This article was originally published in *Learning and Memory: A Comprehensive Reference, 2nd edition*, published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues who you know, and providing a copy to your institution's administrator.

All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

https://www.elsevier.com/authors/author-schemas/artwork-and-media-instructions

http://dx.doi.org/10.1016/B978-0-12-809324-5.21036-5
ISBN: 9780128051597
Copyright © 2017 Elsevier Ltd. All rights reserved.
Academic Press
2.02 Encoding–Retrieval Interactions

Henry L Roediger, III, Eylul Tekin, and Oyku Uner, Washington University in St. Louis, St. Louis, MO, United States

© 2017 Elsevier Ltd. All rights reserved.

2.02.1 Encoding–Retrieval Interactions

Psychologists often give advice about how to improve learning and memory. Consider the following:

- Repetition of information, especially spaced repetition, improves memory.
- Practicing retrieval improves memory.
- Forming images of verbal material improves memory.
- Pictures are better remembered than words.
- Deep (semantic) levels of processing for material are better for memory than shallow (phonemic or orthographic) processing.

All these statements are true. Yet they are all false at the same time. That is, each of these generalizations implies that the variable described always improves learning and memory. However, boundary conditions exist for all claims about memory. In certain contexts, or with certain instructions, or with certain materials and especially with certain types of memory tests, the variables listed above as having a positive effect on memory may have no effect or even the opposite effect. In short, no general laws of memory exist (Roediger, 2008). We consider one form of this problem in our chapter.

2.02.2 The Memory Experiment

Theories and phenomena of memory have been explored for over 2000 years, and the essay by Aristotle and the dialogues of Plato are still worth reading today for their insights. Hundreds of philosophers and other thinkers advanced ideas and theories about memory over the ensuing centuries.

The greatest advance in the history of the study of memory occurred in the late 1800s: The memory experiment was invented (Nipher, 1876; Ebbinghaus, 1885). Before that, people discussed memory solely on the basis of introspecting on their personal experience or from observing and talking to (or reading the works of) other people. These are great means to generate ideas, but poor methods to determine which ideas have merit and which should be dismissed as wrong. Talk is cheap.

Experimental methods were developed and discussed in the middle ages and over the years became widely applied to study natural phenomena (physical and chemical phenomena in particular). However, they were not systematically applied to psychological phenomena until centuries later, in the 1800s. The basic idea of a memory experiment is to have people study material under controlled conditions, holding other factors constant, and observe the outcome on a measure of memory. Broadly speaking, Francis Nipher (1876, 1878) was the first to do this for short-term memory, and Hermann Ebbinghaus (1885) developed a method for examining long-term memory.

The memory experiment usually involves three phases. Events to be remembered occur under controlled conditions (with or without instructions to remember), a retention interval occurs (time, but often events of interest occur during this time), and finally some form of memory assessment (a test) occurs. These phases of the experiment are often called by the names of the hypothetical processes assumed to occur during them: encoding, storage, and retrieval. Variables can be manipulated during any of these stages. Since the invention of the memory experiment of this classic form, a huge amount of information about memory has been discovered. Not surprisingly, many facts uncovered accord with the ideas proposed by philosophers and others; on the other hand, many experimental results do not accord with prior writings (or were completely unanticipated).
and other findings seem counterintuitive even to the people within the experiment who produced the results (Bjork et al., 2013; Roediger and Butler, 2011). In this chapter, we consider an important variation on the basic memory experiment, the encoding–retrieval paradigm.

2.02.3 The Encoding–Retrieval Paradigm

In this paradigm, two sets of factors are manipulated, one set during learning (encoding) and the other set during the test (retrieval). That is, experimental subjects study material under two or more conditions and then are tested under two or more conditions. The very simplest case is represented in Fig. 1, which is adapted from Tulving (1983, p. 220) who first introduced the concept of the encoding–retrieval experiment. Two study (or encoding) manipulations (A and B) are crossed with two test (or retrieval) conditions (X and Y). If each column in the table is considered by itself (e.g., A vs. B manipulated and tested under the X condition), that arrangement is called an encoding experiment. Most experiments in the history of memory research are like this, seeking to determine how some variable manipulated at study affects performance as assessed on one test. This technique is fine as far as it goes, but often the further assumption is made that the variable affects memory in general, in statements such as “studying pictures rather than words improves memory.” Such statements implicitly assume that memory measures are interchangeable – if A produces greater recall than B, then it will reveal the same effect on all other measures. That assumption is generally false, as it turns out.

The rows in the table represent retrieval experiments, where encoding conditions are held constant and the interest is in how performance varies on different tests. Relatively few experiments of this type exist, but there are important exceptions (e.g., Tulving and Pearlstone, 1966).

Finally, when the entire 2 × 2 arrangement is considered, we have an encoding–retrieval experiment. Of course, several factors may be varied during encoding and more than two tests can be compared. The encoding–retrieval experiment permits researchers to assess the generality of encoding effects: What tests will reveal pictures be better remembered than words? What test might show no difference? Are there tests on which words lead to better performance than pictures? The answer to the last question is yes (Weldon and Roediger, 1987; McDermott and Roediger, 1994). Results from this sort of design constitute the topic of this chapter.

We should mention a final type of memory assessment, represented as a memory test in the cell in the lower right-hand corner of Fig. 1. Nothing is varied, but the design is still useful and in fact this is doubtlessly the most widely used memory test. This is a standardized test and there are many sorts. In virtually all Western nations, standardized tests are given to children to assess their knowledge (the SAT, GRE, MCAT, and many others in the United States), and similar standard memory tests are given to older adults to assess their performance. The aim of such tests, of course, is to study variation among people taking the tests, and their performance is assessed against standardization samples so that one can discover, for example, that an older adult is one standard deviation below people of his or her age on the test.

In the next sections, we review evidence on eight different uses of the encoding–retrieval paradigm to examine how study and test conditions do (and do not) interact. The proviso is that we consider only explicit or episodic memory tests in this chapter, ones that direct people to remember experiences from their past (Tulving, 1983). A rich literature exists on a comparison of explicit to implicit memory tests, with implicit tests (see Schacter, 1987) being ones that measure retention indirectly, through measures of transfer. However, these comparisons are beyond the scope of this chapter (see Roediger and McDermott, 1993 for a partial review). The eight manipulations we consider are ones of verbal context, mental operations, global environmental context, global feature context, background context (a specific form of physical context), variations in one’s bodily state (e.g., effects of taking drugs or not), one’s mental state (in particular, mood), and finally manipulation of physical operations performed during study and test. We will clarify the meaning of these terms as we proceed.

