words and nameable pictures for the same-mode and different-mode conditions across serial position 5 to 8 from Experiment 3B.

Figure 22  Mean accuracy (in percentage) on serial probed recall of signed words and nameable pictures for four probe types across serial position 2 to 6 from Experiment 4.
Separate Short-Term Memory Buffers for Input and Output Phonology

The importance of phonology in short-term memory (STM) has been well recognized for more than three decades (e.g., Conrad, 1964; Conrad & Hull, 1964). Numerous studies have shown that both auditorily and visually presented verbal materials are remembered in phonological codes (e.g., Conrad & Hull, 1964). For example, Conrad (1972) found that for both spoken and written lists the immediate recall of rhyming letters (BCDPTV) was much worse than that for non-rhyming letters (KNVXYZ). It has also been observed that participants recalling lists of consonants are more likely to make errors that are acoustically similar to the correct items (Conrad, 1964). This phonological similarity effect is regarded as clear evidence indicating that the representations in STM, for both spoken and written items, are phonology-based. Despite this commonality between STM for spoken and written items, immediate recall is much higher for the last one or two items on auditory lists than on visual lists, which is termed the modality effect (Corballis, 1966). A similarly superior memory as for the spoken items is found when all items are presented visually and the participants read the list aloud (Conrad & Hull, 1968; Crowder, 1970; Murray, 1965).

Different theories have been proposed to account for both the phonological similarity effect and the modality effect. Due to the pronounced contribution of phonology in STM, a capacity for phonological retention is commonly postulated.
Although the modality effect strongly suggests that auditorily and visually presented items are remembered differently in phonological codes, it has usually been assumed that spoken items access the phonological buffer automatically whereas written items require an additional encoding process to be stored in the buffer (e.g., Baddeley, 1986; Penney, 1989). In other words, according to this approach, the difference between STM for spoken and written items derives from the encoding process; once items are encoded, they are maintained in the same buffer.

Baddeley and colleagues (Baddeley, 1986; Baddeley & Hitch, 1974) have proposed a model that accounts for the phonological similarity effect, but does not address the modality effect specifically. This model is a multi-component working memory model which includes a central executive component in addition to peripheral storage components. One of these peripheral storage components is the articulatory loop which is used to retain verbal materials. The articulatory loop is composed of a phonological store and an articulatory subvocal rehearsal component. The phonological store is a passive capacity that is accessible either through auditory presentation or articulatory encoding of visual materials. The articulatory rehearsal process not only serves the function of transforming visual materials into phonological codes, but also functions to keep the contents of the store refreshed so that they can be maintained for a longer time. According to this model, since both auditorily and visually presented items

are maintained in the same phonological store, phonological similarity between both kinds of materials would influence memory performance. No ready explanation for the modality effect is offered by this model, however.

A similar approach of one common phonological buffer for both auditorily and visually presented items is taken by Crowder and Morton (1969). They argued that the modality effect is due to auditory sensory memory, also called the precategorical acoustic store (PAS). This account assumes that the PAS maintains a sensory trace of the last one or two spoken items for a few seconds. Supporting this account is the suffix effect, which is the decrease in recall of the last few spoken items when they are followed by an irrelevant word (e.g., the word “zero” or “recall” following every list). The auditory trace in the PAS supplements the information maintained in a phonological buffer, which is a capacity general to different modalities of presentation. It should be noted that although there are two kinds of capacity for spoken materials in STM (i.e., the PAS and the phonological buffer), the PAS is regarded as a peripheral sensory storage capacity and only contributes to the last one or two items in the list. The phonological buffer is the main capacity to retain verbal information presented either in auditory or visual modality. Thus, this model is similar to Baddeley’s working memory model (1986) in that only one phonological buffer is proposed for all verbal materials.
Although the PAS theory offers an explanation for much empirical data, it has been called into question by several studies using other than spoken presentation. For example, Campbell and Dodd (1980) found that an auditory suffix influenced memory for lip-read items as much as memory for spoken items. Shand and Klima (1981) also reported that for congenitally deaf participants, there was an enhanced recency effect, namely better memory performance for items at the end of the list, for items in sign language relative to written words, which is similar to the better performance on spoken items for hearing people. These studies strongly argue against the claim that only auditorily presented items are remembered better than written items due to their representation in auditory sensory memory. Moreover, a short-lived PAS cannot account for Watkins and Watkins’ (1980) finding that the modality effect extended over four to six serial positions when recall was delayed by 20 seconds and the retention interval was filled by a silent copying task.

Besides the problems mentioned above, the view of one single phonological buffer in STM has been challenged by some patients’ performance reported in neuropsychological literature. Shallice and Vallar (1990) reviewed 11 patients with selective STM deficits who had impaired spans of unconnected auditory-verbal items but intact auditory word perception and speech production. Nine of these patients were reported to show a reversed modality effect. That is, the majority of the patients with
STM deficits performed better on recall of visually presented lists than on auditorily presented ones. If there were only one phonological buffer for both modalities, and the buffer is damaged as observed in the patients with spoken items, the performance on the recall of written items should be equally impaired. None of the theories with single phonological buffer could adequately account for these results.

To explain the reversed modality effect commonly shown by patients with STM deficits, Shallice and Vallar (1990) hypothesized a visual STM capacity that is exploited by these patients, resulting in their better recall of the visually presented lists (see Warrington & Shallice, 1972, for the first proposal along this line). They also speculated that the absence of the reversed modality effect in two of the 11 patients was either due to a mild impairment in the phonological buffer, or due to severe impairments in both the phonological and the visual buffers (Shallice & Vallar, 1990, p. 19). Consistent with this proposal, two patients’ immediate list recall was not influenced by phonological factors and thus was argued to rely on visual STM. For patient PV, there was no effect on the phonological similarity between the letters in a list on the number recalled, and no effect of preventing rehearsal by using articulatory suppression (i.e., continuous uttering an irrelevant speech sound) (Vallar & Baddeley, 1984). For patient KF, his errors on recall of visually presented lists were affected by visual, rather than phonological, similarity (Warrington & Shallice, 1972).
The proposal of visual STM for written items is supported by some experimental data from the neurally intact population as well. Logie, Della Sala, Wynn, and Baddeley (2000) found that participants recalled words that are phonologically similar but visually dissimilar (e.g., guy, sigh, lie) better than those that are phonologically and visually similar (e.g., fly, cry, dry), even though the former words are more frequent than the latter. They also found a visual similarity effect on the recall of letters in upper and lower cases (i.e., better recall of Dd and Qq, which are visually dissimilar than Kk and Ww, which are visually similar). Moreover, this visual similarity effect was not influenced by concurrent articulatory suppression, which in turn suggests that no subvocal encoding to phonology was involved. It should be noted, however, that this visual similarity effect was only observed when the phonological information was not sufficient to differentiate the identity of different items in a list. Specifically, for the letter memory task, since a correct response depended both on letter identity and case, the task actually required recall of both phonological and visual codes.

Although the data from neurally impaired and intact participants are consistent with the proposal of visual STM for written items, it does not rule out the possibility that visual STM is only supplementary to the dominant phonological codes for verbal materials and has a relatively small contribution. The effect of visual codes is only manifested when phonological information is not available (as in brain-damaged patients)
or not useful (as in lists composted of rhymed or homophonic items). That is, written verbal materials normally are remembered in terms of their phonology, as shown in numerous studies (e.g., Conrad & Hull, 1964), and the visual codes are exploited only as a secondary memory source.

The theories that have been discussed so far agree on the idea that there is one single phonological buffer, although some researchers argued for an additional visual STM capacity as well (e.g., Logie et al., 2000; Shallice & Vallar, 1990). In contrast, other researchers (e.g., Allport, 1984; Martin, Lesch, & Bartha, 1999) have proposed that there are two, rather than one, phonological buffers - that is, input and output buffers. Allport (1984) reported two brain-damaged patients who both showed very reduced memory span on list repetition. The two patients differed, however, in their performance on a matching span task, in which they had to judge whether two spoken lists were the same or different. (The different lists differed only in the order of two adjacent items.) Although one patient performed as poorly on the matching span task as on the list repetition task, the other patient showed excellent performance on the matching span task, at least for three- and four-word lists. To account for the difference between the two patients’ performance, Allport proposed that there was a distinction between input and output phonology in STM. Specifically, the input phonological buffer is responsible for the automatic retention of speech input, in which phonology is readily available, and the
output phonological buffer is used in speech production. The output buffer is also involved whenever a phonological code has to be retrieved from long-term memory (as in reading or picture naming) rather than being encoded from the input. To perform the matching span task, the participant only has to maintain the spoken list for a short time and compare it with the other incoming lists, thus only the input phonological buffer is needed. To perform the list repetition task, on the other hand, the participant not only has to remember the spoken list but also needs to reproduce it orally, thus both input and output phonological buffers are required. Based on this model, Allport concluded that while both patients had disrupted output phonological buffer, only the patient with poor performance on the matching span task had an impaired input phonological buffer as well.

Consistent with Allport’s (1984) proposal, Romani (1992) and Martín et al. (1999) reported two cases for whom input retention was preserved relative to output retention. The patient examined by Romani (1992) showed normal performance on those tasks that tap into the input phonological buffer. For example, he showed normal performance on the matching span task even for lists that were composed of nonwords, for which only phonological but no lexical-semantic information is available. He also performed normally on a probe task, in which the participant only has to make a yes/no judgment on the presence of a probe in a previously given list. Contrary to his normal
performance on input retention, the patient performed poorly on list repetition. To account for the selective deficit demonstrated by this patient, Romani reasoned that the patient had a preserved input phonological buffer while the output phonological buffer was disrupted.

Martin et al. (1999) reported a patient MS, who also demonstrated disrupted output phonological retention while his input phonological retention was intact. It was observed that MS performed poorly on those tasks that require verbal output, while his performance on those tasks that do not require verbal output was excellent. For example, MS's list repetition of 3- and 4-word lists was impaired (only 83% and 55% correct), but he performed comparably with the controls' mean on the rhyme probe task, in which he judged whether a probe rhymed with any of the list item. Since MS repeated nonword lists normally, it was argued that MS's deficit was not at the output phonological buffer per se. Rather, he suffered from weakened connection between the semantic system and the output phonological representations. More interestingly, MS's deficit on output phonological retention was more severe when the lists were composed of words that he had difficulty producing in a picture naming task. That is, MS's list recall performance coincided his performance on single word production - he better recalled the lists consisting of the words that he could name, whereas he poorly recalled the lists consisting of the words that he could not name. Based on this case, which provided
strong evidence for a close relation between language processing and STM performance, Martin and colleagues proposed a language-based STM model (see Figure 1; also see Martin et al., 1999, for more discussion on this issue). Indeed, the dissociation of input and output phonological retention mirrors nicely the well-established distinction between input and output phonological representations in language processing (e.g., Howard & Franklin, 1990; Monsell, 1987).

The patients’ selective impairment observed by Romani (1992) and Martin et al. (1999) cannot be attributed to higher difficulty for list repetition than that of the matching span and the probe tasks. Tested at list length where controls not at ceiling, other patients who show the opposite pattern of deficits have been reported to support a double dissociation of input and output phonological retention. Martin, Shelton, and Yaffee (1994) reported a patient EA who performed very poorly on the matching span and probe tasks, thus implying a disrupted input phonological buffer. However, EA’s spontaneous output was normal as assessed by a variety of measures including speech rate, syntactic complexity, pausing, etc. Assuming that spontaneous speech production involves planning several phonological forms simultaneously, her intact production suggested a preserved output phonological buffer while the input phonological buffer was impaired. Another patient JB, reported earlier by Shallice and Butterworth (1977), showed a similar pattern to EA.
Figure 1. A language based model of verbal short-term memory (adopted from Martin et al., 1999).
The STM model with separate input and output phonological buffers was originally motivated by some patients’ dissociative performance on input and output retention (e.g., Allport, 1984; Howard & Franklin, 1988, 1990; Martin et al. 1994; Martin et al. 1999; Romani, 1992). In addition to neuropsychological data, this proposal can also account for memory differences between auditorily and visually presented items. As mentioned above, the output phonological buffer is not only crucial to normal speech output, it also provides the capacity for verbal materials whose phonology is internally generated. Visually presented verbal items are shown to be maintained in phonological codes in STM (e.g., Conrad & Hull, 1964), even though there is no “sound” in the stimulus. To transform the written materials into their phonology, one has to produce the sound via the process of internal generation of phonology.

Cheng (1974) examined the separation between auditory and articulatory codes by manipulating phonological similarity independently across these two codes. Specifically, participants were presented with spoken strings of seven consonants alone or of these consonants paired with a redundant vowel /a/. When named as a letter, these consonants were phonologically different (F, G, K, M, R, V, Y); when paired with the vowel /a/ at the end, they sounded similar. The same consonant string was also visually presented, either with or without a redundant letter A, simultaneously with the spoken string and participants were instructed to silently mouth the written string. By manipulating
phonological similarity in the spoken and the written strings independently, Cheng discovered that the phonological similarity in both codes had detrimental effects on immediate serial recall. More interestingly, acoustic and articulatory similarity had different effects on the serial position curve. Articulatory similarity negatively affected recall at all serial positions equally whereas acoustic similarity negatively affected recall at only the last three serial positions. This finding provided further support for the idea that there are separate phonological buffers, one for acoustic (input) codes and the other for articulatory (output) codes. The two kinds of retention are differentially influenced by phonological similarity across different serial positions.

If there are indeed separate buffers for input and output phonological information, one prediction is that immediate recall capacity should be increased when both buffers are employed. This prediction is confirmed by Frick’s (1984) study in which 10-digit lists were presented to participants in four conditions: 1) all 10 digits were visually presented (VV), 2) all 10 digits were auditorily presented (AA), 3) the first five digits were visually presented with the last five digits auditorily presented (VA), and 4) the first five digits were auditorily presented with the last five digits visually presented (AV). It was found that when participants were asked to recall the last five digits first, they had greater memory span for the VA condition than that for both the AA and VV condition. Frick took this as evidence for different buffers in STM that were addressed by spoken and
written materials respectively. Because the participants were required to produce oral
responses, which might interfere with the memory for the auditorily presented digits, the
AV condition did not differ from the AA and VV conditions. However, it should be noted
that although Frick attributed the enhanced memory performance to the non-redundant
capacity in visual STM, there is no evidence indicating that the visually presented digits
are actually remembered in visual codes. Greene (1989) used a similar methodology of
mixed-modality presentation and found that both silently-read (written) items and read-
aloud items (functionally equivalent to spoken items) in one list were influenced by
phonological similarity.

Greene (1989) discovered that immediate serial recall of visually presented
consonants and words was influenced by phonological similarity, even when half of the
list items were read aloud. In other words, it is the case that written items are still
remembered phonologically despite the presence of other spoken items. Moreover, the
phonological similarity effect of both auditory and visual items was not influenced by
subvocal suppression, suggesting that the phonological representations for both
modalities were independent from rehearsal process. This finding is in sharp contrast
with the phenomenon that subvocal suppression eliminates the modality effect in single-
modality lists. To account for this discrepancy, Greene (1989) proposed that the
"modality differences on mixed lists are based on different information in memory than
that responsible for the modality effect in single-modality lists” (p. 273). Regardless of
the discrepancy, this piece of evidence argues against an important role of visual STM,
even in mixed-mode presentation. Thus, it is more sensible to argue that the non-
redundant capacity observed by Frick (1984) is provided by the output phonological
buffer rather than visual STM. The study by Greene (1989) also converges with many
earlier studies (e.g., Conrad & Hull, 1964) and underscores the importance of phonology
to verbal STM irrespective of presentation modality, although the locus of the modality
effect demands further research.

Instead of focusing on the increased capacity in STM for mixed-modality
presentation, other studies have examined the destructive effects of switching between
auditory and visual formats which results in switching between the underlying
buffers/organizations (e.g., LeCompte & Watkins, 1993; Murdock, 1967; Murdock &
Walker, 1969). For example, Murdock (1967, Experiment 3) devised a serial probed
recall paradigm, in which each item was presented either auditorily or visually and the
actual presentation modality was randomly determined. Instead of asking participants to
recall the whole list of stimuli, a probe was presented after a list and the participants were
asked to recall the item, the target, following the probe in the previous list. This
methodological change bypassed the confounding of the interference caused by responses
and the rapid decay when participants recall the first few items in a list. The mixed-
modality presentation also enabled researchers to independently manipulate the
modality of both the probe and the target. Murdock (1967) found that superior recall was
only observed when both the probe and the target were auditorily presented. Although the
memory of a visual probe with a visual target was poorly retained, in general "recall was
more probable when the probe and target items were in the same modality than when
they were in different modalities" (LeCompte & Watkins, 1993).

Also using mixed-modality presentation, Murdock and Walker (1969) found a
pronounced modality effect (i.e., higher accuracy for spoken than for written words) in
participants’ single-trial free recall. More importantly, the order of recall was organized
more by presentation modality rather than temporal sequence. These studies argued
strongly against the model of one storage capacity for both spoken and written items.
Instead, the findings are consistent with the proposal of separate input and output
phonological buffers, although Murdock (1967; Murdock & Walker, 1969) favored the
interpretation of separate auditory and visual capacities.