---

2.02.3.1 Verbal Context

Verbal context refers to conditions in which subjects are asked to remember target words, such as CHAIR, preceded by another word (e.g., glue–CHAIR). Of course, this is a standard tactic in the classic paired-associate learning paradigm, but in the late 1960s and early 1970s Endel Tulving put it to new use, varying verbal context between encoding and retrieval, which led to the encoding specificity hypothesis (or principle), to be discussed shortly. Tulving and his colleagues asked the fundamental question: What factors determine the efficacy of retrieval cues? This question was addressed in a series of experiments (Thomson and Tulving, 1970; Tulving and Osler, 1968; Tulving and Thomson, 1973) that provided an early use of the encoding–retrieval paradigm. During study or encoding, Thomson and Tulving (1970) asked subjects to study either weakly associated cue–target pairs (e.g., glue–CHAIR) or just the target itself (e.g., CHAIR). At retrieval, subjects either received a free recall test for the capitalized target word (i.e., with no cues), a cued recall test using weak associates to the target word (glue–?), or a cued recall test with strong associates to the target (table–?). For the subjects who studied single target words with no context, strong associates led to the best recall (.68), much better than with weak cues (.43) or no cues (.49). However, for the subjects who studied cue–target pairs with weak associates, cued recall with weak associates led to best recall (.82) and the strong cues actually led to worse recall (.23) than did no cues (free recall of .30). Encoding of CHAIR in the context of glue altered the encoding (perhaps eliciting associations to how furniture is made) so that even a strong associate of CHAIR like table was rendered an ineffective cue. Because of the changed verbal context during encoding, strong associates of the target word were completely ineffective and, if anything, distracting.

Roediger and Adelson (1980; to be discussed below) replicated the surprising finding that normatively strong associates could lead to poorer recall than even having no cues. Thomson and Tulving (1970) concluded that the effectiveness of retrieval cues depends on encoding circumstances (see too Tulving and Osler, 1968).

Tulving and Thomson (1973) asked further whether specific encoding might even lead to greater cued recall for a word not only in comparison with other recall tests (as in Thomson and Tulving, 1970), but also in comparison with a recognition test. Can recall of a word given a weak associate as a cue be greater than recognition of the word itself? Using the example above, when given glue, can recall of CHAIR ever be greater than recognition of the target word itself, i.e., recognition of CHAIR? Of course, at the time (and even now) the assumption is that recognition is a more sensitive memory measure than recall and will measure memories too “weak” to be recalled, so finding that recall can be greater than recognition in some circumstances is an interesting prediction, which Tulving and Thomson (1973) made from the encoding specificity hypothesis. In Experiment 1, during the study phase, subjects again studied weakly associated cue–target pairs (e.g., whisky–WATER). They then were given a generation task in which they generated six related words from strong associates to the target words (e.g., lake). They generated the target words 72% of the time. After the generation phase, subjects were asked to go back to the generated list and try to recognize any target words they had generated. Subjects recognized 23% of the generated words. Finally, they took a cued recall test with the weak associates provided during encoding (whisky). Recall was 68% in this case. In short, subjects recalled more targets with “weak” cues than they could recognize in the context of a strong cue (the word itself). This observation launched the study of the “recognition failure of recallable words” and generally showed that the result held under many conditions (e.g., Watkins and Tulving, 1975), although exceptions do exist (Gardiner and Tulving, 1980) and such exceptions seem to depend on how well integrated materials are during encoding.

Inspired by this prior research, other experiments addressed questions such as how the semantic relation between words affects recall. For instance, Roediger and Adelson (Experiment 3, 1980) addressed the role of semantic congruity between study and test conditions for cued recall, using homographs (words spelled the same but having two distinct meanings) as targets. Subjects studied word pairs in which the encoding context for the target word was manipulated and the retrieval cue was held constant. For example, for the target word ADDRESS the cue used at testing was speech. The encoding conditions involved providing a close synonym of the target with it (e.g., talk–ADDRESS), a weaker synonym (e.g., oration–ADDRESS), or a synonym for a completely different sense of the word (e.g., residence–ADDRESS). The assumption was that the last condition would provide a quite different encoding of the target word that was incongruous with the retrieval cue speech. The results supported the prediction of the encoding specificity hypothesis: When scored for recall of either part of the encoded event (either the cue or the target—talk or address—was scored as correct recall of the unit) to the extralist cue speech, recall was much higher in the two congruous meaning conditions (.43 and .37 for the two similar meaning conditions) and quite low for the incongruous condition (.08). The latter value was lower than free recall (.15), indicating that subjects would have been better off covering up the cues in this case and simply recalling the words using whatever cues that came to mind. Thus, the semantic match between the context word during the study and the retrieval cue during the test determined recall, in line with the encoding specificity hypothesis.

Similar results showing semantic flexibility of materials were observed also with sentences. In their study, Barclay et al. (1974) asked subjects to study sentences such as The man lifted the piano or The man tuned the piano and later presented them retrieval cues such as something heavy or something with a nice sound during recall. Subjects were instructed to recall related nouns from the sentences (e.g., piano) when tested. Although both cues could refer to the target word, when the cue tested the same aspect of the word as that encoded, recall was much greater. The results are in Fig. 2, Panel A. These results do not depend on using homographs; the words in the Barclay et al. (1974) experiments were not ambiguous in meaning, yet the same pattern still occurred. In addition, sentences are more natural language materials than paired associates, helping to extend the encoding specificity principle more widely.

The literature on encoding specificity with verbal materials is vast, and most experiments used the encoding–retrieval paradigm. We report only a few more experiments in this section that extend the principle in interesting ways with verbal materials. Marian and
Neisser (2000) conducted an autobiographical memory study with Russian–English bilingual students who had immigrated to the United States. Subjects were asked to tell a detailed event from their lives according to 16 prompt words, and half of these words were in Russian and the other half in English. When the prompt words were in Russian, subjects recalled more Russian autobiographical memories, but when the prompt words were in English they recalled more autobiographical memories from their time in America. Fig. 2, Panel B shows these results. The language used for the prompt helped determine the probability of recalling a memory from the era of their lives in which they spoke the language on a daily basis.

Another interesting study by Marian and Kaushanskaya (2007) revealed language dependency for semantic knowledge. Mandarin–English bilingual subjects answered 16 general knowledge questions presented either in Mandarin or in English. These questions were considered bivalent questions because they had answers in both of the cultures (e.g., Name a statue of someone standing with a raised arm while looking into distance with two possible answers as the Statue of Liberty and the Statue of Mao). The results revealed that when the questions were asked in Mandarin, subjects responded more with answers coming from general knowledge acquired in Mandarin, and when the questions were asked in English, they responded more with answers coming from general knowledge acquired in English. Again, the language in which the knowledge was learned affected recall of the knowledge. See the results in Fig. 2, Panel C.