Penney (1989) did a thorough review on literature relevant to the modality effect
and proposed a separate-stream hypothesis to account for it. According to this hypothesis,
"the processing of auditorily and visually presented verbal items is carried out separately
in STM. The processing mechanisms are specific to either the auditory or the visual
stream, and the two streams have different properties and capabilities and represent
information in different ways." (Penney, 1989, p. 399) Specifically, Penney argued that spoken items are represented automatically in both acoustic (A) codes and phonological (P) codes while written items are only represented in P codes. The A code contains sensory information that, in the absence of interference, persists for an extended period of time and gives rise to the modality effect. On the other hand, the generation of the P codes in the visual stream is not automatic and can be disrupted. It is also hypothesized that items are organized most strongly along the temporal dimension in the auditory stream, which may also contributes to the superior serial recall of auditorily presented items. The visual stream differs from the auditory stream such that items presented simultaneously seem to be more strongly associated.

To support the separate-stream hypothesis, Penney and Butt (1986) exploited the serial probed recall paradigm with mixed-modality presentation, adopted from Murdock (1967, Experiment 3) as discussed above. In this study, the list was composed of 10 digits, and the presentation modality (visual or auditory) changed every second digit. In one condition (next item, NI), participants were instructed to recall the item that had immediately followed the probe in the original list regardless of presentation modality. In the second condition (next item in the same modality, NISM), participants were instructed to recall the item after the probe that is in the same modality. The results from this study revealed that for both spoken and written targets, participants’ recall was
highest with adjacent probe and target in the same modality, next highest with
temporally distant probe and target in the same modality, and lowest with adjacent probe
and target but in different modalities. This finding strongly suggested that auditorily and
visually presented items are stored in two different buffers, so that a probe in the same
modality with the target, even temporally two items away from the target, is a more
effective cue than a probe that is temporally close to the target but in a different modality.

At first glance, the data obtained by Penney and Butt (1986) seems to fit the
separate-stream processing hypothesis (Penney, 1989). That is, items presented in the
auditory and visual modality are indeed organized in different ways and maintained in
two different kinds of capacity. However, it is not clear that how the A and P codes
proposed by the hypothesis can explain this dissociation. In particular, the hypothesis
argues that the visual stream is based on the P codes, which are also the representations
underlying the auditory stream. Although Penney argued that storage of the A codes
differed from the PAS proposed by Crowder and Morton (1969) in terms of there being a
bigger capacity and a longer duration for A codes, it is not the case that auditorily
presented items are remembered via the A codes while visually presented items are
remembered via the P codes. Even if the observed characteristics of the auditory stream
are solely determined by the A codes, it is still not clear why an adjacent auditory probe
is a less effective cue for a visual target than a temporally distant visual probe, given that
either probe has a representation in the P code. As acknowledged by Penney (1989, p. 418): “there is little evidence for a sensory-based visual code that contributes substantially to performance in typical short-term memory tasks. Instead, it seems that participants generate a phonological or P code, which allows for rehearsal, and that performance is based mainly on this phonological code.” Thus, the privileged memory for two temporally distant written items cannot be explained by visual STM. In contrast with the failure of the separate-stream processing hypothesis, the proposal of separate input and output phonological buffers offers a ready explanation for the empirical data. That is, spoken and written items are maintained in the input and output phonological buffers respectively. A temporally adjacent probe in the other buffer is less effective than a temporally distant probe in the same buffer. This is true for both auditorily and visually presented targets, as observed by Penney and Butt (1986).

It should be noted that, different from Murdock (1967) and Penney and Butt (1986), LeCompte and Watkins (1993) favored another interpretation for the better memory performance when the probe and the target are in the same modality than when they are in different modalities. Instead of proposing different buffers/systems for spoken and written items, they described the observed dissociation in terms of physical similarity. That is, better memory results when the probe and the target are similar, whereas dissimilarity between the probe and the target would decrease recall accuracy.
To support this view, LeCompte and Watkins (1993) reported decreased between-mode relative to within-mode recall between spoken and written words, nonverbal sounds and pictures, male and female voices, and written digits presented in the upper-left and the lower-right corners of the screen. However, the same separation was not found between pictures and written words nor between written digits presented in the upper and the lower half of the screen. Whether there was a separation between spoken words and nonverbal sounds was not clear, since a memory cost caused by the modality switch between the probe and the target was only observed when the target was a spoken word but not when the target was a nonverbal sound.

Although physical similarity can account for these empirical data, it only provides a post hoc explanation but does not say too much about the organization of STM. For example, no explanation was provided as to why the visual dissimilarity between written words and pictures is a less effective organizing principle than physical characteristics of the voice like pitch between male and female voices. In contrast to the physical similarity principle, the hypothesis of separate input and output phonological buffers actually predicts the separation between spoken and written words but no separation between pictures and written words. The hypothesis of separate phonological retention is also compatible with most of the findings from LeCompte and Watkins (1993), even though it may not appear so at first glance. Specifically, LeCompte and Watkins (1993) found that,
at the presentation rate of 500 ms per item, nonverbal sounds were retained differently from pictures, which should both be stored in the output phonological buffer according to the hypothesis of separate phonological retention. However, it has been demonstrated that nameable pictures are retained both visually and phonologically at a fast presentation rate (i.e., 125 and 500 ms per item) (Coltheart, 1999). The dissociation observed by LeCompte and Watkins (1993) could be due to the separation between visual (nonverbal) and phonological (verbal) representations in STM.

Another ostensibly problematic dissociation from LeCompte and Watkins (1993) was between written digits presented at two different corners of the screen. However, this dissociation was observed when the presentation rate was fast (i.e., 200 ms per item) with subvocal suppression. It is unlikely that verbal phonological codes were the only (or even the major) representations for the STM under this manipulation. Hence, the dissociation could actually reflect the characteristics of nonverbal (visual-spatial) STM, and it is reasonable for such STM to be influenced by the spatial location of the stimuli. Since the focus of the current study is the organization of verbal phonological STM, for any experimental paradigm to provide relevant data to this issue, it is important to demonstrate that phonological coding of verbal materials is involved.

Despite the importance of phonological representations in verbal STM, they are by no means the only codes that are used to retain verbal information. Articulatory
suppression during list presentation and recall, which is assumed to prevent phonological encoding of written words, does not reduce recall to zero (e.g., Neath, 1998, p.83 and p.88). Even without suppression, there is evidence of a semantic contribution to memory span (e.g., Bourassa & Besner, 1994; Hulme, Maughan, & Brown, 1991; Martin et al., 1994; Poirier & St. Aubin, 1995; Watkins, 1977). Certainly when phonological codes are not available or useful, other representations (i.e., visual, orthographic and/or lexical-semantic codes) may make a greater contribution to recall. Materials drawing differentially on various non-phonological codes could also possibly result in differentiation between memory for same mode and different mode items. However, it is important to keep in mind that the main issue of this study is whether there is one or two phonological buffers in verbal STM. The separation along dimensions other than phonology would not be of interest here.

In summary, the evidence from neuropsychological studies and memory experiments supports the language–based verbal STM model with separate input and output phonological retention (see Figure 1). The input phonological buffer is automatically accessed by speech input and maintains this phonological information, while the output phonological buffer serves to maintain the internally generated phonological representations for speech production or immediate recall. As has been discussed above, the separate buffers have different characteristics and may be addressed
by items presented in different modalities. However, I do not intend to claim that there is separate STM capacity for every single presentation modality, from which argument the cause of memory difference is trivialized as a peripheral, sensory ("precategorical") one. Instead, I reason that the representations for input and output phonological retention play a central role in verbal STM, and they should be functionally sufficient to explain much of the STM findings reported in the literature.

The proposal of separate input and output phonological buffers is also consistent with theories of language processing, and emphasizes the close relation between the long-term linguistic representations and the representations in STM as depicted in Figure 1. This is not to say that the presentation modality has no influence on STM; rather, the nature of the presentation modality determines whether the information is stored in either the input or output phonological buffer. The documented modality effect simply reflects the distinction between input and output phonological (but not between auditory and visual) retention in STM.

In the current study, the proposed separation between input and output phonological retention in STM was further investigated by using different presentation modalities. In addition to spoken and written words, silent lip-read films, nameable pictures, and words in American Sign Language (ASL) were employed. These verbal materials are interesting because they are all stored in phonological codes, as spoken
items, when presented in a short list and tested for immediate recall (see discussion below). However, for those materials that do not contain acoustic information within the stimuli, the phonology is derived from internal generation, and hence should be stored in the output phonological buffer. On the other hand, for those materials that contain the natural phonology to the language users (e.g., speech for hearing people and ASL for deaf people), the phonological information is readily provided by the stimuli, and hence should be stored in the input buffer.

Memory for silent lip-read films have been examined extensively in the past two decades (e.g., Campbell & Dodd, 1980; Greene & Crowder, 1984). It seems intuitively plausible that lip-read films are stored in terms of phonological codes derived from the lip movements rather than relying on visual codes, since words in lip-read films look extremely similar and are highly confusable. Moreover, the information in lip-read words, despite being visually presented, is revealed over time as is the acoustic information in spoken words. Campbell and Dodd (1980, 1982) provided empirical evidence for this claim. It was found that silently lip-read lists, like auditory lists, showed recency compared to written lists and this effect was abolished by a spoken suffix. It was also found that a lip-read suffix reduced the recency effect for both spoken and lip-read lists. In contrast, a written suffix had no effect on both spoken and lip-read lists.
Despite the commonality that spoken and lip-read words shared, Campbell and colleagues (Campbell, 1987; Campbell, Garwood, & Rosen, 1988) argued that the memory codes for lip-read and spoken lists were not identical as indicated by the differential effects produced by a lip-read suffix and a spoken suffix. A spoken suffix greatly interfered with the recency effect from heard, lip-read, and mouthed lists, whereas a lip-read suffix only decrease the recency effect from these lists to a less pronounced degree (Greene & Crowder, 1984). Therefore, it was reasoned that both lip-reading and auditory presentation give rise to a phonetic code, but the phonetic representation underlying a lip-read stimulus is underspecified, abstract, and amodal (Campbell, 1987, p. 146). This conclusion is consistent with the proposal of separate input and output phonological retention in STM, according to which lip-read items, whose phonology need to be generated internally, are maintained in the output phonological buffer, whereas spoken items are maintained in the input phonological buffer. It should be noted, however, that Campbell (1987, p. 146) conceptualized this phonetic representation to be accessed by spoken items as well, and it "was equivalent to the phonological input buffer proposed by Monsell (1987)" (Campbell, 1980, p. 146). Based on the fact that lip-read words are remembered phonologically, the current study also examined the issue of whether spoken and lip-read items are retained in the input and output buffers respectively, or both are retained in the same buffer as Campbell (1987) has proposed.
Nameable pictures were included in the current study according to the same logic. It has been repeatedly shown that nameable pictures are remembered in phonological rather than visual codes. For example, Schiano & Watkins (1981) found that both memory span for pictures of common objects and for the names of these objects was influenced by phonological similarity and the length of the names, as well as by the participant's engaging in "irrelevant" vocalization (concurrent articulation) during item presentation. Brandimonte, Hitch, and Bishop (1992) also demonstrated that memory span for pictures was decreased by simultaneously verbal suppression. These results clearly indicated that nameable pictures, just like other verbal materials, are retained phonologically but not visually in STM. Since there is no acoustic information in picture presentation, the phonological representations of picture names need to be generated before they can be held in STM. Thus, it is hypothesized that nameable pictures, as well as written and lip-read words, are stored in output phonological buffer.

In addition to lip-read words and nameable pictures, words presented in ASL were also employed in the current study. It has been well documented that ASL is a natural human language possessing its own phonology and syntax just like any spoken language (Corina & Sandler, 1993). Signs in ASL have been analyzed into four basic phonological components: handshape, location, movement, and palm orientation (Wilson & Emmorey, 1997). Wilson and Emmorey (1997, 1998) provided clear evidence that
deaf participants used the phonology of ASL to code verbal materials in STM.

Specifically, they reported a robust phonological similarity effect with signs that was not eliminated by concurrent manual articulation (i.e., analogous to articulatory suppression for hearing people). When signs were presented to participants as static pictures, however, the phonological similarity effect was abolished by concurrent manual articulation. These results suggest that encoding ASL as natural linguistic input is independent from signed rehearsal, but the processes engaged in rehearsal are used to internally generate signs, which is necessary for remembering pictures of ASL. The findings mimic the findings from hearing people – that is, concurrent “articulatory” suppression affects memory for verbal materials that require phonological encoding and are stored in the output phonological buffer (e.g., written words for hearing people and ASL depicted in pictures for deaf people), whereas natural linguistic input (i.e., speech for hearing people and signs for deaf people) is not influenced by such suppression.

Based on the previous studies, it is reasonable to hypothesize that ASL, the natural linguistic input for congenitally deaf people who are native signers, is retained in an input phonological buffer, as speech is for hearing people, although the phonology for ASL is very different from that for speech. In contrast to this commonality, ASL differs from speech in the modality of transmission. That is, ASL possesses no acoustic information but only spatial-gestural information that is perceived visually, as for lip-read
items and namable pictures. If sensory presentation modality determines the
organization of verbal STM, the memory performance for ASL would be more similar to
that for other visually presented materials (e.g., written items, lip-read items, and
nameable pictures). On the other hand, if signed input is coded into an input phonological
buffer despite its visual presentation, then signed input should show a separation from
written words, similar to the separation between spoken and written words. The
prediction of whether the STM for signed words would show a separation from that for
spoken words for hearing signers is less straightforward. It could be the case that both
spoken and signed words are stored in the input buffer, so there is no separation between
them. It could also be the case, however, that because the phonological characteristics of
speech and sign are so different there is a separation between them. (Also, one might
claim an evolutionary basis for a capacity for the retention of input phonology for speech,
but not for sign.) Because of this ambiguity in the predictions for speech and sign,
hearing users of sign were not tested on speech and sign. Instead, the research focused on
deaf signers and their retention of sign versus other visual materials (i.e., written words
and pictures). The results of Wilson and Emmorey (1997) suggest a separation between
the codes involved in perceiving signs and producing them (for instance, in naming a
picture) because they showed that manual suppression affected recall of static pictures of
signs but not naturally perceived signs. Given the unique characteristics of ASL,
examining memory for ASL can provide a means to adjudicate between a modality based organization versus input and output phonology organization of verbal STM.

**Experiment 1**

In this experiment, spoken and written items were employed. I first sought to replicate the phonological similarity effect with visually presented lists to establish the fact that written verbal materials are remembered in phonological codes. Then I adopted the probed serial recall paradigm utilized by Murdock (1967) and Penney and Butt (1986) to replicate the STM performance that implicates the separate buffers for spoken and written items respectively in the mixed-modality lists. This paradigm with different instructions (NI and NISM) was preferred than the one used by LeCompte and Watkins (1993), because the latter paradigm can only inform us the existence of a cost associated with modality switch but not the locus of such a cost. By employing two kinds of instructions, the memory from a same-modality distant probe and a different-modality adjacent probe can be compared directly. If the detrimental STM is caused by perceptual discontinuity, a distant probe in the same modality may not be an effective cue since the modality changed twice between the probe and the target. If the detrimental STM is caused by retention in different buffers, on the other hand, a same-modality distant probe should result in better memory performance than a different-modality adjacent probe, since the former probe is in the same buffer with the target, whereas the latter probe is in
a different buffer from the target. Different presentation rates and capital letters, in addition to digits, were administered to examine the generality of the findings by Penney and Butt (1986).

Experiment 1A

Method

Participants. Eighteen undergraduate students from Rice University participated in this experiment in exchange for course credit. They were all native English speakers and had normal or corrected-to-normal vision.

Materials and design. Three groups of letters were selected such that phonological and visual similarity among the letters within each group was manipulated independently. For phonologically similar and visually dissimilar (PSVD) lists, the capital letters B E G R S T V Z were used. Six out of these eight letters rhymed with each other and they were all visually dissimilar according to Gibson’s (1969) chart of distinctive features and Townsend’s (1971) confusion matrix of capital letters. The average of shared distinctive features for any two PSVD letters was 1.25 and the mean confusibility was 1.61. For phonologically dissimilar and visually similar (PDVS) lists, capital letters E F I K L M R W were used. None of these letters rhymed with each other. The average of shared distinctive features for any two PDVS letters was 2.32 and the mean confusibility was 2.41. As for phonologically dissimilar and visually dissimilar
(PDVD) lists, capital letters H L O R S U Y Z were used. None of them rhymed with any other PDVD letters. The average of shared distinctive features for any two PDVD letters was 0.82 and the mean confusibility between two PSVD letters was 2.04.

Fifty-six lists were generated in advance for each group of the letters, and the three sets of lists fulfilled following constraints: Each list was composed of 8 capital PSVD, PDVS, or PDVD letters, depending on the condition. Every letter only appeared once in a list, and the serial position of that letter was randomly assigned across different lists. One of the 8 letters was selected as the probe and the letter next to the probe was selected as the target, such that every letter was selected once as the probe in serial position 1 to 7 and as the target in serial position 2 to 8 within the 56 lists for each condition. That is, each combination of the probe letter, the target letter, and the serial position of probe and target was tested once for each group of letters.

The 56 lists using the same set of letters were grouped in one block and were randomly presented within the block, therefore every participant received the same set of lists with a different order. The sequence of the administration of the three blocks was counterbalanced across participants.

Procedure. Every participant was tested individually in a quiet room. A Macintosh computer presented the stimuli on the monitor and recorded the responses. In each trial, there was a fixation point appearing in the center of the computer screen
accompanied by a beep for 800 msec. After a blank of 200 msec, 8 capital letters in a list were presented successfully in the center of the screen. Every letter was presented for 800 msec and was followed by a blank for 200 msec. Seven-hundred msec after the end of the last letter, a word “recall” was presented on the screen with a probe letter below it. Participants were instructed to press one of eight pre-defined keys (number key 1 to 8) to indicate the letter that comes right after the probe according to the previously presented list. The probe letter stayed on the screen until a key was pressed. After making a response, participants pressed the space bar to initiate the next trial.

Every participant received 168 trials in total which were grouped in 3 blocks corresponding to the three different conditions. A 2-minute break was allowed between two blocks. There were 7 practice trials, which preceded the experiment, whose procedure was identical to the experimental trials, except that participants received feedback on the correctness of their response in each practice trial. No feedback was provided during the experiment.

Results and Discussion

Both accuracy and reaction time of participants’ responses were recorded, but only accuracy was analyzed and reported in this and the following experiments, since the instruction emphasized accuracy rather than speed. The probed recall accuracy in the three conditions across different serial positions of the target was computed and reported
in Figure 2. A two-way analysis of variance (ANOVA) was conducted on two within-participant variables: Three levels of phonological and visual similarity (PSVD, PDVS, and PDVD) and 7 levels of serial positions of the target (2 to 8). All effects reported as significant in this and the following experiments had a probability value less than .05, unless otherwise noted.

*Figure 2.* Mean accuracy on serial probed recall (in percentage) of written letters for three levels of stimulus similarity across serial positions 2 to 8 from Experiment 1A.

The ANOVA revealed a significant main effect of phonological and visual similarity, $F(2, 34) = 17.14, MSE = 57.67, p < .001$, and a significant main effect of serial position of the probe, $F(6, 102) = 3.71, MSE = 10.17, p < .002$. The interaction between stimulus similarity and serial position was not significant, $F(12, 204) = 1.24, MSE = 1.79, p > .05$. Further analyses were performed to determine the source of the difference for
both main effects. With regard to the main effect of serial position, one-item primacy and recency effects were obtained with memory accuracy of the first and the last letter was significantly better than that of the other serial positions. Since the main interest of the current study was the effect of phonological and visual similarity, and there was no interaction between stimulus similarity and serial position, further discussion was restricted to the effect of phonological and visual similarity.

The memory accuracy of PSVD lists (64% correct) was significantly worse than that of PDVS (80% correct) and PDVD (76% correct) lists, both $p < .001$. Although the difference between the memory for PDVS and PDVD lists was not significant ($p > .05$), it should be noted that, unlike the results reported by Logie et al. (2000), the visually dissimilar lists were actually remembered worse than the visually similar ones. As suggested earlier, their finding may have been due to the lack of useful phonological information. In the present experiment, distinctive phonological information was available for both visually similar and dissimilar lists. Since the phonological information was useful for retaining the list, the supplementary visual codes were not exploited and thus exerted no influence on the memory performance in the current experiment.

The result of this experiment replicated previous studies in that visually presented letters are remembered in phonological but not visual codes in STM. In the present experiment, phonological and visual similarity of capital letters was manipulated
independently. The probed recall was equivalent for lists composed of phonologically
dissimilar letters, regardless of visual similarity of the letters within the lists. On the other
hand, memory was worse for the lists that contained phonologically similar but visually
dissimilar letters. This finding clearly indicated the contribution of phonology in STM for
written items.

Experiment 1B

Given the fact that visually presented letters are retained in phonological codes as
demonstrated in Experiment 1A, the next question is whether these written letters are
stored in the same phonological buffer with spoken letters. Specifically, I hypothesized
that there are separate input and output phonological buffers in STM and they serve to
maintain spoken and written verbal materials respectively. Penney and Butt (1986) have
shown that spoken and written digits are remembered in two separate “streams”,
therefore a temporally distant cue in the same modality is more effective than a
temporally adjacent cue in a different modality. The current experiment sought to
replicate this finding using the same paradigm. The procedures employed by Penney and
Butt (1986) were used with phonologically and visually dissimilar capital letters as
stimuli. The presentation rate was changed from 0.5 second per item to 1 second per item
to be comparable with that used in Experiment 1A.

Method
Participants. Sixteen undergraduate students from Rice University participated in this experiment in exchange for course credit. They were all native English speakers and had normal or corrected-to-normal vision and hearing.

Materials and design. Two hundred and eighty lists were prepared in advance for the current experiment. These lists consisted of ten phonologically and visually dissimilar (PDVD) letters, A H L N O R S U Y Z, which were ordered in a random sequence within each list. The presentation modality within a list changed after every two letters, resulting in four list types in which the first four letters could be AVVA, AAVV, VAAV, or VVAA (A indicated auditory presentation and V indicated visual presentation). Two list types (AAVV and VVAA) contained 5 letters in each modality, whereas the other two list types (AVVA and VAAV) contained 6 letters in one modality and 4 letters in the other modality.

A probed-recall paradigm was employed in which only target position 4 to 10 were tested. Specifically, for each list type, each letter appeared in every tested serial position as the target one time. The combination of four list types, ten letters, and seven serial positions of the target resulted in 280 lists in total. Among these trials, each letter was selected as the probe the same number of times, and each letter was paired with different serial positions. The choice of the target letter was as variable as possible. The 280 lists were evenly divided into two sets, so that each letter was selected as the target at
each serial position from 4 through 10 within each list type equally often. Each of the
list sets was paired with one of the two instructional conditions. In the “next item (NI)”
instructional condition, participants were instructed to recall the target immediately
following the probe, regardless of the presentation modality. In the “next item in the same
modality (NISM)” instructional condition, participants were instructed to recall the target
following the probe that was also presented in the same modality as the probe. The
combination of the two list sets and the instructional conditions was counterbalanced
across participants.

Procedure. In this experiment, each list was presented in a mixed-modality
fashion so that the presentation modality changed after every two letters. Four list types
were created in which the first two letters exhausted all possible combinations of two
presentation modalities (i.e., AA, AV, VA, and VV). The trials of this experiment had a
similar timeline with that of Experiment 1A, except that 10 letters rather than 8 were
presented successively. In each trial, a fixation point first appeared in the center of the
computer screen accompanied by a beep for 800 msec. Then 200 msec after the fixation
point, 10 letters in a list were presented successively either in the center of the screen or
through the computer’s speakers, depending on the pre-determined presentation modality.
The duration of each spoken letter was between 505 msec and 749 msec, with the mean
of 612 msec. The duration of each written letter remained to be 800 msec and was
followed by a blank of 200 msec as in Experiment 1A. The presentation rate of the list, irrespective of the presentation modality, was fixed at 1 second per item. Fifteen hundred msec after the beginning of the last letter, the word “recall” appeared on the screen simultaneously with the probe in its original presentation modality. Participants pressed one of the 10 pre-specified keys (number key 1 to 0) to indicate their response. The correct response for each trial depended on the type of the list in which the 10 letters were presented and also the instructional condition (NI or NISM). When the probe was a written letter, it stayed on the screen until a key was pressed. Then participants pressed the space bar to initiate the next trial.

The two list sets paired with the two instructional conditions were administered in two different days, around one week apart. The order of two instructional conditions was counterbalanced across participants. For each set, the 140 lists were divided into five blocks in which four list types appeared equally often. A two-minute break between blocks was allowed. There were 8 practice trials before the experiment. The procedures for these trials were the same as the experimental trials, except that participants received feedback about the correctness of their responses. No feedback was provided for the experimental trials.
Results

Reaction time and accuracy of responses were measured for all experiments, but only the data for accuracy was analyzed. Memory accuracy under different conditions was computed across serial positions of the target, and depicted in Figure 3 for both target modalities. A 2 (target modality) x 4 (probe type) x 7 (serial position) within-participant ANOVA was conducted. Among the four probe types, there were three "close" conditions in which the probe was immediately prior to the target while there was one "distant" condition in which the probe was three items before the target. In two of the three close conditions the probe and the target were presented in the same modality ("same-mode under the NI instruction" and "same-mode under the NISM instruction"). In the other close condition, the probe and the target were in different modalities ("different-mode"). The probe and the target were in the same modality ("same-mode but distant") in the distant condition. Because the probe type and instruction were not orthogonal, they were not treated as two separate factors.

The ANOVA results revealed that the main effect of target modality was highly significant, $F(1,15) = 22.19, \text{MSE} = .35, p < .001$, replicating the well-known modality effect such that spoken targets were remembered better than written targets. The main effect of different probe type was also highly significant, $F(3,45) = 17.56, \text{MSE} = .17, p < .001$, indicating that the probes in different conditions were of different effectiveness as a
Figure 3. Mean accuracy (in percentage) on serial probed recall of spoken and written letters for four probe types across serial position 4 to 10 from Experiment 1B.
cue for the recall of the target. Direct comparisons between different probe types indicated that while the two same-modality adjacent conditions had equivalent accuracy \( (p > .96) \) as did the different-modality adjacent condition and the same-modality distant condition \( (p > .36) \), the two same-modality adjacent conditions resulted in significantly better memory performance than the other two conditions (all \( ps < .003 \)).

Since the interaction between target modality and probe type was also significant, \( F(3,45) = 6.12, MSE = .14, p = .001 \), further analyses on each target modality were conducted separately. For auditory targets, the mean accuracy for the four probe types were as follows: 76.6% for same-mode under the NI instruction (AA(NI)), 81.1% for same-mode under the NISM instruction (AA(NISM)), 48.6% for same-mode but distant (A..A), and 49.5% for different-mode (VA). The two same-mode and close conditions, AA(NI) and AA(NISM), did not differ from each other \( (p > .18) \), and the other two probe types, A..A and VA, showed no difference \( (p > .88) \), either. The other differences were all significant at \( p < .001 \). For visual targets, the mean accuracy for the four probe types were as follows: 52.5% for same-mode under the NI instruction (VV(NI)), 47.9% for same-mode under the NISM instruction (VV(NISM)), 37.0% for same-mode but distant (V..V), and 44.1% for different-mode (AV). Although the memory for the V..V condition seemed to be worse than that for the other probe types, it did not differ significantly from the AV condition \( (p > .26) \). The VV(NI) condition was significantly better than the V..V
(p < .025) and the AV (p < .004) conditions, while the VV(NISM) condition was only significantly better than the V..V condition (p < .012). All the other differences were not significant.

The main effect of serial position was found to be significant, \( F(6,90) = 18.54, MSE = .06, p < .001 \), as were the interactions of this variable with both target modality, \( F(6,90) = 2.46, MSE = .05, p < .03 \), and probe type, \( F(18,270) = 1.76, MSE = .04, p < .03 \). The main effect of serial position was due to the recency effects seen for both target modalities. The serial position by target modality interaction reflected a flatter serial position curve for the visual modality as clearly illustrated in Figure 3. Also, recall accuracy for spoken targets showed a rising trend except for serial position 4 whereas this was true only for the last two serial positions for written targets. The three-way interaction between target modality, probe type, and serial position was also significant, \( F(18,270) = 3.21, MSE = .04, p < .001 \). Further analyses revealed that for spoken targets, no recency effect was observed for the different-mode (VA) condition. On the other hand, for written targets, the recency effect was absent only for the temporally distant (V..V) condition.

**Discussion**

The main interest of this experiment was to examine whether a same-mode but distant probe was a more effective cue than an adjacent probe in a different modality. The
results from the current experiment did not replicate Penney and Butt’s (1986) findings. In particular, the same-mode but distant probe did not produce better recall than the different-mode probe for both target modalities. It should be noted, however, that the same-mode but distant probe was not significantly less effective than the different-mode probe, either. If there were only one buffer for both spoken and written items, one would expect that the effectiveness of a cue would be heavily influenced by the temporal distance between the probe and the target. Although for both spoken and written targets the same-mode but distant probe produced numerically worse recall accuracy, it behaved no worse than the different-mode probe statistically.

For the three probe types in which the probe and the target were adjacent to each other, the temporal distance between the probe and the target was the same. Since the different-mode probe was less effective than the same-mode probe from both instructional conditions, it was reasonable to hypothesize that there was “memory cost” associated with the modality switch of list presentation. This memory cost could result from one or the combination of several sources, such as tuning to different input modality, attentional discontinuity, and storing incoming information in two different buffers. Although the memory cost did not necessarily implicate separate buffers for spoken and written verbal materials, it should still be the case that a same-mode but distant probe suffered more memory cost than an adjacent different-mode probe, unless
there was some modality-specific capacity that would store closely the probe and the
target in the same modality without including a temporally adjacent probe in a different
modality.

*Experiment 1C*

In Experiment 1B, no strong evidence for the separation between the input and
output phonological buffers in verbal STM was obtained. The results were inconsistent
with the findings from Penney and Butt (1986). Two methodological differences, namely
the stimuli (letters vs. digits) and presentation rate (1 sec per item vs. 0.5 sec per item),
might have resulted in the discrepancy in the data. In Experiment 1C, I conducted an
exact replication of Penney and Butt (1986) to establish the reliability of the previous
findings.

*Method*

*Participants.* Sixteen undergraduate students from Rice University participated in
this experiment in exchange for course credit. They were all native English speakers and
had normal or corrected-to-normal vision and hearing.

*Materials, design, and procedures.* The current experiment is similar to
Experiment 1B, except that digits from 0 to 9 were used instead of capital letters. The
duration of each spoken digit was between 337 msec and 453 msec, with the mean of 403
msec. Another procedural change was the presentation rate of the successive digits. In
each trial, a fixation point first appeared in the center of the computer screen accompanied by a beep for 500 msec. Then 200 msec after the fixation point, 10 digits in a list were presented successively either in the center of the screen or by the computer’s speakers, depending on the pre-determined presentation modality. The presentation rate of the list, irrespective of the presentation modality, was fixed at 0.5 second per item. One second after the beginning of the last digit, the word “recall” appeared on the screen simultaneously with the probe in its original presentation modality. The other procedures were the same as in Experiment 1B.

Results

The same analyses were conducted as in Experiment 1B. The memory accuracy under different conditions was computed across serial positions of the target, and depicted in Figure 4 for both target modalities. The 2 (target modality) x 4 (probe type) x 7 (serial position) within-participant ANOVA revealed a significant main effects of target modality, $F(1,15) = 20.64, MSE = .37, p < .001$, of probe type, $F(3,45) = 35.68, MSE = .12, p < .001$, and also of serial position, $F(6,90) = 27.63, MSE = .07, p < .001$. As in Experiment 1B, a robust modality effect and a recency effect were again found. Different from Experiment 1B, however, direct comparisons between different probe types indicated that although the accuracy for the two same-modality adjacent conditions remained equivalent ($p > .96$), all other comparisons were significant. That is, both the
Figure 4. Mean accuracy (in percentage) on serial probed recall of spoken and written digits for four probe types across serial position 4 to 10 from Experiment 1C.
same-modality adjacent conditions had higher accuracy than either of the other two conditions (all $ps < .001$). More importantly, a temporally distant cue in the same modality with the target resulted in better recall than a temporally adjacent cue in a different modality ($p = .019$). This pattern basically replicated the findings of Penney and Butt (1986).

Since the interaction between target modality and probe type was also significant, $F(3,45) = 4.91$, $MSE = .13$, $p < .005$, separate analyses were performed on each target modality. For spoken targets, the mean accuracy for the four probe types were as follows: 84.5% for AA(NI), 87.3% for AA(NISM), 64.6% for A..A, and 47.0% for VA. There was no difference between the two adjacent same-mode conditions ($p > .33$). Every other comparison was significant (all $ps < .009$). In particular, the same-mode but distant probe produced better recall than the different-mode but adjacent probe. The same pattern was observed for written targets, as the recall accuracy for the four probe types were 60.7% for VV(NI), 57.7% for VV(NISM), 49.5% for V..V, and 41.8% for AV. The two adjacent same-mode conditions showed no difference as for spoken targets ($p > .41$). The same-mode but distant probe also resulted in numerically better memory than the different-mode probe; however, this difference did not reach the level of significance ($p > .30$).

The significant interaction between probe type and serial position ($F(18,270) = 2.19$, $MSE = .04$, $p < .004$) should be understood in light of the significant three-way
interaction between target modality, probe type, and serial position ($F(18.270) = 2.29, MSE = .03, p < .003$). For spoken targets, all probe types produced a significant recency effect ($ps < .006$). This was true for written targets ($ps < .001$) except that the temporally distant probe (V..V) only produced a marginally significant recency effect ($p < .056$).

Although the difference between the V..V and AV conditions was not significant, visual inspection revealed an obvious difference in the trend of these two conditions across different serial positions. A trend analysis on the V..V and AV conditions confirmed this observation and indicated a significant difference between the trend, $L_{V..V} = 0.79, L_{AV} = 2.21, t(15) = 2.31, p < .035$. Consistent with this result, the V..V condition resulted in numerically better memory performance than the AV condition except for the very last two serial positions.

Discussion

Using the same methodology, the current experiment replicated the findings of Penney and Butt (1986). Specifically, the two adjacent same-mode conditions resulted in equivalent recall accuracy for both target modalities, indicating that different instructions did not influence participants’ performance level. The worse memory in the different-mode condition than in the two adjacent same-mode conditions again indicated the cost from modality switch. The most important implication of the present data was that the same-mode but temporally distant probe was a more effective cue than the different-
mode adjacent probe. This result suggested that spoken items were retained in a
different buffer from written items so that digits in the same modality, albeit temporally
distant, were better related than with temporally adjacent digits in a different modality.

Although the current data were consistent with the hypothesis of separate input
and output phonological retention, there was also the other possible account that might
give rise to this pattern. Participants' accuracy in the current paradigm depended on the
probability of remembering not only the target but also the probe. Due to the robust
modality effect (that is, better recall of auditory items), all else being equal, a target
would be remembered better following a spoken probe than following a written probe.