The two studies by Marian and colleagues helpfully extend the encoding specificity principle to autobiographical memory (Marian and Neisser, 2000) and semantic memory (Marian and Kaushanskaya, 2007; see also Muter, 1984). In these cases, the encoding manipulation (language of acquisition during learning of personal events or semantic knowledge) is provided by nature and then the retrieval cues are selected to match one or the other encoding context, so the studies do conform to the encoding–retrieval paradigm. In addition, Marian and Fausey (2006) found language-dependent recall in a study in bilinguals in which they manipulated both encoding (i.e., the language of a story) and retrieval (i.e., the language of the questions about the story), thus conforming to the standard encoding–retrieval paradigm more closely. Once again, the recall was better when the language of the questions matched the language used in learning the story than when the two mismatched.

Another factor that determines the effectiveness of retrieval cues is the number of events that are tied to a specific cue. For example, Tulving and Pearlstone (1966) had students learn lists of categorized words with varying numbers of words in the different categories. They showed that the more words studied in each category, the lower the probability that a specific word would be recalled. Watkins and Watkins (1975) called this the cue overload principle, and later Nairne (2002) referred to it as cue discriminability. We turn now to other types of encoding–retrieval experiments.

### 2.02.3.2 Mental Operations

The encoding specificity hypothesis assumes that it is not the event or stimulus from the outside world that determines later memorability, but rather how that event is encoded and how well the cues in the retrieval environment match and discriminate the encoded event. A natural extension is to examine encoding–retrieval interactions manipulated by other kinds of mental operations or types of processing used during encoding and retrieval.

Transfer-appropriate processing, a term coined by Morris et al. (1977), is one such extension of the encoding specificity ideas to broader mental operations. This approach was inspired by (and in some ways critical of) the levels-of-processing approach of Craik and Lockhart (1972), which argued that remembering becomes better due to the “depth” or level of encoding. For example, words might be coded in terms of orthographic characteristics (e.g., judging the case of the letters in the word, upper or lower), the sound of the word (phonemic properties, such as words it rhymes with), and the meaning of the words (the semantic properties). This order from superficial appearance to sound to meaning is from shallow to deep levels of analysis within the
levels-of-processing framework, and many experiments show that when words are processed in these various ways, recall and recognition on standard tests are markedly affected (e.g., Craik and Tulving, 1975). Morris et al. (1977) argued that this standard finding arose because the types of test typically used relied on meaning-based information. They posited that if a test required knowledge of the orthographic or sound properties of words, then the so-called shallower levels of processing would lead to better performance than would deep, semantic-based encoding. That is, Morris et al. argued that there was nothing inherently "better" in any encoding condition; rather, the learning goals and the type of retrieval task also helped determine performance. Put another way, performance on a test will be enhanced to the extent that the encoding operations transfer appropriately to the retrieval task, hence the name of the framework.

Morris et al. (1977) tested this hypothesis in three experiments that yielded similar results. Subjects studied a list of words, half with a semantic and half with a phonetic acquisition task. The semantic acquisition task was to decide whether the presented word fit in a given sentence, whereas the phonetic acquisition task was to decide whether the presented word rhymed with a given word. Equal number of questions from each acquisition task had yes and no as correct responses, but we will focus only on the results from yes responses. Critically, subjects were tested in two different ways: a standard recognition and a rhyme recognition test. Half of the subjects took a standard recognition test, where they indicated the words they studied among other unstudied words. The other half took a rhyme recognition test that included words they never studied, but some words that rhymed with what they studied and some that did not. Subjects’ task was to indicate the ones that rhymed with the words they studied. Morris et al.’s prediction was that the levels-of-processing effect would occur in the standard recognition test, where the words studied with a semantic acquisition task would be recognized more than those studied with a phonetic acquisition task. However, they predicted the opposite pattern for the rhyme recognition test, because it was more transfer appropriate to test memory for words studied with a phonetic acquisition task with a rhyme recognition test. Overall, semantic acquisition words were recognized more than phonetic acquisition words, but the interaction between the encoding and retrieval conditions fit well with the predictions: Semantic acquisition words were recognized more than phonetic acquisition words in the standard recognition test, but the pattern reversed for the rhyme recognition test. Matching the mental operations of study and test (i.e., semantic acquisition with standard recognition and phonetic acquisition with rhyme recognition) enhanced memory.

Another way to manipulate mental operations during encoding and retrieval is to contrast generation and reading. Glisky and Rabinowitz (1985) showed that when subjects generated words during recognition that they also generated during study, they recognized more of those words compared to when they just read the words and made a recognition judgment. During study, subjects generated half of the study list from word fragments and read the other half of the words that were intact. Subjects then took a recognition test, but half of the words were in fragments and the other half were intact. They generated words from the fragments and read the intact words, and then made a recognition judgment for each word. Overall, there was a generation effect (Slamecka and Graf, 1978), where words generated during study were recognized more than words that were read. However, generating words at test only helped recognition of words that were also generated during study, not for words that were originally intact. Put another way, reinstating the mental operations of study during test (i.e., generate at study–generate at test) enhanced recognition memory. Note that this was not the case for the match between the reading conditions at study and test.

Dewhurst and Brandt (2007) replicated the findings of Glisky and Rabinowitz (1985) in two experiments and extended them by changing the study material and asking subjects’ recollective experience at test. In Experiment 1, subjects generated some words from fragments at study and test, and in Experiment 2, subjects generated some words from anagrams at study and test. Both experiments showed similar results: Generating words during study increased recognition of words generated and read during test, but more importantly recognition memory was enhanced when the same words were also generated at test. In both experiments, they asked subjects to indicate their recollective experience during recognition by giving a remember, know, or guess response after each recognition judgment (Gardiner et al., 1996; Tulving, 1985). The results of Experiment 1 are in Fig. 3. Subjects gave more remember

Figure 3 Results of Dewhurst and Brandt (2007). Error bars indicate standard error of the mean. Recognition results are provided only for items given remember responses at test.
responses when words generated during study were also generated at test, but not when words read during study were also read at test. Dewhurst and Brandt (2007) concluded that matching effortful encoding and retrieval conditions (i.e., generation) enhanced recognition memory and was accompanied by a conscious recollection of encoding (i.e., more remember responses), whereas matching automatic encoding and retrieval conditions (i.e., reading) did not show a similar pattern.

Reinstating mental operations of encoding during retrieval may not always enhance memory; therefore generalized statements should be made with caution. For instance, matching reading conditions during encoding and retrieval did not increase recognition accuracy (Dewhurst and Brandt, 2007; Glisky and Rabinowitz, 1985). Similarly, memory was not enhanced when generation and reading conditions at study and test were manipulated between subjects (Dewhurst and Knott, 2010; Mulligan and Lozito, 2006).

In line with the general theme of this chapter, the way information is encoded interacts with the retrieval task and determines subsequent memory performance. According to Hunt and Einstein (1981), there are two qualitatively different ways to process information: relational processing and item-specific processing. Relational processing integrates and organizes the input by focusing on commonalities of separate events and leads to an effective search strategy, whereas item-specific processing increases distinctiveness by focusing on features discriminating an event from others in the input. Hunt and Einstein (1981) argue that both are useful during retrieval and show that the effectiveness of the type of processing is related to the type of retrieval task.