For the spoken target, the worse performance for the VA condition than the A..A
condition might simply be caused by the inferior memorability of the written probe
relative to the spoken probe. The data from the written target argued against this
possibility, however. If there were no cost associated with the modality switch, better
performance in the AV condition than the V..V condition would be expected based on the
modality effect and also the closer temporal distance. Contrary to this prediction, memory
for the AV condition was worse than that for the V..V condition, although the difference
failed to reach significance. The lack of significance might have resulted from the more
memorable spoken probe and hence the improved memory for the AV condition.
Expected vs. observed memory performance. To formally test the argument that there was memory cost caused by the modality switch of list presentation, I computed the expected recall accuracy for targets following an adjacent probe at different serial positions for each participant. As discussed above, the logic behind this computation was that to make a correct response in the current probed serial recall paradigm, the participant had to remember both the probe and the target. That is, the recall accuracy for the target at a certain serial position N was actually the product of the recall accuracy for the probe at the serial position N-1 and for the target at the serial position N. Assuming that the recall accuracy for the probe and for the target at a certain serial position was the same, the expected recall accuracy for the spoken and the written items at the serial position 4 through 10 was separately obtained from the following equation:

$$RA_N = \sqrt{\frac{RA_N^{(NI)} + RA_N^{(NISM)}}{2}}$$

where $RA_N$ was expected recall accuracy at the serial position N, $RA_N^{(NI)}$ was observed recall accuracy from the same-mode (NI) condition at the serial position N, and $RA_N^{(NISM)}$ was observed recall accuracy from the same-mode (NISM) condition at the serial position N. It should be noted that this estimation of $RA_N$ would be valid if the recall accuracy for the serial position N-1 and N were identical. Based on the empirical data obtained in the current study, however, I realized that the recall accuracy through the
serial positions 4 to 10 was usually greater than that for the previous serial position. As a result, the expected $RA_N$ would be an underestimation.

After obtaining the expected recall accuracy for different serial positions, the expected recall accuracy for both the same-mode and the different-mode conditions through the serial position 5 to $10^1$ was computed via the following equation:

$$RA_N(XY) = RA_{N-1}(X) \times RA_N(Y),$$

where $RA_N$ was expected recall accuracy at the serial position $N$ when the probe was in $X$ modality and the target was in $Y$ modality, $RA_{N-1}(X)$ was expected recall accuracy for the probe in $X$ modality at the serial position $N-1$, and $RA_N(Y)$ was expected recall accuracy for the target in $Y$ modality at the serial position $N$. The derived expected memory accuracy, as well as the observed performance, was depicted in Figure 5.

To compare the accuracy between the expected and the observed memory performance for both the same-mode and the different-mode conditions, a $2 \times 2 \times 6$ (expected vs. observed) x (same-mode vs. different-mode) x (serial position 5 to 10) within-participant ANOVA was conducted for each target modality separately. The observed memory performance for the same-mode condition was the mean of the observed accuracy for the same-mode (NI) and same-mode (NISM) conditions. The results of the ANOVA revealed that, for both spoken and written targets, the observed performance differed significantly from the expected performance ($F(1,15) = 17.47$, $MSE = .03$, $p =$
Figure 5. Expected and observed accuracy (in percentage) on serial probed recall of spoken and written digits for the same-mode and different-mode conditions across serial position 5 to 10 from Experiment 1C.
.001 for spoken targets, and $F(1,15) = 18.05, MSE = .05, p = .001$ for written targets).

The difference between the same-mode and the different-mode conditions was significant for spoken targets, $F(1,15) = 45.74, MSE = .19, p < .001$, while it was not significant for written targets, $F(1,15) < 1, MSE = .09, p > .41$. The effect of serial position was significant for both target modalities ($F(5,75) = 28.44, MSE = .03, p < .001$ for spoken targets, and $F(5,75) = 22.72, MSE = .09, p < .001$ for written targets). Although the same-mode and the different-mode conditions did not differ significantly for written targets, the highly significant interaction between this variable and the variable of expected vs. observed performance for both spoken targets ($F(1,15) = 30.13, MSE = .03, p < .001$) and written targets ($F(1,15) = 58.07, MSE = .03, p < .001$) indicated that further analyses were needed.

To understand the significant interaction obtained above, the 2 (expected vs. observed) x 6 (serial position) within-participant ANOVA was performed on the same-mode and the different-mode conditions separately for both target modalities. For the same-mode condition with spoken targets (AA), the difference between the expected (85%) and the observed (87%) performance, albeit very small, was significant, $F(1,15) = 33.00, MSE = .0006, p < .001$. This was actually consistent with the computation I employed to derive the predicted accuracy. As I pointed out earlier, the expected memory performance consistently underestimated the accuracy. Since it was a close but biased
estimation, the small deviation from the observed accuracy was significant. As depicted in Figure 5, the effect of serial position was significant, $F(5,75) = 25.32, MSE = .02, p < .001$, again replicating the standard recency effect. The interaction between the two variables was also significant, $F(5,75) = 4.12, MSE = .003, p = .002$, indicating that the biased underestimation of the predicted accuracy was not as severe as for the later serial positions.

Contrary to the results for the same-mode condition, the ANOVA on the different-mode condition for spoken targets (VA) revealed a different pattern. The difference between the expected (65%) and the observed performance (48%) was still significant, $F(1,15) = 23.49, MSE = .06, p < .001$, but with the opposite direction. That is, the observed memory accuracy was much worse than the expected one, even though the expected accuracy tended to underestimate performance. An effect of serial position was detected, $F(5,75) = 14.00, MSE = .04, p < .001$. The significant interaction between two variable, $F(5,75) = 3.62, MSE = .03, p = .005$, indicated that the memory cost caused by the modality change was only present for the serial positions from 7 to 10.

The results for written targets showed a similar pattern as those for spoken targets. Specifically, the ANOVA for the same-mode (VV) condition revealed that the expected performance (57%) was significantly lower than the observed one (62%), $F(1,15) = 20.86, MSE = .006, p < .001$, although the difference was very small. A robust serial
position effect ($F(5,75) = 17.95, MSE = .05, p < .001$) and the significant interaction between two variables ($F(5,75) = 2.37, MSE = .006, p = .047$) were found. For the different-mode (AV) condition, on the other hand, the significant effect of memory accuracy indicated that the expected performance (69%) was much better than the observed performance (45%), $F(1,15) = 35.78, MSE = .08, p < .001$. The effect of serial position was significant, $F(5,75) = 19.91, MSE = .05, p < .001$. There was no interaction between the two variables for written targets in the different-mode condition.

In summary, the comparison between the observed and the expected memory performance for spoken targets unambiguously argued against the idea that the inferior accuracy for the VA condition than the A..A condition was caused by the less memorable written probe. The observed VA performance was significantly worse than the expected VA performance, which tended to underestimate memory accuracy and already took the memorability of the written probe into account. The same argument could be made for the results from written targets. Although the probe for the AV condition was more memorable, the observed accuracy for this condition was still worse than the expected one. The evidence from these analyses provided strong support for the memory cost caused by modality switch.

According to the model of separate input and output phonological retention, this cost was associated with the change between two buffers. It was also possible, as it has
been discussed earlier, that the cost implicated some other perceptual discontinuity and/or attentional blink caused by the change of presentation format. These alternatives were addressed in the following experiments, in which spoken and written materials were mixed with lip-read items and nameable pictures in Experiment 2 and 3, respectively. If the memory cost observed in the current experiment were simply caused by the change of presentation format, it was expected to discover such cost whenever the list changes from one format to another format. On the other hand, if the memory cost was actually resulted from the switch between input and output phonological buffers, then it was expected to only observe such cost between spoken items and items in other formats (i.e., lip-read items and nameable pictures), but not between written and lip-read items, for example.

*Experiment 1D*

As mentioned above, Experiment 1B and Experiment 1C differed in the stimuli used and the presentation rate. Since both digits and letters are very familiar verbal materials, it seems unlikely that the stimuli played an important role in the different memory performance between the two experiments. Instead, it is speculated that the faster presentation rate in the Experiment 1C is crucial to the demonstration of separate input and output phonological buffers, due to the rapid decay and the transient trace of STM storage, especially for written items. To verify the importance of the presentation rate, in Experiment 1D I again used the capital letters as stimuli and employed the faster
presentation rate at 2 items per second. A similar pattern of results as in Experiment 1C was expected.

Methods

Participants. Twenty undergraduate students from Rice University participated in this experiment in exchange for course credit. They were all native English speakers and had normal or corrected-to-normal vision and hearing.

Materials, design, and procedures. The current experiment used the same list materials and design as Experiment 1B, except that the spoken letters were recorded at a shorter duration such that every spoken letter lasted 453 msec. The procedures of the current experiment were identical to that of Experiment 1C. Specifically, the presentation rate of 0.5 second per item was administered.

Results

The same analyses were conducted as in Experiment 1C. The memory accuracy under different conditions was computed across serial positions of the target, and depicted in Figure 6 for both target modalities. The data revealed a very similar pattern with that of Experiment 1C. The 2 (target modality) x 4 (probe type) x 7 (serial position) within-participant ANOVA confirmed the significant main effects of target modality, \( F(1,19) = 47.50, \text{MSE} = .31, p < .001 \), of probe type, \( F(3,57) = 31.33, \text{MSE} = .19, p < .001 \), and also of serial position, \( F(6,114) = 30.56, \text{MSE} = .07, p < .001 \). A robust
Figure 6. Mean accuracy (in percentage) on serial probed recall of spoken and written letters for four probe types across serial position 4 to 10 from Experiment 1D.
modality effect and a recency effect were again found. Direct comparisons between
different probe types also replicated the pattern observed in Experiment 1C: Recall for
the two same-modality adjacent conditions was equivalent ($p > .15$), which was superior
than that for both the same-modality distant condition and the different-modality adjacent
condition (all $ps < .002$). Critically, the same-modality distant condition also had higher
recall accuracy than the different-modality adjacent condition ($p = .016$).

Since the interaction between target modality and probe type was also significant,
$F(3,57) = 8.42, MSE = .12, p < .001$, separate analyses were performed on each target
modality separately. For spoken targets, the mean accuracy for the four probe types were
as follows: 82.6% for AA(NI), 86.1% for AA(NISM), 61.1% for A..A, and 42.3% for
VA. Further analyses showed that recall accuracy was highest for both the adjacent same-
mode conditions, which was significantly better than that for the same-mode but distant
probe ($ps < .001$), which in turn was better than that for the different-mode probe ($p <
.002$). As in Experiment 1C, the same pattern was found for written targets, with a less
pronounced effect. The recall accuracy of the four probe types were 50.7% for VV(NI),
54.0% for VV(NISM), 40.6% for V..V, and 35.3% for AV. The two adjacent same-mode
conditions resulted in equivalent recall accuracy ($p > .25$), which were both higher than
that for the different-mode probe ($ps = .001$). The memory from the V..V condition,
although worse than that from VV(NISM) \( (p < .003) \) and no better than that from AV \( (p > .41) \), was only marginally worse than that from VV(NI) \( (p > .066) \).

All interactions between the three variables were significant: target modality and probe type, \( F(3,57) = 8.42, MSE = .12, p < .001 \), target modality and serial position, \( F(6,114) = 2.41, MSE = .05, p < .032 \), probe type and serial position, \( F(18,342) = 2.63, MSE = .03, p < .001 \), and target modality, probe type, and serial position, \( F(18,342) = 2.12, MSE = .03, p < .005 \). Further analyses indicated that all probe types produced a significant recency effect for both spoken and written targets \( (ps < .001) \), except for the V..V condition \( (p > .23) \). Due to the apparently different trend between the V..V and the AV conditions across different serial positions, a trend analysis was conducted which confirmed the significant difference between these two conditions, \( L_{V..V} = 0.42, L_{AV} = 1.56, t(19) = 3.35, p < .003 \). Similar to the results of Experiment 1C, the V..V condition resulted in numerically better memory performance than the AV condition except for the very last two serial positions.

Discussion

As it has been expected, the current experiment replicated the findings of Experiment 1C using capital letters instead of digits. The results speak to the reliability of the differential memory performance between auditorily and visually presented items. They also provided further evidence for the hypothesis of separate input and output
phonological retention in STM. The lack of significant difference between the V..V and the AV conditions is a little problematic to this hypothesis. However, as pointed out earlier, this finding might possibly arise due to the more memorable spoken probe in the AV condition. Given the fact that the written probe in the V..V condition was less memorable and temporally distant from the target, its numerically better memory performance was actually consistent with the idea that the written target related better with a written probe in the output phonological buffer than with a spoken probe in the input phonological buffer. In other words, the benefit of being stored in the same buffer overcame the disadvantages caused by a less memorable probe and the distance between the probe and the target.

_Expected vs. observed memory performance._ As in Experiment 1C, I computed the expected recall accuracy for targets following an adjacent probe at different serial positions for each participant (see Figure 7). A 2 (expected vs. observed) x 2 (same-mode vs. different-mode) x 6 (serial position 5 to 10) within-participant ANOVA was conducted for each target modality separately. The results from these analyses were very similar to that from Experiment 1C. Specifically, both target modalities showed a significant effect between the expected and the observed memory performance ($F(1,19) = 15.10, MSE = .04, p = .001$ for spoken targets, and $F(1,19) = 22.49, MSE = .06, p < .001$ for written targets), and a significant effect of serial position ($F(5,95) = 30.25, MSE = $
Figure 7. Expected and observed accuracy (in percentage) on serial probed recall of spoken and written letters for the same-mode and different-mode conditions across serial position 5 to 10 from Experiment 1D.
.04, \( p < .001 \) for spoken targets, and \( F(5.95) = 23.19, \text{MSE} = .06, p < .001 \) for written targets). Although the difference between the same-mode and the different-mode conditions was only significant for spoken targets, \( F(1,19) = 72.51, \text{MSE} = .18, p < .001 \), but not for written targets, \( F(1,19) = .35, \text{MSE} = .09, p > .56 \), the interaction between this variable and the variable of expected versus observed accuracy was significant for both modalities; \( F(1,19) = 26.93, \text{MSE} = .04, p < .001 \) for spoken targets, and \( F(1,19) = 53.16, \text{MSE} = .06, p < .001 \) for written targets.

Further analyses were performed on the same-mode and the different-mode conditions for each target modality separately. The results basically replicated the data from Experiment 1C. For the same-mode condition, the difference between the expected and the observed performance (83% vs. 85% for spoken targets, and 50% vs. 55% for written targets), albeit very small, was significant \( (F(1,19) = 34.27, \text{MSE} = .0008, p < .001, \text{and } F(1,19) = 54.12, \text{MSE} = .003, p < .001, \text{respectively}) \). The effect of serial position was significant for both modalities as well, \( F(5,95) = 24.42, \text{MSE} = .02, p < .001, \text{and } F(5,95) = 21.31, \text{MSE} = .04, p < .001 \). The interaction between the two variables was also significant, \( F(5,95) = 5.64, \text{MSE} = .004, p < .001 \) for spoken targets, and \( F(5,95) = 2.41, \text{MSE} = .007, p < .042 \) for written targets.

In contrast with the small but consistent underestimation of the expected performance relative to the observed one, the results from the different-mode conditions
showed much better expected accuracy than the observed memory performance. The ANOVA results indicated that the difference between the expected and the observed accuracy for spoken targets (60% vs. 43%) and for written targets was both significant, $F(1,19) = 20.66, MSE = .08, p < .001$, and $F(1,19) = 36.90, MSE = .11, p < .001$, respectively. The effect of serial position was also significant for both modalities, $F(5,95) = 17.20, MSE = .04, p < .001$, and $F(5,95) = 15.33, MSE = .04, p < .001$. The interaction between the two variables was only significant for written targets, $F(5,95) = 2.59, MSE = .03, p < .031$, but not significant for spoken targets, $F(5,95) < 1, MSE = .04, p > .51$.

The results from the comparisons between the expected and the observed memory accuracy replicated the pattern found in Experiment 1C and confirmed the existence of the memory cost associated with modality change. Even though the current data could not determine the exact source of this memory cost, it clearly demonstrated that the worse memory performance for the VA condition relative to the A..A condition was not simply due to the less memorable written probe. At the same time, the same reason could be applied to the fact that the AV condition was numerically worse but not significantly different from the V..V condition. That is, if the spoken probe in the AV condition were not more memorable than the written probe in the V..V condition, the difference between these two conditions might reach significance and further support the model of separate input and output phonological retention.
Presentation rate. The data from Experiment 1D, in contrast with that from Experiment 1B, indicated the importance of presentation rate in STM experiments. Inspection of the data from both experiments revealed two interesting points. First, comparing the two adjacent same-mode conditions, the faster presentation rate improved recall accuracy in the NISM instructional condition for both modalities (8.75% for spoken targets and 4.85% for written targets). Second, the faster presentation rate reduced recall accuracy particularly in the different-mode probe condition. That is, recall accuracy was reduced by 7.2% for the VA condition and 8.8% for the AV condition. Although the VV(NI) condition also had a lower recall accuracy in the faster than the slower presentation rate (50.7% and 52.5%, respectively), the difference was only 1.8%.

One possible reason for the ostensibly conflicting results described above was that the slower presentation rate allowed participants to internally rehearse the list while the faster presentation rate prevented it. The significance of internal rehearsal was that it generated a phonological representation in the output buffer, regardless of presentation modality of the information. When materials were presented in mixed-modality lists and rehearsed over time, both the input and output phonological buffers were exploited to store spoken items, while only the output phonological buffer was used to store written items. This was detrimental to the NISM condition, because the rehearsal process “equalized” the items in the list by storing all of them in the output phonological buffer.
On the other hand, this was particularly helpful to recall accuracy for the different-mode condition, since both the probe and the target had a representation in the output buffer when internal rehearsal was possible. In this sense, the AV and VA conditions were actually similar to the VV condition at a slower presentation rate. This speculation was consistent with the data from Experiment 1B and 1D. The validity of this claim was examined directly in Experiment 2.

In summary, the data from Experiment 1 provided strong support for the hypothesis that spoken and written verbal materials are remembered in two different phonological buffers in STM. Specifically, Experiment 1A replicated the well-known phonological similarity effect with written verbal materials, which indicated the phonology-based representations for visually presented items in STM. In Experiment 1B, it was found that a temporally distant probe in the same modality produced no worse serial recall than a temporally adjacent probe in a different modality. More importantly, in both Experiment 1C and 1D, the same-mode but distant probe served as a more effective cue than the adjacent different-mode probe. Also, the observed memory performance for the different-mode conditions was much worse than the expected performance, which took the memorability of the probe into account and tended to underestimate memory accuracy. In Experiment 2, lip-read words were mixed with spoken and written words separately to determine whether the memory cost induced by
different presentation modalities for the probe and the target was actually caused by switching between phonological buffers or by switching between presentation formats.

Experiment 2

As discussed earlier, words presented in lip-read films are appropriate stimuli to examine the issue of phonological retention in STM. Given the extremely high visual confusability between lip-read items, it is unlikely that the memory for them is maintained in visual codes. Also, it has been demonstrated that STM for lip-read items are reduced by an auditory suffix (Campbell & Dodd, 1980), thus indicating the phonological nature of their retention. According to the hypothesis of separate input and output phonological buffers in STM, lip-read films, as well as other materials that require internal generation of phonology, should be stored in the output buffer. On the other hand, spoken words, within which the phonological information is readily available, should be stored in the input buffer. The memory for these two kinds of materials should show dissociation.

Experiment 2 employed the same serial probed recall paradigm with mixed-modality list presentation as used in Experiment 1B to 1D. Specifically, lists consisting of spoken and lip-read words and lists consisting of written and lip-read words were used. If the hypothesis of separate input and output phonological retention in STM is correct, then it should be found that spoken words and words in lip-read films are not retained in the
same buffer, whereas written words and words in lip-read films are retained in the same (output) buffer. That is, the hypothesis leads to the prediction that a temporally distant probe in the same modality with the target is a more effective cue than a temporally adjacent probe in a different modality when spoken words and lip-read films were used in the mixed-modality lists. When written words and lip-read films were used in mixed-modality lists, on the other hand, worse recall should be observed in a temporally distant probe in the same modality with the target, since memory performance is determined by the temporal distance within the same buffer.

Due to the difficulty mainly caused by lower familiarity imposed by lip-reading films, the list length was reduced from 10 to 8. Two presentation rates, namely 1 item and 2 items per second, was examined in Experiment 2A and 2B, respectively. This was motivated by the findings from Experiment 1. If the difference caused by the two presentation rates in fact resulted from internal rehearsal allowed by the slower presentation rate, improved recall was predicted for the condition of different-mode probe with the slower presentation rate from lists consisted of spoken words and words in lip-read films. Although the same rehearsal process would be involved for lists consisted of written words and words in lip-read films as well, the recall from the condition of different-mode probe would not be benefited by the slower presentation rate, since all list items are already stored in the output phonological buffer regardless the presentation rate.
Experiment 2A

Method

Participants. Forty undergraduate students from Rice University participated in this experiment. They were all native English speakers who had normal or corrected-to-normal vision and hearing. Twenty students received the mixed-modality lists of spoken and lip-read words, while the other 20 students received the mixed-modality lists of written and lip-read words.

Materials. The stimuli used for Experiment 2A were the following eight words: bread, ears, felt, jump, live, owl, three, and wool. These words were chosen because they are easily differentiated from each other in lip-read films. One female native English speaker saying these eight words was videotaped and played to the participant on a computer screen silently as lip-read words. The voice of the same speaker saying these eight words was also recorded and played to the participant via the computer speakers as spoken words without any visual input. The printed words were presented to the participant in the center of the computer screen as written words. The duration of each word in all three modalities was edited to be 800 ms. The presentation rate of each list was 1 item per second.

Design and procedures. The design and procedure employed in Experiment 1B was adopted in Experiment 2A, except for the following necessary changes. Due to the
reduced length of each list, only target serial position 4 to 8 were tested in the probed serial recall paradigm. Two kinds of mixed-modality lists were prepared, with one kind consisting of spoken and lip-read words, and the other kind consisting of written and lip-read words. For each kind of mixed-modality lists, four different list types were prepared in which the first four words were WLLW, LLWW, LWWL, or WWLL (L indicates words in lip-read films and W indicates spoken or written words, depending on the kind of the mixed-modality lists). For each kind of mixed-modality presentation, 160 lists were needed to be exhaustive of the combination of four list types, eight words, and five serial position of the target. Each kind of the 160 mixed-modality lists were evenly divided into two sets and administered in the NI and NISM instructional conditions respectively. Half of the participants received lists consisting of spoken and lip-read words, while the other half of the participants received lists consisting of written and lip-read words. For each participant, two sessions were administered, such that the NI instruction was given in one session while the NISM instruction was given in the other session. The order of the two instructional conditions was counterbalanced across participants.

Results

The same analyses were conducted as in Experiment 1B to 1D for lists consisting of spoken and lip-read words and for lists consisting of written and lip-read words,
respectively. The memory accuracy under different conditions was computed across serial positions of the target, and depicted for two kinds of lists for both target modalities in Figure 8 and Figure 9. For memory accuracy from lists consisting of spoken and lip-read words (Figure 8), the 2 (target modality) x 4 (probe type) x 5 (serial position) within-participant ANOVA revealed significant main effects of target modality, $F(1,19) = 37.76, MSE = .27, p < .001$, of probe type, $F(3,57) = 24.31, MSE = .12, p < .001$, and also of serial position, $F(4,76) = 29.22, MSE = .09, p < .001$. As in previous experiments, the standard recency effect was found. Moreover, lip-read words in the current experiment were more similar to written words in the sense that they were retained less well than spoken words.

Since the interaction between target modality and probe type was also significant, $F(3,57) = 14.73, MSE = .10, p < .001$, separate analyses were performed on each target modality. For spoken targets, the mean accuracy for the four probe types were as follows: 78.3% for AA(NI), 81.3% for AA(NISM), 55.3% for A..A, and 41.0% for LA; the difference between these four conditions was significant, $F(3,57) = 34.97, MSE = .11, p < .001$. Paired comparisons revealed that the two adjacent same-mode conditions, although not different from each other, were remembered better than both the A..A and the LA condition ($p < .001$). More importantly, the A..A condition was significantly better than the LA condition ($p < .018$). The same pattern was observed for written targets, as the
Figure 8. Mean accuracy (in percentage) on serial probed recall of spoken and lip-read words for four probe types across serial position 4 to 8 from Experiment 2A.
recall accuracy for the four probe types was 36.5% for LL(NI), 49.8% for LL(NISM), 46.3% for L..L, and 33.5% for AL. The LL(NISM) condition yielded significantly better memory performance than the LL(NI) condition ($p < .008$) and than the AL condition ($p < .005$). As observed with spoken targets, the same-modality but distant probe (in the L..L condition) produced better memory performance than the adjacent different-modality probe (in the AL condition), $p < .04$.

The significant interaction between probe type and serial position ($F(12,228) = 2.03, MSE = .05, p < .023$) should be understood in light of the significant three-way interaction between target modality, probe type, and serial position ($F(12,228) = 1.81, MSE = .05, p < .047$). For spoken targets, all probe types produced a significant recency effect ($ps < .001$). For written targets, the LL(NI), LL(NISM), and AL conditions showed a recency effect ($ps < .001$) but the temporally distant probe (L..L) did not produce such an effect ($p = .27$). In summary, the results from the lists consisting of spoken and lip-read words supported the proposal of separate input and output phonological buffers in demonstrating that memory for spoken words and lip-read words is differentially retained.

When considering the memory performance from lists consisting of written and lip-read words (Figure 9), the results told a very different story. A 2 (target modality) x 4 (probe type) x 5 (serial position) within-participant ANOVA indicated no difference
Figure 9. Mean accuracy (in percentage) on serial probed recall of written and lip-read words for four probe types across serial position 4 to 8 from Experiment 2A.
between written and lip-read targets, \( F(1,19) < 1, \text{MSE} = .16, p > .88 \). There were main
effects of probe type, \( F(3,57) = 7.08, \text{MSE} = .19, p < .001 \), and of serial position, \( F(4,76) = 16.39, \text{MSE} = .10, p < .001 \). The interaction between probe type and serial position was
also significant, \( F(12,228) = 2.61, \text{MSE} = .06, p < .003 \). Different from previous
experiments, only a serial position effect was found and there was no difference between
written and lip-read target modalities.

The mean memory accuracy for different probe types was as follow: for written
targets, 59.3\% for VV(NI), 56.8\% for VV(NISM), 39.5\% for V..V, and 54.3\% for LV;
for lip-read targets, 51.5\% for LL(NI), 61.3\% for LL(NISM), 41.0\% for L..L, and 54.3\%
for VL. The data suggested that temporal distance between the probe and the target
determines memory performance, whereas the modality of the probe and the target has no
influence on recall. Further analyses confirmed this pattern: The distant same-modality
probe resulted in worse memory performance than the other three probe types (\( ps < .031 \))
while there was no difference among the other three probe types. Further analyses also
revealed that the significant interaction between probe type and serial position reflected
the significant serial position effect for all probe types (\( ps < .001 \)) except for the distant
same-modality probe (\( p > .60 \)). Consistent with the prediction from separate input and
output phonological STM, written words and lip-read words showed no dissociation with
each other, and the performance of serial recall was mainly determined by the distance between the probe and the target rather than the presentation modality.

When comparing the results for lip-read words mixed with spoken words to that for lip-read words mixed with written words, another important point was noted. In the current experiment, despite the superior memory for spoken words than for written words, as demonstrated in higher accuracy for the AA(NI) and the AA(NISM) than for the VV(NI) and the VV(NISM) condition, the memory for the LA and the AL condition was actually lower than that for the LV and VL condition. That is, even though the spoken word was more memorable, the memory cost caused by switching between buffers in the LA and the AL condition eliminated such an advantage and in turn produced worse recall than the LV and the VL condition, which contained a less memorable written item. This pattern further supported the separation between the input and output phonological buffers in STM.

Discussion

The current experiment employed verbal materials in lip-read films as well as spoken and written words to examine the performance of probed serial recall. The data from this experiment provided further support for the separation between input and output phonology in STM. The pattern implies that lip-read words, whose phonology like that for written words requires internal generation, were stored separately from spoken
words, whose phonology is readily available in the stimulus. As a result, when spoken words (in the input buffer) and lip-read words (in the output buffer) were mixed in the same list, a distant probe in the same modality was a more effective cue than an adjacent probe in a different modality. This pattern was observed for both spoken and lip-read targets. When lip-read words were mixed with written words, however, since both modalities were stored in the output phonological buffer, the distant probe yielded the worst memory performance because of the greater distance between the probe and the target. Switching modality between the probe and the target did not produce any significant memory cost, as evident by the equivalent memory performance between the adjacent same-modality and different-modality conditions.

The results of this experiment also provided evidence against Campbell’s (1987) proposal that spoken and lip-read words are both retained in phonetic codes and stored in the capacity similar to an input phonological buffer. Contrary to the common characteristics that a lip-read suffix shared with a spoken suffix but not with a written suffix (e.g., Campbell & Dodd, 1980, 1982; Greene & Crowder, 1984), lip-read items showed inferior STM than spoken items (i.e., the modality effect), and were similarly retained in the output phonological buffer just like written words (at least for mixed-modality lists). One possible account for this apparent discrepancy is that the suffix effect is a phenomenon rooted from sensory memory for very brief intervals (see Neath, 1998,
p. 48), whereas the memory for mixed-modality lists relies more on STM and is
different from that for single-modality lists (see Greene, 1989). To discern the exact
relations between sensory memory and STM, further research is needed.

The current experiment also provided evidence against the possibility that the
separation between STM for spoken and written words observed in Experiment 1B to 1D
was due to perceptual dissimilarity between these two modalities (LeCompte & Watkins,
1993). In the current experiment, although lip-read words were visually presented while
spoken words were auditorily presented, these two kinds of stimuli were both revealed
over time. On the other hand, although lip-read words and written words were both
visually presented, the former involved lip movements and were perceptually dissimilar
from the latter. If the separation between spoken and lip-read words was simply caused
by the perceptual dissimilarity between the stimuli, there is no good reason not to expect
such a separation between written and lip-read words. The lack of such a separation
between written and lip-read words, which is expected by the model of separate input and
output phonological buffers, poses a problem for the organization in STM based on
perceptual similarity.

*Expected vs. observed memory performance.* When spoken and lip-read words
were mixed together within one list, the STM for spoken words was significantly better
than that for lip-read words. Thus, it is possible to speculate that the superior memory
performance of the A..A condition than the LA condition was simply caused by higher memorability of the spoken probe than the lip-read probe. This account seems unlikely, however, because the L..L condition also yielded better recall than the AL condition, although the former condition had a less memorable probe than the latter.

To formally examine the influence from the memorability of the probe, the expected recall accuracy for targets following an adjacent probe at different serial positions was computed as in previous experiments. When the mixed-modality presentation lists consisted of spoken and lip-read words, the results were similar to the results that were observed for lists consisting of spoken and written words (see Figure 10). For spoken targets, the 2 (expected vs. observed) x 2 (same-mode vs. different-mode) x 4 (serial position 5 to 8) within-participant ANOVA revealed a significant effect between the same-mode and the different-mode conditions, $F(1,19) = 134.90, MSE = .07, p < .001$, and a significant main effect of serial position, $F(3,57) = 35.10, MSE = .04, p < .001$. For lip-read targets, the same ANOVA analysis revealed a significant effect between the expected and the observed memory performance conditions, $F(1,19) = 6.05, MSE = .04, p < .024$, and a significant main effect of serial position, $F(3,57) = 24.68, MSE = .08, p < .001$, but not a main effect of the difference between the same-mode and the different-mode probes, $F(1,19) = 1.75, MSE = .05, p > .20$. Although the difference between the expected and the observed memory performance was only significant for lip-
Figure 10. Expected and observed accuracy (in percentage) on serial probed recall of spoken and lip-read words for the same-mode and different-mode conditions across serial position 5 to 8 from Experiment 2A.
read targets but not for spoken targets, $F(1,19) = .41, MSE = .05, p > .53$, the interaction between this variable and the variable of same-mode and different-mode probe was significant for both modalities; $F(1,19) = 7.26, MSE = .04, p = .014$ for spoken targets, and $F(1,19) = 51.50, MSE = .03, p < .001$ for lip-read targets.

Further analyses were performed on the same-mode and the different-mode probe conditions for each target modality separately. The results basically replicated the pattern observed for lists consisting of spoken and written words in previous experiments. For the same-mode condition, the difference between the expected and the observed performance (78.3% vs. 82.9% for spoken targets, and 38.9% vs. 46.9% for lip-read targets), albeit relatively small, was significant ($F(1,19) = 31.91, MSE = .003, p < .001$, and $F(1,19) = 46.81, MSE = .005, p < .001$, respectively). These findings indicated that when the probe and the target were in the same modality, the expected recall underestimated memory performance, although the difference was relatively small. Despite this systematic underestimation, the results from the different-mode probe condition showed much better expected accuracy than the observed memory performance (51.0% vs. 43.1% for spoken targets, and 55.8% vs. 36.6% for lip-read targets). The ANOVAs indicated that the difference between the expected and the observed accuracy for lip-read targets was significant, $F(1,19) = 22.72, MSE = .07, p < .001$, but not for spoken targets, $F(1,19) = 2.68, MSE = .09, p = .12$. 
In sharp contrast to the findings from lists consisting of spoken and lip-read words, the results from lists consisting of written and lip-read words did not show any difference caused by modality switch between the probe and the target (see Figure 11). The recall accuracy was 54.6% for the expected LV condition, 56.6% for the observed LV conditions, 55.4% for the expected VV condition, 61.6% for the observed VV conditions, 54.3% for the expected VL condition, 57.8% for the observed VL conditions, 55.2% for the expected LL condition, 60.7% for the observed LL conditions. The first thing to note was that recall accuracy did not differ too much depending on the target modality. This was consistent with the finding that there was no modality effect for lists consisting of written and lip-read words. Second, the expected performance always underestimated the memory performance relative to the observed one, even for the different-mode probe condition. This result suggested that there was no detectable memory cost induced by the modality switch between the probe and the target.

Separate ANOVAs on written and lip-read words indicated no difference between the same-mode and the different-mode conditions \((F(1,19) < 1, \text{MSE} = .07, p > .34, \text{and} F(1,19) < 1, \text{MSE} = .06, p > .51, \text{respectively})\), and neither the interaction between this variable and the variable of expected versus observed memory performance \((F(1,19) = 2.13, \text{MSE} = .02, p > .16, \text{and} F(1,19) < 1, \text{MSE} = .03, p > .58, \text{respectively})\). The difference between the expected and the observed memory performance was significant
Figure 11. Expected and observed accuracy (in percentage) on serial probed recall of written and lip-read words for the same-mode and different-mode conditions across serial position 5 to 8 from Experiment 2A.
for written targets, $F(1,19) = 5.65$, $MSE = .02$, $p = .028$, and also for lip-read targets, $F(1,19) = 5.35$, $MSE = .03$, $p = .032$, indicating higher recall accuracy for the observed performance than the expected one. The serial effect was also significant for both target modalities, $F(3,57) = 18.46$, $MSE = .05$, $p < .001$ for written targets, and $F(3,57) = 20.06$, $MSE = .05$, $p < .001$ for lip-read targets.