In Experiment 1, subjects studied one of two word lists. One was a categorized list of six nouns from six categories, which by nature allowed relational processing due to obvious categories (e.g., fruits, animals). The other list was also categorized with six nouns from six categories, but the categories were less obvious (e.g., liquids, things that women wear) and the words were perceived to be unrelated, which allowed item-specific processing due to less obvious categories. Subjects studied the words with one of two orienting tasks. The categorization task required subjects to sort the words into the appropriate category, which emphasized relational processing, whereas the pleasantness rating task required subjects to rate the pleasantness of each word, which emphasized item-specific processing. Hunt and Einstein argued that memory performance will be highest when the structure of the list (i.e., related or unrelated words) and the orienting task (i.e., categorization or pleasantness rating) encourage a different type of processing. Put another way, they expected an interaction between the two variables, where rating pleasantness of related words will increase memory for those words compared to categorizing them and where categorizing unrelated words will increase memory for those words compared to rating their pleasantness.

The results confirmed their prediction, but differed for recall and recognition, which again showed that memory performance was determined by how the retrieval task interacted with the encoding conditions. For instance, related words were better recalled, but unrelated words were better recognized. Also, engaging in both orienting tasks for the same list (related or unrelated) increased recall compared to engaging in either one of the orienting tasks twice; however, engaging in pleasantness rating that allowed item-specific processing increased recognition whether or not subjects also categorized (i.e., engaged in relational processing). These results suggest that the effectiveness of the type of processing (i.e., relational or item-specific information) depends on the type of retrieval task. Recall benefits from both types of processing and recognition benefits from item-specific processing.

Overall, when mental operations used to encode information are reinstated during retrieval, retention is likely to be enhanced. Put another way, retrieval success is not dependent only on the encoding or the retrieval task, but also the match between mental operations engendered during encoding and retrieval.

2.02.3.3 Global Environmental Context

When we go back to places such as our old high school or neighborhood, we all have the experience of remembering events that occurred at those places, often quite vividly. Seeing a house where you used to live or your old classroom triggers memories that might never have been recollected otherwise. In these cases, the physical environment itself or some feature of that environment that occurred at those places, often quite vividly. When we go back to places such as our old high school or neighborhood, we all have the experience of remembering events that occurred there.

The phenomenon of physical context–dependent memory refers to the case of when re-presentation of an environmental context enhances retention for the events originally encoded in that environment relative to their retention in a different context. Many experiments on this subject used reinstatement paradigms following the logic of the encoding–retrieval paradigm. That is, researchers use two (or more) different physical contexts during learning of material and then testing occurs in either matching or mismatching contexts; so the physical context is either reinstated or not. According to the encoding specificity principle, better retention is expected to occur in the matching context condition. Yet, the literature is rather inconsistent on whether this context reinstatement effect occurs with short retention intervals; at the very least, it is quite variable. One problem at the outset is whether the context in which information was learned is indeed part of the encoding of the event. If the context is incidental to learning, it may not be encoded with the information to be learned and thus manipulating the context at retrieval may have little or no effect.

The first studies exploring the role of environmental context using the encoding–retrieval paradigm produced impressive results. For example, Godden and Baddeley (1975) reported an oft-cited study conducted with scuba divers who learned a list of unrelated words either underwater or on dry land (while wearing a wet suit in both cases) and then recalled them a short time later either in...
a matching context (i.e., water–water, land–land) or in a mismatch context condition (i.e., water–land, land–water). As the encoding specificity principle would predict, the number of recalled words was higher in the matching contexts than in the mismatching contexts. See the results in Fig. 4, Panel A.

In the same vein, Smith et al. (1978) obtained similar results when they used two distinct rooms as different contexts. In Experiment 3, subjects either studied categorized words in room A or in room B and recalled the words either in the same or the other room a day later. Room A was on the second floor, decorated with technical equipment, had tile floors, and the experimenter wore a T-shirt and jeans, whereas room B was in the basement, decorated with pictures and plants, had carpet floor, and the experimenter wore a suit and a tie. The test was free recall, and the results confirmed the physical context dependency of memory, showing enhanced memory performance under the same context condition relative to the encoding–retrieval mismatch condition.

Building on these robust findings of context dependency, Smith (1979) found that being physically present in the previous environment was not a necessity for enhanced retention by showing that just imagining the previous environment in which the materials were studied or viewing photos of it was enough to enhance recall. He also found that when the mental reinstatement of the initial learning room was made more difficult by presentation of distractor rooms during study, the effect disappeared. If one has difficulty remembering features of the room, then attempting to reinstate the context did not work.

The initial studies reviewed above generally produced robust effects. Yet the replicability of these promising reinstatement experiments was challenged by new experiments that failed to replicate some of them during the 1980s. Notably, Fernandez and Glenberg (1985) manipulated physical environment during encoding and retrieval conditions in a series of eight experiments similar to those of Smith et al. (1978); not even one obtained the encoding–retrieval interaction with same and different physical contexts in recall. They concluded that the effect was not reliable. These failures to replicate were in between-subjects conditions, so other researchers tried manipulating room context within subjects. Bjork and Richardson-Klavehn (1989) found no effect of same and different room contexts using a within-subject manipulation.

In addition to these, Saufley et al. (1985) could not find any effect using actual university classroom settings with large numbers of students at the University of California at Berkeley. The students all heard lectures in a large university lecture hall; then, for the tests, they were randomly assigned to the same large lecture hall or to ancillary rooms. (This step was taken to permit students to spread out with intervening desks between them.) They found no difference in performance despite testing thousands of students over several semesters on multiple choice tests. However, as we will see in a later section, often encoding–retrieval interaction effects that are routinely obtained in recall vanish when a recognition test is used. Because multiple choice is one type of recognition test, this feature may have been the culprit in this finding. Other findings also supported the claim that context dependency does not occur with recognition tests (Eich, 1985; Godden and Baddeley, 1980; Smith et al., 1978). The idea is that recognition cues are so strong and item specific that the more subtle environmental cues are blocked from exerting any effect (Smith, 1988).

Why do recall results of experiments manipulating room context differ from one another so greatly? One idea, just mentioned, is that the nature of cues in the retrieval environment helps determine whether or not the effect will occur. Depending on the number of noncontextual cues, the effect of physical context dependency might increase or decrease. Recognition tests provide powerful item-specific cues that can provoke recollection no matter what the location of their presentation, which may render subtle context cues ineffective as retrieval cues. On the other hand, free recall does not provide as many strongly associated cues, so perhaps the context cues would come into play in guiding retrieval. Yet in a series of experiments, McDaniel et al. (1989) demonstrated that even when encoding involved orienting tasks such as imagery, organization, and self-referencing, the physical context–dependency effect disappeared in free recall. They concluded that these elaborative activities led to a greater and stronger number of cues and thus overshadowed the context cues.