In summary, the results from the comparisons between the expected and the observed memory accuracy clearly demonstrated the existence of the memory cost associated with modality change only when the different modalities of the probe and the target were in two different buffers. That is, even the prediction systematically underestimated true memory performance, the observed recall accuracy for the different-mode condition was still lower than the expected one, only when spoken and lip-read words were used but not when written and lip-read words were used. If this memory cost was simply caused by switching attention between two modalities or by perceptual discontinuity, lower recall accuracy should be observed as long as there was a modality switch, regardless which modalities were used. The current findings again argued against this possibility because the memory cost only appeared when the two modalities were retained in the input and output phonological buffers respectively, but not when the two modalities were retained in the same buffer.

Experiment 2B
Method

Participants. Forty undergraduate students from Rice University participated in this experiment. They were all native English speakers who had normal or corrected-to-normal vision and hearing. Twenty students received the mixed-modality lists of spoken and lip-read words, while the other 20 students received the mixed-modality lists of written and lip-read words.

Materials, design, and procedures. The method employed in this experiment was identical to Experiment 2A, except for two changes mentioned below. First, the presentation rate of the list items was 0.5 second per item. Second, the stimuli used in this experiment were 8 letters that are easily differentiated from each other in lip-read films (B, F, G, L, M, O, Y, and Z). This change was necessary because, compared to words, they are shorter in time as in lip-lead films and can be presented at a faster rate. The stimuli were prepared and displayed to the participants as the methods specified in Experiment 2A.

Results

The same analyses as in Experiment 2A were performed, and the results from this experiment mainly replicated the findings from Experiment 2A. For memory accuracy from lists consisting of spoken and lip-read words (Figure 12), the 2 (target modality) x 4 (probe type) x 5 (serial position) within-participant ANOVA confirmed the superior
Figure 12. Mean accuracy (in percentage) on serial probed recall of spoken and lip-read letters for four probe types across serial position 4 to 8 from Experiment 2B.
memory for spoken words than for lip-read words \((F(1,19) = 123.24, MSE = .17, p < .001)\), the effect of probe type \((F(3,57) = 55.64, MSE = .10, p < .001)\), and a robust serial position effect \((F(4,76) = 15.98, MSE = .07, p < .001)\). More importantly, when comparing the memory accuracy from the same-mode but distant condition with that from the different-mode adjacent condition, the A..A condition (61.3% correct) was significantly better than the LA condition (36.3% correct), \(p < .001\), and the L..L condition (36.0% correct) was marginally better than the AL condition (27.5% correct), \(p = .08\).

As evident in Figure 12, the L..L condition had higher recall accuracy than the AL condition except for the last serial position. This was confirmed by the significant interaction between probe type and serial position for written targets, \(F(12,228) = 2.96, MSE = .05, p = .001\). Further analyses revealed that while LL(NI) and the AL condition showed a significant serial position effect \((ps < .001)\), the LL(NISM) and the L..L condition did not \((p = .17\) and \(p = 26\), respectively). Although this pattern was similar to that in Experiment 2A, the failure to find a significant difference between the L..L and the AL conditions could be due to the floor effect and/or the superior memorability of the spoken probe.

Contrary to the pattern with lists consisting of spoken and lip-read words, memory performance with lists consisting of written and lip-read words was very
different (Figure 13). A 2 (target modality) x 4 (probe type) x 5 (serial position) within-participant ANOVA indicated no difference between written and lip-read targets, \( F(1,19) = 1.10, MSE = .10, p > .31 \), although the main effects of probe type (\( F(3,57) = 14.05, MSE = .15, p < .001 \)) and of serial position (\( F(4,76) = 20.47, MSE = .10, p < .001 \)) were both highly significant. All interactions were not significant (\( ps > .40 \)). The mean memory accuracy for different probe types were as follow: for written targets, 57.3% for VV(NI), 53.5% for VV(NISM), 32.8% for V..V, and 52.8% for LV, and for lip-read targets, 57.5% for LL(NI), 47.3% for LL(NISM), 33.8% for L..L, and 48.3% for VL. Paired comparisons between different probe types across two target modalities clearly indicated that the distant same-modality probe produced the worse memory performance than the other three probe types (\( ps < .003 \)).

Discussion

The current experiment replicated Experiment 2A in demonstrating the STM pattern indicating separate buffers for spoken and lip-read words but not for written and lip-read words. Specifically, a distant probe in the same modality was a more effective cue than an adjacent probe in a different modality when the two modalities were stored in two different buffers (e.g., spoken words and lip-read words in the input and output buffer, respectively). Also, recall accuracy of the target after an adjacent probe in a different modality suffered from a memory cost caused by the modality switch. On the
Figure 13. Mean accuracy (in percentage) on serial probed recall of written and lip-read letters for four probe types across serial position 4 to 8 from Experiment 2B.
other hand, if the two modalities were both retained in the same buffer (e.g., written and lip-read words in the output buffer), the distant probe resulted in the worse memory performance and received no advantage from being in the same modality with the target. Moreover, when the probe was adjacent to the target, there was no memory cost associated with the modality switch between the probe and the target. These results ruled out the possibility that the dissociation observed in the present study was simply caused by perceptual dissimilarity between different presentation modalities. They provided further evidence for the model of separate input and output phonological STM.

Similar to Experiment 2A, spoken words were remembered better than written words, as clearly demonstrated by higher accuracy for the AA(NI) and the AA(NISM) condition than for the VV(NI) and the VV(NISM) condition. However, the conditions which involved switching between buffers (i.e., the LA and the AL condition) resulted in worse recall accuracy than the conditions which did not involve such switch (i.e., the LV and the VL condition), despite the fact that spoken items were more memorable, and all these conditions changed presentation modality between the probe and the target. This pattern again spoke to the advantage of being retained in the same buffer. The memory cost particularly associated with changing buffers (but not with changing presentation modality) was formally tested in the following analyses.
Expected vs. observed memory performance. The same analyses comparing the
expected and the observed memory performance were conducted as in Experiment 2A.
When the lists consisted of spoken and lip-read words, the recall accuracy was 88.9% for
the expected AA condition, 91.1% for the observed AA conditions, 55.3% for the
expected LA condition, 38.1% for the observed LA conditions, 39.2% for the expected
LL condition, 45.7% for the observed LL conditions, 60.3% for the expected AL
condition, 31.3% for the observed AL conditions (see Figure 14). Overall, spoken targets
were remembered better than lip-read targets, as discovered in previous analyses. More
importantly, while the expected accuracy was lower than the observed one for the same-
mode condition, it was higher than the observed one for the different-mode condition.
This was confirmed by the significant interaction between the variable of expected versus
observed accuracy and the variable of same-mode and different conditions ($F(1,19) =
26.75, MSE = .03, p < .001$ for spoken targets, and $F(1,19) = 59.60, MSE = .04, p < .001$
for lip-read targets).

When separate paired comparisons were conducted on the same-mode and the
different-mode conditions for both target modalities, it became clear that even for the
same-mode condition, the expected recall accuracy underestimated the observed one
($F(1,19) = 44.81, MSE = .0004, p < .001$ for spoken targets, and $F(1,19) = 36.14, MSE =
.005, p < .001$ for lip-read targets), for the different-mode condition the expected recall
Figure 14. Expected and observed accuracy (in percentage) on serial probed recall of spoken and lip-read letters for the same-mode and different-mode conditions across serial position 5 to 8 from Experiment 2B.
accuracy was significantly higher than the observed one \( F(1,19) = 21.70, MSE = .05, \) 
\( p < .001 \) for spoken targets, and \( F(1,19) = 45.20, MSE = .07, p < .001 \) for lip-read 
targets. The memory cost associated with the modality change between the probe and the 
target was not caused by some perceptual or attentional factors because the same memory 
cost was not observed when the lists consisted of written and lip-read words (see below).

When the lists consisted of written and lip-read words, the recall accuracy was 
52.7\% for the expected VV condition, 59.7\% for the observed VV conditions, 51.4\% for 
the expected LV condition, 56.9\% for the observed LV conditions, 49.3\% for the 
expected LL condition, 56.8\% for the observed LL conditions, 50.6\% for the expected 
VL condition, 52.5\% for the observed VL conditions (see Figure 15). Different from the 
results from lists consisting of spoken and lip-read words, there was no modality effect. 
The interaction between the variable of expected versus observed accuracy and the 
variable of same-mode and different conditions was not significant, either \( F(1,19) < 1, \) 
\( MSE = .02, p > .64 \) for written targets, and \( F(1,19) = 1.60, MSE = .04, p > .22 \) for lip-read 
targets). The highly significant difference between the expected and the observed 
memory performance confirmed that the expected accuracy systematically 
underestimated the observed one \( F(1,19) = 10.19, MSE = .03, p < .005 \) for spoken 
targets, and \( F(1,19) = 5.61, MSE = .03, p < .029 \) for lip-read targets). Consistent with the 
findings in Experiment 2A, the memory cost associated with modality change between
Figure 15. Expected and observed accuracy (in percentage) on serial probed recall of written and lip-read letters for the same-mode and different-mode conditions across serial position 5 to 8 from Experiment 2B.
the spoken and lip-read words but not with that between written and lip-read words could not be attributed to some peripheral processing discontinuity. Rather, it supported the hypothesized distinction between materials respectively retained in the input and output phonological buffers in verbal STM.

Presentation rate. When comparing the results for slow and fast presentation rates from lists consisting of spoken and written words, as in Experiment 1B and 1D respectively, it was noted that while the faster presentation rate in general improved recall accuracy, it reduced recall accuracy particularly in the condition with a different-mode probe. This pattern was explained by the fact that the slower presentation rate allowed participants to internally rehearse the list while the faster presentation rate prevented it. Since the product of internal rehearsal, regardless its original presentation modality, was retained in the output phonological buffer, the different-probe condition in the slow presentation rate functionally behaved more like a same-mode condition in which both the probe and the target were both stored in the output phonological buffer. Thus, recall accuracy was better when the presentation rate was slow.

To examine the effect of presentation rate on recall accuracy, the data from Experiment 2A and 2B were pooled together, and a 2 (presentation rate) x 2 (target modality) x 4 (probe type) x 5 (serial position) ANOVA was conducted on the lists consisting of spoken and lip-read words, and on the lists consisting of written and lip-
read words, respectively. In the following discussion, I only concentrated on the effects relevant to the variable of presentation rate, since all other effects were similar to the results reported previously. For the data from the lists consisting of spoken and lip-read words, although there was no main effect of presentation rate, $F(1,38) < 1$, $MSE = .54$, $p > .90$, the interaction between presentation rate and target modality was significant, $F(1,38) = 4.23$, $MSE = .22$, $p = .047$, which was caused by the more pronounced modality effect with the fast presentation rate (32.0%) than the slow presentation rate (22.5%). The interaction between presentation rate and probe type was also marginally significant, $F(3,114) = 2.60$, $MSE = .11$, $p = .056$. No other effect relevant to presentation rate approached significance.

To further understand the interaction between presentation rate and probe type, a simple comparison between slow and fast presentation rate in each probe type was performed. Although none of these comparisons were significant ($ps > .063$), the marginal significant interaction indicated the different trends. For the same-mode condition under NI instruction, recall accuracy was 57.4% for slow presentation rate and 64.1% for fast presentation rate. For the same-mode condition under NISM instruction, recall accuracy was 65.6% for slow presentation rate and 68.3% for fast presentation rate. These results were consistent with Experiment 1B and 1D in that memory performance improved with a faster presentation rate. For the same-mode but distant probe condition,
recall accuracy was 50.8% for slow presentation rate and 48.7% for fast presentation rate. Although memory performance reduced slightly with a faster presentation rate, the difference was very minimal. For the different-mode probe condition, recall accuracy was 37.3% for slow presentation rate and 31.9% for fast presentation rate. As discovered before, the slow presentation rate improved recall accuracy particularly when the probe and the target were in two different modalities.

It is not the case, however, that internal rehearsal which is more likely to happen with a slow presentation rate simply helps the different-mode probe condition regardless of the modalities involved. If the two modalities in which the probe and the target are presented are already retained in the output phonological buffer, internal rehearsal allowed by a slower presentation rate would not specifically help memory performance. This prediction was confirmed by the ANOVA on the lists consisting of written and lip-read words. The variable of presentation rate did not produce a main effect, $F(1,38) = 1.31, MSE = .58, p > .26$, nor did it produce any significant interaction, all $ps > .16$.

Recall accuracy for both the slow and fast presentation rate was 55.4% and 57.4% for the same-mode condition under NI instruction, 59.1% and 50.4% for the same-mode condition under NISM instruction, 40.3% and 33.3% for the same-mode but distant probe condition, and 54.3% and 50.6% for the different-mode condition. Although the different-mode condition still resulted in higher recall accuracy with the slower
presentation rate, the magnitude was relatively small, and other conditions also benefited from the slow presentation rate with these materials.

The current study was not originally planned to examine the effect of presentation rate on recall accuracy. Thus, the results obtained can only be interpreted with caution and they did not provide clear evidence for or against any specific model of verbal STM. Nevertheless, the findings regarding presentation rate was indeed consistent with the model of separate input and output phonological retention in that compared to the effect on the same-mode adjacent probe conditions, slow presentation rate produced a reversed effect on the different-mode adjacent probe condition when the two modalities involved were retained in the input and the output buffer respectively (e.g., spoken and lip-read words). On the other hand, slow presentation rate did not cause any significant effect when the two modalities involved were both retained in the output phonological buffer (e.g., written and lip-read words).

Experiment 3

In addition to lip-read words, nameable pictures are also appropriate stimuli to examine the issue of phonological retention in verbal STM. As it has been discussed earlier, previous studies have shown that memory for nameable pictures is influenced by phonological similarity and word length of the names of the pictures (Schiano & Watkins, 1981) and is reduced by simultaneously verbal suppression (Brandimonte,
Hitch, & Bishop, 1992). These results suggest that nameable pictures, albeit visually presented, are retained in phonological codes. Since there is no acoustic information afforded in pictures, the phonological representation of these pictures must be generated internally, and thus should be stored in the output phonological buffer as predicted by the hypothesis of separate phonological retention in STM.

In Experiment 3A, I first sought to replicate the phonological similarity effect with nameable pictures to demonstrate that pictures are retained phonologically in STM. Specifically, phonological and visual similarity among pictures within a list was manipulated independently. Both slow and fast presentation rate (1 second and 0.5 second per item, respectively) was examined, since the fast presentation rate could reduce the likelihood of internal rehearsal, which might mask the effects for separation between input and output phonological retention. However, it has been reported that phonological similarity between picture names within the to-be-remembered list only influenced STM at the rate of 1 item per second but not at the rate of 8 items per second (Coltheart, 1999). If picture names were not retrieved and retained phonologically in STM at the fast presentation rate, then the results would not be relevant to the organization of verbal STM.

In Experiment 3B, the serial probed recall paradigm with mixed-modality list presentation was once again used as in Experiment 2. Memory for lists consisting of
spoken words and nameable pictures and for lists consisting of written words and
nameable pictures was examined. According to the hypothesis of separate phonological
retention, when the mixed-modality list is composed of items retained in the input and the
output buffers respectively (e.g., spoken words and nameable pictures), a temporally
distant probe in the same modality with the target is expected to be a more effective cue
than a temporally adjacent probe in a different modality. On the other hand, when items
in both modalities are retained in the same buffer (e.g., written words and nameable
pictures), a temporally distant probe in the same modality with the target is expected to
result in the worse recall due to the greater distance between the probe and the target.

Experiment 3A

Method

Participants. Thirty undergraduate students from Rice University were recruited
to participate in this experiment. Only native English speakers who have normal or
corrected-to-normal vision and hearing were qualified. Eighteen students were tested
with the presentation rate at one second per item, while the other 12 students were tested
with the presentation rate at 0.5 second per item.

Materials, design, and procedures. The methods used in this experiment were
identical to those of Experiment 1A, except for two changes. First, all stimuli were
nameable pictures from Snodgrass and Vanderwart (1980) instead of capital letters. For
phonologically similar and visually dissimilar (PSVD) lists, the pictures of stool, scissors, screwdriver, seal, snail, spider, squirrel, and sweater were used. These pictures are visually dissimilar, while their names share one initial phoneme /s/ and one of the two final phonemes /l/ and /er/ with one exception (scissors). For phonologically dissimilar and visually similar (PDVS) lists, the pictures of axe, comb, fork, knife, pencil, spoon, toothbrush, and wrench were used. All these pictures have a similar shape with the same orientation, but none of their names share the initial phoneme or rhyme with each other.

For phonologically dissimilar and visually dissimilar (PDVD) lists, the pictures of anchor, cherry doorknob, kite, giraffe, snowman, mouse, whistle were used. All these pictures are visually dissimilar, and none of their names share the initial phoneme or rhyme with each other.