![Figure 4](results.png) Results of Godden and Baddeley (1975), Isarida and Isarida (2004), and Isarida et al. (2012), respectively. Error bars indicate 95% confidence intervals.
Another possible reason for not obtaining environmental context effects in the lab relative to vivid examples in our lives is that the retention interval between the encoding and retrieval of an event is much shorter in lab experiments. With longer retention intervals, the power of the environment as a cue might increase and this might lead to better recollection. This possibility has not been much examined because it is difficult to instantiate realistic retention intervals, on the order of years, in lab experiments, although we refer to some relevant findings below that suggest that retention interval does have an effect even when varied only over a week or so.

Yet another possible reason relates to one of first points mentioned: Perhaps the encoding context does not actually form an integral part of the encoding to the to-be-learned material. Eich (1985) asked subjects to generate presented words either as isolated visual images or as images integrated with the feature of the environment (e.g., if elephant were presented, they would imagine an elephant sitting on a chair in the room). Eich found that subjects showed a context-dependency effect only in the integrated image condition, showing how associative encoding instructions could influence context dependency. Thus the effect did occur when the context was tied to the to-be-remembered event during encoding. Still, such a finding could not explain why some experiments using room contexts obtained large context-dependency effects (e.g., Smith et al., 1978), whereas close replications of those experiments (some of which used the same rooms as used in Smith et al.) did not (Fernandez and Glenberg, 1985). The failures to replicate seemed to indicate the context-dependency effect using room context was, at best, a finicky phenomenon. Isarida and Isarida (2014) reported that during 1990s they could only find three studies that used physical context reinstatement paradigms.

Due to these conflicting results in the physical context–dependency literature, Smith and Vela (2001) attempted to resolve the issue by conducting a metaanalysis via estimating the effect sizes for different variables. They used 93 effect sizes from 75 experiments that were conducted between the years 1935 and 1997. They included experiments in which the environmental context was incidentally manipulated and global (i.e., slow changing), which fits our definition of physical context. They also only included environmental contexts that were extrinsic to the target stimuli, meaning they did not include intrinsic context manipulations such as visual or auditory manipulation of the material. Lastly, they excluded studies outside of the laboratory like classroom reinstatement because it was harder to control for possible confounds. The metaanalysis revealed that the physical reinstatement paradigm had a reliable but modest effect size (d = .28), meaning the match of the environment between the encoding and retrieval had an effect on memory performance. Also, if the remembered items were not semantically or associatively related, the effect size was larger (d = .33) relative to when the materials were associatively encoded (d = .13). Again in line with the previous explanation of mixed results, the effect size increased with longer retention intervals, with 1 day to 1 week resulting in the largest effect (d = .63). Surprisingly though, unlike some findings taken alone, the metaanalysis did not find any difference between recall and recognition as a function of environmental context. The authors concluded that environmental context had modest effects in both tests. However, the limitation in this and all metaanalyses is that they can only be conducted on experiments that are published, and null results are often not published. Thus, the metaanalysis may overestimate the actual effect. Nonetheless, their metaanalysis provided some clarification for the mixed results of environmental context and opened the way for new studies.

Inspired by the Godden and Baddeley (1975) study with scuba divers, Thompson et al. (2001) conducted a similar study with professional skydivers. In Experiment 1, subjects learned 20 unrelated words either on land or in the air in parachutes and recalled them after 8 min in one of the two environments. When learning occurred on land, physical context dependency was obtained, as being tested on land yielded higher recall than being tested while parachuting. However, when learning occurred in the air, recall was really poor for both the same and the different context conditions (probably due to the stress of parachuting leading to lack of attention to encoding the words). In Experiment 2, they used a video clip of skydiving instead of the actual act to minimize this problem. Subjects studied words presented through audiotape either while watching a skydiving video or not, and this time, physical context dependency was obtained. In both encoding conditions (learning while watching a video or not), testing in the same context condition produced better recall. Using a more modest manipulation, McKone and French (2001) had students study words in a quiet room in a library or in an open outdoor and grassy space. Ten minutes later, they recalled more words when tested in the same context as the one they learned in. This study is notable because they used stem cued recall (e.g., cha___ for chair) and stem cues are item specific (like recognition tests).

Isarida and Isarida (2004) conducted a series of experiments to clarify the conditions in which room-dependent memory occurs. In Experiment 1, they manipulated the place (room A and room B) and the task performed before and after studying words (task A and task B). The two tasks were a 30-s calculation task and a 30-s fine motor skill task. Subjects studied 20 unrelated words and were required to generate sentences with the words during study in either room A with task A or in room B with task B. The tasks were completed before and after studying of a block of five words, with four blocks of words (and five 30-s blocks of tasks). After a 10-min retention interval, subjects recalled words either in room A with task A or in B with task B; before the test they completed the assigned task again as part of reinstatement of environment. The results revealed that subjects recalled more words when the study and test contexts matched in terms of place and task, but of course in this experiment the authors deliberately confounded the room and the task, so no firm conclusion about room dependency can be drawn. Fig. 4, Panel B shows these results. In Experiment 2, they kept the same procedure but just manipulated the place without the task, and in Experiment 3, they just manipulated the task but not the room. In neither of these experiments were context-dependency effects obtained, leading to the conclusion that both factors operated together to produce the effect. Of course, it would be comforting to see the original effect, where context dependency was obtained when both the encoding and retrieval tasks and room matched, replicated.
Based on the results coming from this study and other physical context manipulations, Isarida and Isarida (2014) differentiated between a complex-place context and a simple-place context. Manipulations that involve not just place but also the experimenter or the subsidiary task are considered complex-place manipulations, whereas manipulations that involve only the place are considered simple-place manipulations. They point out that studies have a higher chance of finding an effect of physical context dependency when they use complex-place manipulations (a room combined with some other tasks, such as watching skydiving or performing another task). This is in line with previous studies that revealed context dependency. Most of them used a physical context that fits with the complex-place definition (Smith et al., 1978) or if they were just manipulating the place, the change was profound (Godden and Baddeley, 1975). From the encoding specificity perspective, complex-place manipulations may provide more features during encoding and more retrieval cues, which together may permit a better match between conditions of encoding and retrieval and hence greater recall under the match conditions than the mismatch conditions.

In their metaanalysis, Smith and Vela (2001) reported a reliable effect of physical context dependency with recognition tests despite the equivocal findings in the literature, and one possible reason for this discrepancy is type of material. They included studies with both meaningful and meaningless or unfamiliar stimuli, and the context-dependency effects in recognition were mostly observed in studies that used unfamiliar faces. In their Experiment 1, Smith and Vela (1992) staged a live event of a person coming and talking in front of an intro psychology lecture for a minute. Subjects were asked to identify the confederate from 10 mug shots (i.e., a recognition test) either after 1, 2, or 7 days in either the lecture class or in a quite different room. At every retention interval, subjects in same context condition recognized the confederate better than those in the different context condition, demonstrating a context-dependency effect in recognition. One important feature of this study was that subjects only had to recognize a single target.