The second methodological deviation of this experiment from Experiment 1A was that two presentation rates, namely 1 second and 0.5 second per item, were employed. The presentation rate of 1 second per item was comparable with that of Experiment 1A and was commonly used in previous studies to study STM for pictures (e.g., Schiano & Watkins, 1981). The findings from Experiment 1B to 1D, however, suggested that a faster presentation is crucial to the manifestation of the differential memory performance between input and output phonological retention. Thus, both presentation rates were examined.
Results and Discussion

The same analyses were performed as in Experiment 1A. The probed recall accuracy in the three conditions across different serial positions of the target was computed for slow (i.e., 1 second per item) and fast (i.e., 0.5 second per item) presentation rates and reported in Figure 16 and Figure 17, respectively. The 3 (PSVD, PDVS, and PDVD) x 7 (serial position from 2 to 8) within-participant ANOVA for the slow presentation rate indicated a significant main effect of phonological and visual similarity among list items, $F(2, 34) = 6.11, MSE = .08, p = .005$, as well as the main effect of serial position of the target, $F(6, 102) = 5.99, MSE = 0.08, p < .001$. The interaction between these two variables was not significant, $F(12, 204) = 1.07, MSE = 0.03, p > .38$. Paired comparisons between different conditions revealed that memory accuracy for PSVD lists (46.1%) was significantly worse than that of PDVS (56.2%) and PDVD (57.5%) lists, $t(17) = 3.18, p = .005$, and $t(17) = 2.76, p = .013$, respectively. The difference between the PDVS and PDVD lists was far from being significant ($t(17) = .39, p > .70$). These results clearly demonstrated that nameable pictures were retained phonologically rather than visually when presented at 1 second per item. Also the memory for nameable pictures, just like that for other verbal materials, showed a robust recency effect.
Figure 16. Mean accuracy on serial probed recall (in percentage) of nameable pictures for three levels of stimulus similarity across serial positions 2 to 8 at 1 second per item from Experiment 3A.

The ANOVA for a faster presentation rate (i.e., 0.5 second per item) revealed a somewhat different pattern (see Figure 17). The main effects of stimulus similarity and serial position were again found, $F(2, 22) = 4.53, MSE = .02, p = .023$, and $F(6, 66) = 12.58, MSE = .07, p < .001$, respectively. The interaction between these two variables was not significant, $F(12, 132) = 1.09, MSE = .03, p > .37$. Further analyses indicated that while recall accuracy for PSVD lists (40.9%) was worse than that of PDVD lists (46.6%), $t(11) = 2.42, p = .034$, the recall accuracy of PDVS lists (40.5%) was also worse than that of PDVD lists, $t(11) = 2.93, p = .014$. No difference was found between PSVD and PDVS lists, $t(11) = 0.19, p > .85$. That is, when nameable pictures were presented at 0.5
second per item, both phonological and visual similarity influenced memory performance.

*Figure 17.* Mean accuracy on serial probed recall (in percentage) of nameable pictures for three levels of stimulus similarity across serial positions 2 to 8 at 0.5 second per item from Experiment 3A.

The results of the current experiment from the slow presentation rate were consistent with other studies (e.g., Coltheart, 1999; Schiano & Watkins, 1981) in showing that phonological, but not visual, similarity influenced STM for nameable pictures. The results from the fast presentation rate also extended previous findings to demonstrate that when presented in an intermediate speed between 1 item and 8 items per second, pictures might be retained both visually and phonologically. Since the main goal of this study was to examine whether the phonological STM responsible for pictures is the same or
different capacity from that for spoken and written words, only the slow presentation rate (i.e., 1 second per item) in which pictures are clearly retained in a phonological code was employed in the following experiment.

*Experiment 3B*

In Experiment 3B, the serial probed recall paradigm was again employed as in Experiment 2. Nameable pictures were mixed with spoken and written words respectively, and the recall accuracy under different probe conditions was examined. The presentation rate was 1 second per item, since according to the results from Experiment 3A, STM for pictures was clearly based on phonological rather than visual codes at this speed.

*Method*

*Participants.* Forty-eight undergraduate students from Rice University were recruited to participate in this experiment. They were all native English speakers who had normal or corrected-to-normal vision and hearing. Twenty-four students received the mixed-modality lists of spoken words and nameable pictures, while the other 24 students received the mixed-modality lists of written words and nameable pictures.

*Materials, design, and procedures.* The methods employed in this experiment were identical to those in Experiment 2A, except that lip-read words were replaced by the PDVD pictures from Experiment 3A. That is, two kinds of mixed-modality lists were
prepared, with the first kind consisting of spoken words and nameable pictures, and
the second kind consisting of written words and nameable pictures. For each kind of
mixed-modality lists, four different list types were prepared in which the first four words
were WPPW, PPWW, PWWP, or WWPP (P indicates nameable pictures and W indicates
spoken or written words, depending on the kind of the mixed-modality lists). Each list
was composed of axe, comb, fork, knife, pencil, spoon, toothbrush, and wrench, half in
spoken (or written) words and the other half in pictures. The pictures and written words
was displayed via a computer, and the names of these pictures spoken by a female native
English speaker was recorded and displayed to the participant via the computer speakers.
The presentation rate was 1 second per item.

Results

The same analyses were conducted as in Experiment 2A for lists consisting of
spoken words and namable pictures and for lists consisting of written words and namable
pictures, respectively. The memory accuracy under different conditions was computed
across serial positions of the target, and depicted for two kinds of lists for both target
modalities in Figure 18 and Figure 19. For memory accuracy from lists consisting of
spoken words and namable pictures (Figure 18), the 2 (target modality) x 4 (probe type) x
5 (serial position) within-participant ANOVA revealed significant main effects of target
modality, \( F(1, 23) = 91.79, \text{MSE} = .18, p < .001 \), of probe type, \( F(3, 69) = 19.57, \text{MSE} = \)
Figure 18. Mean accuracy (in percentage) on serial probed recall of spoken words and nameable pictures for four probe types across serial position 4 to 8 of Experiment 3B.
.12, \( p < .001 \), and also of serial position, \( F(4, 92) = 28.36, MSE = .08, p < .001 \). As in previous experiments, the standard recency effect was found. Pictures, like written and lip-read words, were also remembered less well than spoken items.

Since the interaction between target modality and probe type was also significant, \( F(3, 69) = 6.41, MSE = .11, p = .001 \), separate analyses were performed on each target modality. For spoken targets, the mean accuracy for the four probe types was as follows: 86.5\% for AA(NI), 93.3\% for AA(NISM), 75.2\% for A..A, and 58.3\% for PA; the difference between these four conditions was highly significant, \( F(3, 69) = 21.28, MSE = .13, p < .001 \). Paired comparisons revealed that the AA(NISM) condition had the highest recall accuracy and was significantly different from the AA(NI), the A..A, and the PA condition (all \( ps < .007 \)). Again, the difference between the AA(NI) and AA(NISM) conditions might be due to participants’ different strategies under different instructions. The AA(NI) condition also resulted in significantly better memory performance than the A..A and the PA condition (\( ps < .036 \)). More importantly, recall accuracy was higher for the A..A condition than for the PA condition (\( p = .008 \)), as expected by the hypothesis of separate input and output phonological retention.

When considering picture targets, the same pattern was observed although the difference between conditions was less significant. Recall accuracy for the four probe types was 51.9\% for PP(NI), 59.0\% for PP(NISM), 50.4\% for P..P, and 47.1\% for AP.
Only the PP(NISM) condition yielded significantly better memory performance than the P..P and the AP conditions \((ps < .015)\), and no other comparison was significant. However, visual inspection of Figure 18 revealed that the recall accuracy for the P..P condition was higher than the AP condition except for the last two serial positions. This observation was confirmed by the highly significant interaction between probe type and serial position, \(F(12, 276) = 2.38, \text{MSE} = .06, p = .006\). Further analyses indicated that while the P..P condition did not have a significant serial position effect \((p = .10)\), all the other three conditions did \((all \, ps < .002)\). The failure to obtain a significantly better memory performance in the P..P than the AP condition could be caused by higher memorability of a spoken probe than a picture probe. This possibility was addressed in the following discussion by examining the expected and the observed memory performance.

In sharp contrast with the higher recall accuracy for the same-mode distant probe condition than the different-mode adjacent probe condition observed from lists consisting of spoken words and namable pictures, the data from lists consisting of written words and namable pictures had the worst memory performance for the condition of a same-mode distant probe (see Figure 19). The mean recall accuracy for different probe types was as follow: for written targets, 63.3% for VV(NI), 65.8% for VV(NISM), 53.5% for V..V, and 59.2% for PV; for picture targets, 59.6% for PP(NI), 62.1% for PP(NISM), 48.3% for
Figure 19. Mean accuracy (in percentage) on serial probed recall of written words and nameable pictures for four probe types across serial position 4 to 8 of Experiment 3B.
P..P, and 59.8% for VP. The ANOVA indicated no difference between written and picture targets, $F(1, 23) = 1.81, MSE = .12, p > .19$, and no interaction between this variable with probe type, $F(3, 69) < 1, MSE = .06, p > .61$. There were main effects of probe type, $F(3, 69) = 5.87, MSE = .13, p = .001$, and of serial position, $F(4, 92) = 20.78, MSE = .12, p < .001$. The interaction between probe type and serial position was also significant, $F(12, 276) = 3.47, MSE = .06, p < .001$.

Paired comparisons between different probe types confirmed the idea that temporal distance between the probe and the target determines memory performance, whereas the modality of the probe and the target has no influence on recall. The same-mode but distant probe resulted in worse memory performance than all the other three probe types ($ps < .037$), while there was no difference among the other three probe types. Further analyses also revealed that the significant interaction between probe type and serial position reflected the significant serial position effect for all probe types ($ps < .001$) except for the same-mode but distant probe ($p > .55$). Consistent with the prediction from the hypothesis of separate input and output phonological STM, written words and nameable pictures showed no sign of being retained in two different buffers, and the performance of serial recall was mainly determined by the distance between the probe and the target rather than the presentation modality.
Discussion

Consistent with the predictions from the hypothesis of separate input and output phonological retention, the current experiment provided further support by employing nameable pictures. When spoken words (in the input buffer) and nameable pictures (in the output buffer) were mixed in the same list, a distant probe in the same modality was a more effective cue than an adjacent probe in a different modality, although the difference failed to reach significance for picture targets (which was likely masked by differential memorability of the probe in different conditions). When nameable pictures were mixed with written words, on the other hand, since both modalities were stored in the output phonological buffer, the distant probe yielded the worse memory performance due to the greatest distance between the probe and the target. Switching modality between the probe and the target did not produce any significant memory cost, as evident by the equivalent memory performance between the adjacent same-modality and different modality conditions.

Expected vs. observed memory performance. To verify whether the superior recall accuracy for the A..A than the PA condition and the lack of significant difference between the P..P and the AP conditions were caused by the higher memorability of a spoken probe than a picture probe, the expected memory performance for targets following an adjacent probe at different serial positions was computed as in previous
experiments. A 2 (expected vs. observed) x 2 (same-mode vs. different-mode) x 4 (serial position 5 to 8) within-participant ANOVA was conducted for both target modalities in lists consisting of spoken words and nameable pictures, and in lists consisting of written words and nameable pictures, respectively (see Figure 20 and Figure 21).

When the mixed-modality presentation lists consisted of spoken words and nameable pictures (Figure 20), the difference between expected and observed performance was not significant for spoken targets, $F(1, 23) < 1, MSE = .04, p > .91$, but significant for picture targets, $F(1, 23) = 6.25, MSE = .04, p = .020$. Further analyses for spoken targets revealed that although the same-mode conditions had higher memory performance than the different-mode conditions, $F(1, 23) = 56.46, MSE = .0005, p < .001$, this variable did not interact with the variable of expected versus observed recall, $F(1, 23) = 1.96, MSE = .04, p = .18$. That is, although the recall accuracy of the expected same-mode condition (90.0%) was lower than that of the observed same-mode condition (92.7%) while the recall accuracy of the expected different-mode condition (65.6%) was higher than that of the observed different-mode condition (62.5%), the reversing effect between the two variables was not strong enough to reach significance. It should be pointed out, however, that the general pattern of these data did conform to the findings from previous experiments, indicating the memory cost associated with the modality.
Figure 20. Expected and observed accuracy (in percentage) on serial probed recall of spoken words and nameable pictures for the same-mode and different-mode conditions across serial position 5 to 8 from Experiment 3B.
switch between the probe and the target, despite the expected performance systematically underestimated memory performance.

Further analyses on picture targets from the lists consisting of spoken words and nameable pictures showed results similar to those found with spoken targets. The same-mode conditions had higher memory performance than the different-mode conditions, $F(1, 23) = 9.31, MSE = .04, p = .006$. More importantly, this variable significantly interacted with the variable of expected versus observed recall, $F(1, 23) = 28.40, MSE = .04, p < .001$. Paired comparisons indicated that the recall accuracy of the expected same-mode condition (51.8%) was significantly lower than that of the observed same-mode condition (57.7%), $F(1, 23) = 37.92, MSE = .004, p < .001$, whereas the recall accuracy of the expected different-mode condition (69.0%) was significantly higher than that of the observed different-mode condition (53.4%), $F(1, 23) = 16.50, MSE = .07, p < .001$. These results indicated for the AP condition, although the spoken probe was more memorable, which might in turn have eliminated the difference between this condition and the P.P condition, there was still a memory cost associated with the modality switch between the probe and the target.

Contrary to the findings from lists composed of spoken words and nameable pictures, the expected memory performance from lists composed of written words and nameable pictures showed systematic underestimation for the observed memory
performance \( F(1, 23) = 4.38, MSE = .03, p = .048 \) for written targets, and \( F(1, 23) = 10.12, MSE = .03, p = .004 \) for picture targets) (see Figure 21). There was no interaction between the variable of expected versus observed performance and the variable of same-mode and different-mode probe \( (ps > .055) \). The mean recall accuracy was 62.1% for the expected VV condition, 67.7% for the observed VV conditions, 59.3% for the expected PV condition, 61.5% for the observed PV conditions, 57.4% for the expected PP condition, 64.5% for the observed PP conditions, 60.1% for the expected VP condition, 65.1% for the observed VP conditions. As clearly indicated by these results, the method used to derive the predicted memory performance slightly but systematically underestimated the memory performance. More importantly, the modality switch between written words and nameable pictures did not produce any significant memory cost.

In summary, the data from Experiment 3B was similar to that from Experiment 2 and further supported the hypothesis of separate input and output phonological retention in STM. Specifically, while spoken words were stored in the input phonological buffer, nameable pictures as well as written words were stored in the output phonological buffer. Therefore, a distant probe in the same buffer resulted in higher recall accuracy than an adjacent probe in the other buffer (e.g., better recall for the A..A than the PA condition). This effect was not simply caused by perceptual organization of different presentation
Figure 21. Expected and observed accuracy (in percentage) on serial probed recall of written words and nameable pictures for the same-mode and different-mode conditions across serial position 5 to 8 from Experiment 3B.
modalities, because the same finding was not obtained when the two modalities were both retained in the same buffer (e.g., written words and nameable pictures). Rather, a distant probe in the same modality produced the worst memory performance when written words and nameable pictures were used in a list. The comparison between the expected and the observed recall accuracy also indicated that there was memory cost associated with modality change only when the two modalities involved were retained in the input and the output buffers respectively. Switching between two modalities that were both stored in the same buffer did not cause any detectable effect.

Experiment 4

The differential STM performance between spoken and written verbal materials observed in Experiment 1 argued strongly for the hypothesis of separate input and output phonological retention. This hypothesis was further supported by the separation of spoken items with lip-read words (from Experiment 2) and with nameable pictures (from Experiment 3), but not between written words and lip-read words or between written words and nameable pictures. It might still be possible to argue, however, that the observed separation in the previous experiments reflected the distinction between the perceptual channels in which the stimuli are presented in, namely audition and vision, rather than the distinction between the input and output phonological buffers.
Although this speculation seems very unlikely, given the evidence of the phonology-based STM for written items (Experiment 1A), lip-read words (Campbell & Dodd, 1980, 1982), and nameable pictures (Experiment 3A), Experiment 4 tested this possible confounding empirically by using words in ASL. Since ASL is the natural linguistic input for some congenitally deaf people, the hypothesis of separate phonological retention predicts that signed words should be stored in the input STM buffer for these people, while written words and nameable pictures should be stored in the output STM buffer. That is, for deaf signers, the STM for remembering incoming signs is retained in the input buffer while the STM for generating signs to remember other verbal materials is retained in the output buffer. On the other hand, if a distinction between auditory and visual STM is an adequate account for the results from Experiment 1 to 3, then written and signed words, as well as nameable pictures, should be stored in the same buffer for visually presented items.

A pilot study was conducted in which signed and written words were presented in mixed-modality lists to 15 deaf participants. Although these participants could read words well when there was no time pressure and they understood the instructions, their STM performance for written words at the rate of 1 second per item varied to a great degree. Only five of these participants achieved a mean accuracy of 50% or higher correct in the STM task with a list length of 6 items. Due to the difficulty of this task to a
majority of deaf participants and the great variability among their performance on this task, only a strong target modality effect, namely better recall for signs than for written words, and a robust serial effect were identified. All other effects were not significant. The poor STM performance, especially for written items, was consistent with previous research (e.g., Shand, 1982). It has also been suggested that the codes used for printed words by deaf ASL signers might depend on a variety of task and participant factors (Hanson & Lichtenstein, 1990; Lichtenstein, 1998).

To reduce the difficulty of the task, and to use verbal materials that are clearly retained in ASL codes, nameable pictures were employed mixed with signed words. Wilson and Emmorey (1997) reported that nameable pictures showed a phonological similarity effect in ASL codes for deaf signers. Specifically, pictures whose names were similar in ASL phonology were remembered worse than pictures whose names were dissimilar. Moreover, this effect was eliminated when manual suppression was introduced, similar to the elimination of the phonological similarity effect from written words when subvocal suppression was introduced (Wilson & Emmorey, 1997). This piece of evidence is consistent with the proposal that STM for verbal materials other than signs is retained in ASL-phonology based codes generated internally and stored in the output buffer. According to the hypothesis of separate input and output phonological retention, nameable pictures and static pictures of ASL should be stored in the output
buffer just like written words. Therefore, it should result in differential STM performance from signed words, which are stored in the input buffer.

Method

Participants. Seven congenitally deaf participants who are fluent ASL signers were recruited for this experiment. All of these participants had learned and used ASL for at least 6 years on a daily basis, and ASL is one of the major ways they communicate. Each participant was reimbursed at the rate of 10 dollars per hour for their participation.

Materials, design, and procedures. The methods used in this experiment were identical to those in Experiment 3B, except for three major changes. First, the list length was reduced from 8 to 6 to be more comparable with previous studies with deaf participants in which a shorter list length was commonly used (e.g., Wilson & Emmorey, 1997, 1998). Second, different stimuli were used in this experiment. Six common words (frog, girl, heart, idea, queen, rat) that are easily differentiated from each other in ASL and in pictures were employed. One native ASL signer signing each of the 6 words was videotaped and edited as signed words in individual short films. The duration of each of these films was 1 second. Both the films and the pictures of these 6 words were displayed in the center of the computer screen to the participant. The third major methodological change was presentation rate. Due to the minimal time required to sign a word and the
limitation of the computer to load a film, the presentation rate of list items was 1.2
second per item, which was slightly longer than previous studies.

Only the mixed-modality lists composed of signed words and nameable pictures
were prepared. As in previous experiments, there were four conditions for each target
modality (i.e., a same-modality adjacent probe under the NI instruction, a same-modality
adjacent probe under the NISM instruction, a same-modality distant probe, and a
different-modality adjacent probe). While target serial positions 2 to 6 were examined for
the three same-modality conditions, only target serial positions 4 to 6 were tested for the
different-modality condition. The combination of 4 list types (SPPSSP, PPSSPP,
PSSPPS, and PPSSPP, in which S indicates the signed word and P indicates the nameable
picture), 6 words, and 5 serial positions of the target (except for the different-modality
condition) resulted in 120 lists for the NI instructional condition and 96 lists for the
NISM instructional condition.

Results

The same analyses were conducted as in Experiment 3B for lists consisting of
signed words and nameable pictures. Memory accuracy under different conditions was
computed across serial positions of the target, and depicted for both target modalities in
Figure 22. Although the STM performance at serial positions 2 to 6 was obtained for
some conditions in the present experiment, only the data from the serial position 4 to 6
Figure 22. Mean accuracy (in percentage) on serial probed recall of signed words and nameable pictures for four probe types across serial position 2 to 6 from Experiment 4.
was considered in the following analyses because only these serial positions were tested in the same-modality distant probe condition.

Visual inspection of the data revealed some interesting trends. First, memory performance for signed targets following different types of probes was equivalent across different serial positions. This was confirmed by the far-from-significant main effect of probe type for signed targets, \( F(3, 18) < 1, MSE = .06, p = .82 \). The mean accuracy for the four probe types was as follows: 80.5\% for SS(NI), 81.6\% for SS(NISM), 76.0\% for S..S, and 83.0\% for PS. Although no memory cost was associated with the PS condition relative to the SS(NI) and the SS(NISM) condition, the distance between the probe and the target did not seem to reduce recall accuracy in the S..S condition (except for the last serial position), either.

Second, for picture targets, the pattern of recall accuracy was more similar to the findings in Experiment 1 to 3 in which verbal materials from two buffers were presented in mixed-modality lists. The mean accuracy for the four probe types was 83.3\% for PP(NI), 80.2\% for PP(NISM), 68.1\% for P..P, and 67.1\% for SP. The main effect of probe type was significant, \( F(3, 18) = 4.98, MSE = .03, p = .011 \), and further comparisons verified that the PP(NI) condition was significantly better than the P..P (\( p = .024 \)) and the SP (\( p = .020 \)) condition, and the PP(NISM) condition was significantly better than the P..P (\( p = .039 \)) condition and marginally better than the SP (\( p = .068 \)) condition. Given
that signed items were remembered better than nameable pictures for the majority of
the participants, the lack of difference between the P..P and the SP conditions might be
caused by the higher memorability of the probe in the latter condition.

The 2 (target modality) x 4 (probe type) x 3 (serial position) within-participant
ANOVA only revealed a significant main effect for serial position, $F(2, 12) = 6.88$, $MSE$
$= .08, p = .01$. The main effect of target modality ($F(1, 6) = 1.99$, $MSE = .07, p = .21$) and
probe type ($F(3, 18) = 1.84$, $MSE = .05, p = .18$) were not significant. None of the
interactions was significant either (all $ps > .19$). Although the accuracy for signed targets
was better than that for picture targets, the magnitude of this modality effect (6.2%) was
substantially smaller than that from previous experiments. Five of the seven participants
showed the same pattern as the group data, while the other two participants showed the
opposite pattern. More participants were needed to determine the validity of the modality
effect between signed words and nameable pictures. The lack of significant effects may
be due to small sample and the highly variable performance.

Discussion

Because it was difficult to obtain many deaf ASL signers who can perform at an
acceptable level on the serial probed recall task to participate in the current experiment,
the data from this experiment should be interpreted with caution and the results need to
be replicated with more participants. Although no clear differentiation between the STM
for signed words and nameable pictures, as that between speech and other verbal materials, was observed, Experiment 4 did provide some evidence for such a separation. First of all, a distant probe in the same modality with the target did not result in worse recall than an adjacent probe in a different modality from the target. Moreover, while both the probe and the target were visually presented, an adjacent signed probe was a less effective cue to a picture target than an adjacent picture cue, even though signs were usually remembered better by these participants. These findings were more similar to those observed for lists relying on both the input and output phonological buffers, but not like the results for lists consisting of two kinds of visually presented items which were both retained in the output phonological buffer.

It should be emphasized again that, given the difficulty of the task and the small, variable sample, the findings from this experiment were by no means conclusive. However, they suggested that some interesting characteristics of ASL do provide researchers a tool to explore the organization of verbal STM. Despite its visual presentation as written words and nameable pictures, ASL is revealed over time as speech and is the natural linguistic input for deaf people. The hypothesis of separate input and output phonological retention predicts that signed words should be stored in deaf signers’ input phonological buffer, just like speech is stored in hearing people’s input
phonological buffer. The current experiment presented some preliminary data indicating the possible differentiation between STM for signed words and nameable pictures.

One caveat of the current experiment was that the difficulty of this task requires participants with better STM ability. However, it has been reported that working memory capacity is positively correlated with deaf people's ability to use speech-based encoding processes (Lichtenstein, 1998). If these high ability deaf participants were trying to internally generate the “speech” codes for the items in the list, all items would be stored in the output phonological buffer, hence eliminating the differentiation between input and output phonological retention. To avoid this problem in future research, a simpler serial probed recall task with mixed-modality presentation can be used so that participants with a wide range of STM ability can participate in the study. For example, the paradigm used by LeCompte and Watkins (1993) in which participants just need to remember the sequence of the list but not the presentation modality can be useful to determine whether there is a memory cost resulting from the modality switch for lists consisting of signed words and nameable pictures.

On retrospect, I should first examine whether deaf signers use ASL-phonology based codes to remember nameable pictures as in Experiment 3A. Although Wilson and Emmorey (1997) have demonstrated that this is the case in a serial recall paradigm, it is
worthwhile to replicate the same effect in the serial probed recall paradigm. The reason that I did not investigate this effect specifically in the current study was that it was difficult to recruit deaf signers and ask them to participate multiple experiments on the same nature. More participants would allow empirical validation of the phonological similarity effect of nameable pictures in STM for deaf signers. It would also be helpful to obtain a measure for the reliance of sign-based and speech-based STM of different types of verbal materials for each participant to determine whether the separation between input and output phonological retention is more pronounced for people who rely more on sign-based representations in verbal STM.

General Discussion

The current study employed different verbal materials in a serial probed recall paradigm to examine the phonological representation for retaining these items in STM. Specifically, spoken words, written words, words in lip-read films, and nameable pictures were presented in mixed-modality lists to college students in Experiment 1 to 3. In Experiment 4, words in ASL and nameable pictures were mixed together and were presented to deaf signers. According to the hypothesis of separate input and output phonological retention, spoken words and words in ASL, whose phonology is readily available for hearing and deaf people respectively, are retained in the input phonological
buffer, whereas other verbal materials, whose phonology requires internal generation, are retained in the output phonological buffer.

In general, the results from hearing participants are consistent with previous research in supporting this hypothesis. It was found that when the probe and the target were stored in two different buffers (e.g., spoken words versus lip-read words), recall accuracy of the target was higher following a same-modality distant probe than following a different-modality adjacent probe. There was also a memory cost associated with the modality switch between the probe and the target. On the other hand, when the probe and the target were in the same buffer (e.g., written words versus lip-read words), the same-modality distant probe was less effective than the different-modality adjacent probe, and there was no significant memory cost associated with the modality switch between the probe and the target.

The results based on deaf participants’ STM performance in Experiment 4 were not conclusive, however. The insufficient power of this experiment was mainly due to the small sample size and the great variability of participants’ STM ability. The pattern of the data showed that when words in ASL and nameable pictures were presented in a mixed-modality list, a distant probe in the same modality with the target did not result in better recall than an adjacent probe in a different modality from the target. The memory cost associated the modality switch between signed words and nameable pictures was not
strong enough to be detected, either. Close inspection of the data for signed and picture targets separately nevertheless revealed some evidence for the possible differentiation between these two presentation modalities. In particular, although the results for signed targets did not resemble the pattern observed for spoken targets mixed with other materials retained in output phonological retention, the results from picture targets did demonstrate the pattern implying some separation between these two kinds of materials. Given the uniqueness of the stimuli employed in this experiment and the great variability among the participants, no strong conclusion can be reached based on the current data. More research using a similar paradigm is needed to verify the prediction from the hypothesis of separate input and output phonological retention regarding the STM for ASL.

In the hypothesis of separate input and output phonological retention, it is reasoned that the phonology for lip-read items, as well as that for written words and nameable pictures, is generated internally and hence is an output phonological code. As a result, these items are retained in the output phonological buffer. However, the influence of visual information, specifically lip movements, on speech perception has long been documented as the well-known McGurk effect (McGurk & MacDonald, 1976). For example, when hearing the syllable /ba/ while seeing the articulation of /ga/, a majority of English speakers would perceive /da/ rather than the visual /ga/ or the acoustic /ba/.
Although the nature and the level of audiovisual integration reflected by the McGurk effect is not well understood yet, it has been reported that this effect was not influenced by word meaning or sentence context (Sams, Manninen, Surakka, Helin, & Katto, 1998). That is, even audiovisual integration would make a word become a nonword or make a sensible sentence become non-sensible, a robust McGurk effect was still observed. Based on these findings, Sams et al. (1998) concluded that audiovisual speech integration occurs at the phonetic perceptual level before the word meaning is extracted.

Given the influence of visual information on speech perception, one might argue that lip-read words would activation the input phonological representations for single word processing (i.e., speech perception), which in turn would be retained in the input phonological buffer. Although there is no empirical data available to evaluate this possibility, other accounts are equally capable of explaining the McGurk effect and output phonological retention for lip-read words. One possible explanation is that the McGurk effect is the product of audiovisual integration. While lip-movements directly activate output phonological representations for the words presented in lip-read films, they can only modify the input phonological representations whenever there is simultaneously compatible auditory information. In Experiment 2 of the current study, lip-read words were presented silently without any auditory input. To reach a phonological representation for the intended word, the participants had to generate the
phonology solely relying on the lip-movements. Therefore, it is more likely that the output, rather than input, phonological representation is activated and then retained in the output buffer in STM.

The hypothesis of separate input and output phonological retention is motivated by research on brain-damaged patients (e.g., Allport, 1984; Martin et al., 1999). Following this hypothesis, the separation between speech and other verbal materials whose phonology is internally generated is conceptualized as a distinction between input and output phonological retention. However, other organizational principles in verbal STM have been proposed to account for the observed differential memory performance for spoken and written items. For example, a visual STM capacity has been postulated to be responsible for increased digit span in mixed-modality presentation (Frick, 1984) and some patients’ reversed modality effect (Shallice & Vallar, 1990). However, as discussed earlier, the visual STM may be supplementary to phonological STM for verbal materials, and its influence is only observed when the phonological information is disrupted (as in brain-damaged patients) or not useful (as for homophonic stimuli). Moreover, Experiment 1A and 3A in the current study argued against an important role of visual codes in STM for written items and nameable pictures. As a result, it is unlikely that the findings from this study reflected the distinction between phonological and visual STM.
LeCompte and Watkins (1993) took a different approach in which the separation between spoken and written words, as well as other separation between different materials from their study, was explained by the perceptual similarity principle. However, the separation between spoken words and other verbal materials (e.g., written words and nameable pictures) cannot be solely accounted for by perceptual dissimilarity between these items, since no such separation was observed in the current study between other verbal materials which are also perceptually distinct from each other (e.g., lip-read words and written words).

As discussed earlier, some of the empirical findings from LeCompte and Watkins (1993) were not necessarily relevant to phonological codes in verbal STM (e.g., written digits presented rapidly under articulatory suppression). Otherwise, the hypothesis of separate input and output phonological retention offers an adequate account for most of the other results (e.g., the separation between spoken and written words but not between pictures and written words). One of their findings cannot be readily accounted for in terms of the use of non-phonological coding strategies – specifically, the dissociation between recall of words spoken by male and female voices. That is, one might have expected both to be coded in terms of input phonology and hence show no separation. One possible explanation of the observed differentiation could be due to a disruption in switching attention between two perceptual streams which caused a decrement in the
encoding of an item when a switch was made. A switch between male and female voices may be more detrimental to perception than a switch between written digits in two different locations because of the ambiguity of spoken input, which makes it harder to encode the stimuli. Martin, Breedin, and Damian (1999) found that participants did worse on same-different syllable judgments (e.g., ra - la vs. ra - ra) when the voice changed across the two syllables. Mullenix, Pisoni, and Martin (1989) also found that for words presented in noise, identification accuracy was higher and repetition of the word was faster if the same voice was used than if the voice changed across stimuli. The exact cause of the separation between male and female voices observed by LeCompte and Watkins (1993) is beyond the scope this study. Further experiments designed to examine the nature of the representations underlying these materials in STM are needed to directly address this issue.

Another candidate of the theories to account for the current results is the working memory model proposed by Baddeley and colleagues (Baddeley, 1986; Baddeley & Hitch, 1974). In this model, two components in the articulatory loop have been proposed, namely, a phonological store and an articulatory subvocal rehearsal process. Although these two components can function independently and can be selectively influenced by different factors, only one capacity is postulated for all verbal materials. Even if all written words, lip-read words, and nameable pictures require rehearsal to be stored in
phonological forms while spoken items do not, once the materials are stored in the
phonological buffer, memory accuracy should only depend on the temporal relation
between items. The better memory performance for two distant items in the same
modality than that for two adjacent items in different modalities cannot be accounted for
by the two components in the articulatory loop.

Campbell et al. (1988) have also offered an account for the separation between
STM in different modalities. According to their proposal, both lip-reading and auditory
presentation give rise to a phonetic code, but the phonetic representation is
underspecified for lip-read stimuli. Different from the hypothesis of separate input and
output phonological retention, however, Campbell et al. argued that despite the
differential linguistic representations for spoken and lip-read words, they are both
retained in one capacity, namely an input phonological retention capacity. The findings
from Experiment 2 in the current study provided a strong challenge to this view.
Specifically, it was found that spoken words were stored in a different buffer/capacity
with lip-read words whereas there was no such distinction between lip-read and written
words. The proposal of one capacity for both spoken and lip-read words cannot
adequately explain the fact that an adjacent probe in a different modality (albeit still in
the same buffer according to this proposal) results in worse recall than a distant probe in
the same modality.
Overall, the hypothesis of separate input and output phonological retention in verbal STM seems to provide an adequate account for neuropsychological evidence and findings from STM experiments. It is also supported by the data from Experiment 1 to 3 on hearing participants in the current study. In addition to replicating the differential memory performance for spoken and written items in serial probed recall, the findings from this study further extend the model of STM to account for other verbal materials (i.e., lip-read words and nameable pictures) based on whether the phonological information is readily available in the stimuli or requires internal generation. More research employing non-speech language (e.g., ASL) is needed to evaluate the generalizability of the hypothesis of separate input and output phonological retention in verbal STM for signed materials.
References


signers: The primary language coding hypothesis reconsidered. *Cognitive
Psychology, 22*, 211-224.


Shallice (Eds.), *Neuropsychological impairment of short-term memory*, pp. 287-
318. Cambridge, UK: Cambridge University Press.

words: Evidence for a long-term memory contribution to short-term memory


of deaf college students. *Journal of Deaf Studies and Deaf Education, 3*, 80-134.

effects in immediate verbal serial recall. *The Quarterly Journal of Experimental
Psychology, 53A*, 626-646.


Footnotes

1Since only the serial positions from 4 to 10 were tested and I did not know the recall accuracy for the serial position 3, the expected recall accuracy for the serial position 4 could not be computed.

2The difference between the LL(NI) and the LL(NISM) conditions was possibly due to participants’ different strategies under the NI and the NISM instructions. In the NI condition, participants might allocate memory resource to each item in the list evenly, regardless of presentation modality; in the NISM condition, on the other hand, they might allocate more memory resource to lip-read items, since less effort is needed to retain spoken items. As a result, the LL(NISM) condition might receive more attention than the LL(NI) condition, hence the better memory performance.