Other studies have used multiple targets in old/new recognition tests. Dalton (1993) investigated the role of familiarity in face recognition in a room context paradigm. She familiarized the subjects with 36 of 72 faces 1 week before the study and test sessions, so half would be somewhat familiar and half would not. Then she had subjects study all 72 in one of two physical contexts and then tested them after a short delay in either the same or different context. Dalton obtained a context-dependency effect in face recognition, but only for the novel faces (i.e., the ones not prefamiliarized). Russo et al. (1999) replicated Dalton’s (1993) finding with faces (Experiment 1) and then extended the same logic to verbal materials in later experiments. In Experiment 2, they used pronounceable nonwords (items such as flirp), reasoning that nonwords were not familiar to subjects and thus would show the context-dependency effect even with recognition testing. As they predicted, subjects were better at recognizing nonwords when they were tested in the same room where they had studied them relative to being tested in a different room. In Experiment 3, they used familiar words as materials and failed to find a physical context–dependence effect using a recognition test. They concluded that familiarity of materials mediates the context-dependency effect in recognition tests.

More recently, Isarida et al. (2012) attempted to clarify the equivocal findings of physical context dependency with recognition tests. They noted that most of the studies that used meaningful words as materials and failed to find context dependency used study times of at least 3 s per item. They suggested that study time might be another factor that mediates context dependency in recognition. In a first experiment, they used 80 unrelated words, and half of the subjects studied them at a 1.5-s rate and the other half at a 4-s rate. For the same context, they used a match between the room, experimenter, and subsidiary task across the study and test sessions, and for the different context condition, the room, experimenter, and task were different. In the 1.5-s condition, subjects had higher recognition rates for the words when they were tested in the same context condition; however, the context-dependency effect was not found for the 4-s condition. See the results in Fig. 4, Panel C. In another experiment, they compared words and nonwords at a slow rate and replicated previous findings that the context-dependency effect occurred for nonwords only at a slow rate. Thus they concluded that study time and meaningfulness/familiarity of items contribute to the context-dependency effect in recognition tests. They argued that the power of copy cues provided on the recognition test is reduced at fast presentation rates during study or with unfamiliar items, permitting the global environment to act as a cue even during a recognition test. Of course, these results and conclusions await further testing and confirmation.

2.02.3.4 Global Feature Context

A related line of research to that just reviewed has examined the power of particular environmental features while holding the physical context (e.g., the room in which the experiment is conducted) constant. These features are still considered global contexts because they apply to all stimuli in that environment (Isarida and Isarida, 2014). For instance, Geiselman and Glenny (1977) familiarized subjects with a male and a female voice before their study session and asked subjects to rehearse each visually presented word either in the male voice, female voice, or in their own voice for 10 s during the study session. Subjects were tested auditorily in a recognition test with words presented in either the male or the female voice. When the voice of the word during the test matched with the rehearsed voice during study (i.e., male–male or female–female), subjects recognized the words better relative to the mismatch conditions. In another interesting study using the same idea, Grant et al. (1998) manipulated background noise of the room while keeping the room constant between study and test sessions. Subjects studied an article either in silence or with university cafeteria noise at the background, and they were tested with short-answer questions (i.e., cued recall) either in the match or the mismatch noise context. The results revealed feature dependency with both matching conditions producing better recall than mismatching condition. It seems surprising that even subjects in the noise–noise match condition performed better than subjects in the mismatch conditions. The results are shown in Fig. 5, Panel A.
Odor is one of the most common features that is used in these feature-context studies, because it is considered a strong retrieval cue especially in longer retention intervals (Isarida and Isarida, 2014). With retention intervals ranging from 5 min to 48 h, many studies have demonstrated that odor is an effective physical cue at retrieval that leads to better recollection through free recall and recognition, and even with prose passages (Cann and Ross, 1989; Herz, 1997; Isarida et al., 2014; Pointer and Bond, 1998; Schab, 1990; Smith et al., 1992). For example, Parker et al. (2001) found an odor context-dependency effect with a retention interval of 4 weeks. Subjects rated 30 words for pleasantness and familiarity either while breathing a lemon odor, a lavender odor, or no odor. Four weeks later, they were given two tests under conditions of the same or different odor – free recall then recognition. The results revealed that if there was an odor present at the study, subjects performed better overall than if there was no odor. In addition, if the odor during the test matched the original odor during the study session, subjects performed better both on recall and recognition (compared to the mismatch odor and no odor conditions). These results support the power of particular environmental features as well-encoded and good retrieval cues.

Aggleton and Waskett (1999) conducted a remarkable study on physical context dependency with odors. Their subjects were people who had visited the Jorvik Viking Centre, a museum on Vikings in York, England, almost 7 years before the study. One of the features of the museum involved having a combination of seven evocative smells for the full “Viking experience.” As a recall test, they created a questionnaire of 20 questions about the items in the museum that were visible from the cars that carried people around during their visits. They obtained all of the seven smells (i.e., the Jorvik odors) of the museum and assigned subjects to three conditions to test them with the questionnaire; all groups took the questionnaire twice. The first group got Jorvik odors during the first test, and for retest they got control odors. The second group got the control smells during the first test and then Jorvik odors for the second test. The third group did not receive any odors with either of the questionnaires. Both the first and second groups were encouraged to smell the odors during test and retest. From the first test to retest, only the second group slightly improved their performance with the addition of Jorvik odors, whereas the other two groups showed no significant difference between tests. The results suggest that odors aided recall of relevant information after 7 years. See Fig. 5, Panel B.

Background music is another global context that produces a context-dependency effect in free recall (Balch et al., 1992; Smith, 1985). However, whether background music leads to better recall due to physical context dependency or rather to possible changes in mood induced by the music (and therefore a mood-dependency effect, discussed later in the chapter) is still up for debate in the literature. On the one hand, previous research indicated that better recall only occurred in conditions where music tempo or tonality matched between encoding and retrieval, which were also two aspects of music that affected the arousal and pleasantness of experience for the subjects (i.e., mood of subjects), respectively (Balch and Lewis, 1996; Mead and Ball, 2007). These studies used the same music piece and only manipulated either the tempo or the tonality of the music. They concluded that tempo and tonality lead to mood changes, and the mood match between encoding and retrieval leads to better recall. On the other hand, more recently, Isarida et al. (2016) addressed some methodological concerns using the same piece of music with only tempo or tonality manipulations. In their experiments, to insure that it is only tempo or tonality that mediates better recall, they had three conditions for background music during retrieval: same music piece with same tempo or tonality (SS), different music piece with same tempo or tonality (DS), and different music piece with different tempo or tonality (DD). If the effect is caused purely by tempo or tonality and not by the specific piece of music being played, the SS and DS conditions should lead to similar and greater levels of recall compared to the DD condition; if the effect is not mediated by tempo or tonality, SS should lead to better recall than both DS and DD. The results confirmed the second possibility, because subjects in the SS condition recalled more than did subjects in both the DS and DD conditions. Further, subjects in DS and DD conditions did not differ from one another. Based on their results, Isarida et al. concluded that tempo or tonality, and therefore mood, did not mediate the observed context-dependency effect. Rather, the effect seems to be a true context-dependency effect created by the specific piece of music being played during both encoding and retrieval.
Across many studies, physical context match between encoding and retrieval either in the form of place, odor, or mixture of a couple of dimensions has a reliable but modest effect and can serve as a retrieval cue for recall and even for recognition. However, the findings on feature dependency should be interpreted with caution because the published effects are usually small and null results might be hidden in file drawers. In the next section, we consider local context manipulations that lead to context dependency.

### 2.02.3.5 Background Context

Background contexts are different from general physical contexts on several dimensions. One of the differences is related to how specific each context is. On the one hand, physical contexts are more general and capture multiple aspects of the environment. Consider the classic study by Godden and Baddeley (1975), in which subjects studied and recalled a list of words underwater or on land. The dry/wet dimension, the light, and the type of living things in the environment are only a few of the many different aspects of each environment. On the other hand, background contexts are more specific. Unlike physical contexts, they have one or only a few aspects that can vary (much like the research on specific environmental contexts described in the previous section). A background context typically refers to the screen color, the location on the screen, or the picture or the ongoing video used as background for to-be-remembered items. The items are presented in the foreground, as the primary task, but the background varies. A second difference between the two contexts comes from the speed at which they can change. Physical contexts change slowly if at all; however, background contexts can change rapidly. Third, the two contexts differ in the number of items associated with them. In physical contexts, more items are typically associated with a single context; on the contrary, in some background context-dependency studies, only one item is presented on each context. Broadly, the difference between physical and background contexts can be understood as the difference between global and local contexts (Glenberg, 1979; Isarida and Isarida, 2007). In sum, background contexts are more specific, faster to change, and associated with fewer items compared to physical contexts, which makes it easy to study background context manipulations within laboratory settings.

Although types of background context vary among experiments, most studies converge on one conclusion: When the background context of encoding is the same with the background context of retrieval, subjects usually remember more compared to when the contexts do not match (i.e., context dependency). We will start with simpler background contexts and move on to more complex ones.

One of the simplest background contexts to manipulate is the color of the screen on which the words are presented. Dulsky (1935) conducted a series of experiments to examine how the change of background color affected recall. Subjects learned 10 pairs of nonsense syllables and recalled the target syllable given the cue syllable immediately after study across three study/test trials. In each experiment, subjects learned the pairs in different ways: the background color of the cue was different for each pair but was always gray for the target, the background color of the cue was always gray but was a different color for each target, the background color of each pair was different or the background color for all pairs was gray. Recall occurred in one of three conditions: The cue was shown on the same background color (control), on a different color (interchanged), or on a gray background (neutral). There was no mention of background color to the subjects in any of the experiments. Overall, recall was lower when the background color during testing differed from that in learning. In addition, recall from a cue on an interchanged (conflicting) background was worse than recall on a neutral background, suggesting that the target was associated with the background color. Errors in recall could be one of two types: incorrect based on the cue or incorrect based on the whole list. There were more errors of the first type when background color in the test was different from that used during learning. Among the two conditions in which recall background was different, there were more errors based on the cue type when the background color was interchanged than when it was neutral, also suggesting that the target was associated with the background color during learning. In addition, overall recall was worse when only the target background color changed, compared to when only the cue background color or the pair’s background color changed, further supporting the claim that the target was associated with the background color. Based on these results, Dulsky (1935) concluded that background color could serve as an effective retrieval cue and a lack of this cue could be detrimental for recall.

Others have also examined the effect of background color change in retention. For instance, Weiss and Margolius (1954) showed that, when background color of paired associates stayed the same from learning to test, recall was higher even after a 24-h retention interval. Dulsky (1935) and Weiss and Margolius (1954) both used a unique background color for each pair subjects studied. In a free recall study, Isarida and Isarida (2007) showed that a unique background color for each word was not necessary to obtain a background context effect. In fact, recall was greater when background color did not change from learning to test even when there were only two different colors during learning. However, the effect occurred only when background color alternated for each word. No context-dependency effect occurred when the researchers used only one background color or when a background color was presented for numerous successive words.

Murnane extended findings concerning background color in a series of experiments by adding text color and screen location as further variables (Murnane and Phelps, 1993, 1994, 1995). Murnane and Phelps (1993) gave subjects word pairs to learn with instructions that they should form an association between the words, hoping to prevent subjects from associating the words with the background context. Subjects studied word pairs either on one, two, three, or six different background contexts that were unique combinations of background color, text color, and screen location. Immediately after learning, subjects completed an associative recognition test formed of intact and rearranged word pairs (for the first three experiments) and an item recognition test formed of single target and distractor words (for the last two experiments). Half of the words in the recognition test were
presented with the same background context, and the remaining half were presented in a different context that was not used during learning. Overall, subjects correctly recognized more targets and falsely recognized more distractors when the recognition context was the same as learning. The number of different contexts did not reliably change the results. Informing the subjects about the background contexts before learning did affect recognition, but informing them between learning and the test did not have a reliable effect on recognition. This finding is different from other studies that showed an increase in context dependency when subjects are informed about the context; however, Murnane and Phelps (1993) argued that mentally reinstating an arbitrary learning context might be hard especially when subjects were presented with another arbitrary context on the screen.

In the previous study, Murnane and Phelps compared recognition in the same background context to recognition in a different background context that was not presented during learning. A follow-up study included recognition in a different context that was presented for a different word pair during learning (Murnane and Phelps, 1994). Subjects studied word pairs in two different background contexts (i.e., unique combinations of background color, text color, and screen location) and were given an item recognition test immediately after learning. Some items in the recognition test were in the same context, some were in a completely new context, and some were in the other context that subjects experienced during learning. Similar to their previous study, Murnane and Phelps found context dependency when recognition in the same context was compared to recognition in the completely new context; however, there was no context dependency when recognition in the same context was compared to recognition in the other context subjects experienced during learning. The authors concluded that it was necessary to use a context not experienced during learning on a recognition test to obtain context dependencies. However, this outcome differs from that reported with simpler contexts (shifting color backgrounds).

Another study by Murnane and Phelps (1995) showed that the effectiveness of a background context did not decrease as item strength increased. The argument was that if a to-be-remembered item was better encoded, the background context would not serve as an effective retrieval cue and would not be as important. In a series of experiments using their previous paradigm (Murnane and Phelps, 1993, 1994), Murnane and Phelps (1995) increased item strength via either spaced repetition, presenting word pairs longer or increasing the depth of processing of the word pairs. Overall, context dependency did not decrease as item strength increased, eliminating the argument that background context will not serve as an effective retrieval cue when to-be-remembered items are better encoded. Note that this finding is different from findings arising from the physical context–dependency literature.

Macken (2002) extended previous findings by showing that context dependency occurred for both words and nonwords during recognition, but only when recognition was accompanied by recollection, not familiarity. He compared recognition of both words and nonwords presented either in the same background context as learning on the recognition test or in a completely new background context while asking subjects about their recollective experience during recognition using Tulving’s (1985) remember–know paradigm. Macken obtained context dependency for hits (i.e., subjects correctly recognized more words and nonwords when the background context was the same than when it was new), but only when subjects reported conscious recollection of the items (i.e., with remember responses) and not when subjects based their response on knowing. In addition, Macken (2002) showed that context dependency also occurred when recognition in the same context was compared to recognition in a different but previously experienced context. This outcome differed from the findings of Murnane and Phelps (1994). Aside from background color and simple visual contexts, other studies examined richer background contexts, such as presenting items to be remembered on scenes and pictures. In one such study, Murnane et al. (1999) gave subjects word pairs embedded in four different plausible scenes (e.g., a word pair on a chalkboard in a classroom scene). Subjects rated the relatedness of the words in the word pair. During recognition, they were presented a word from each pair either in the same background context or in a completely new background context for a yes/no recognition test. Distractors were equally distributed among the four previously presented and the new contexts. Overall, context dependency was observed, where subjects correctly recognized target words (a word from each pair) more and false-alarmed to distractors (new words) more when the words were presented in the same context than when they were presented in a new context.

In a separate study, Hayes et al. (2007) showed poorer recognition of objects when they occurred in the absence of the context in which they were studied. They presented subjects pictures of objects in scenes and on a white background during learning and used the same backgrounds during the recognition test. Subjects correctly recognized more objects when they were presented in the same scene compared to when they were presented on a white background. Instructing the subjects to focus only on the objects did not reduce context dependency, so Hayes et al. (2007) concluded that subjects bind the object and context automatically.

Reder et al. (2013) compared recognition of famous and unknown faces in the same context they were studied or in a different context that had been previously experienced by subjects. They saw faces of famous and unknown people superimposed on familiar locations and rated how likely it was for each person to visit that location. There were either 3 or 12 faces associated with each location. During the recognition test, the faces from the study phase as well as distractor faces were presented either on the same background context or on one of the other background contexts from the study phase. As expected, famous faces were recognized more accurately than unknown faces, but critically the context-dependency effect occurred only for famous faces and more so when there were fewer faces associated with a background context. Subjects correctly recognized more famous faces when they were presented on the same background context during the test than when they were presented on a background context that had presented with a different face during study. Recognition of unknown faces was reliably above chance and neither context reinstatement nor the number of faces associated with a background context mattered. Reder et al. (2013) concluded that context dependency is observed for background contexts, but only when the stimuli were familiar to subjects. Note that this finding contradicts those of environmental context dependency, where the effect was obtained primarily for meaningless or unfamiliar materials. This discrepancy may be due to poor recognition of unfamiliar faces overall in the Reder et al.’s experiment.
Although a picture is a richer context than a combination of background color, text color, and screen location, it is still relatively simple compared to being underwater or being in a different room. In attempting to study context dependency within laboratory settings, Smith and Manzano (2010) came up with a clever method to help bridge the gap between physical context and background context. They presented subjects with 30 words superimposed on an unrelated video with color, movement, and sound (e.g., a restaurant scene, a softball game) and told them to remember both the words and the videos for a later test. They manipulated the target-to-context ratio, so either 15 words, 3 words, or 1 word was associated with 2, 10, or 30 contexts, respectively. Immediately after study, subjects were asked to recall all the words they studied. Critically, videos corresponding to half of the words were presented during recall, with no background context presented for the remaining words. Fig. 6 shows recall performance across each target-to-context ratio. Overall, context dependency was observed, as more words were remembered when the videos associated with them were presented during recall than when they were not. There was also an effect of target-to-context ratios, where subjects remembered more words when the target-to-context ratio was smaller (i.e., performance was best when there was only one word associated with a context). Critically, context dependency was larger when the target-to-context ratio was smaller, with the effect size increasing when only one word was associated with a context (Cohen’s $d = 3.02$).

How specific are these effects to the context presented during learning? Do similar contexts—one softball game during study, a different game during test—lead to similarly strong effects? Several experiments suggest that when the reinstated context is similar to, but not the same as, the original context, context dependency is observed, although to a lesser degree. Using videos as background contexts, Smith et al. (2014) explored the effect of context similarity. Subjects studied words superimposed on unrelated videos, and there was only one word associated with each video. Across two experiments, recall was tested by presenting some of the original videos, some different videos, or without the presentation of any videos. In Experiment 1, the different videos were from the same event but were recorded at a different time period within that event. In Experiment 2, the different videos were from a new event similar to the original one (e.g., a different restaurant). Overall, context dependency was observed, where subjects remembered more words when the same associated videos were presented during recall relative to no context during recall. Critically, there was also context dependency, though smaller, in the test conditions that used videos of different time periods from the original contexts and videos of scenes similar to the original contexts. Based on the findings of Smith et al. (2014), similar contexts can also lead to background context effects, though to a lesser degree. This outcome is similar to stimulus generalization in classical and operant conditioning: stimuli similar to the original stimulus situation elicit the response, although to a lesser extent.

In summary, the evidence is consistent: Background contexts—be it background color, simple visual contexts, pictures, or videos, aid retrieval when those present during encoding are also present during retrieval. Context dependency seems to occur regardless of instructions related to the contexts, is usually greater when there are fewer to-be-remembered items associated with a context, and can also occur when the contexts presented during retrieval are similar to (but not the same as) the original context.

### 2.02.3.6 State Dependency

Similar to context dependency with external cues, remembering can be dependent on our internal states. Most research in this area focused on the pharmacological and the mood state of the individual. Again, encoding–retrieval designs are used in these experiments. Most research converges on the conclusion that remembering is enhanced to the extent that encoding and retrieval states match; however, the literature regarding mood states is less clear. Therefore, the relations between mood and memory will be discussed in the next section, and this section will focus on other forms of state dependency.
The typical state-dependent retrieval memory experiment has subjects learn information in one of two pharmacological states (e.g., after imbibing alcohol or marijuana) and then testing them later either with the same drug in their system or no drug (sober). Results from experiments in the 1960s and 1970s were mixed, with some finding state dependency with various manipulations but others obtaining no effect. Eich (1980 p. 165) examined the literature to shed light on what he referred to as the "unpredictability of human state-dependent retrieval." He discovered that whether state-dependent retrieval occurs depends on the nature of the retrieval task. According to his review, state-dependent retrieval is a cue-dependent phenomenon, meaning that it is largely determined by the presence or absence of cues during retrieval. Most studies using free recall have obtained the effect, whereas those that used recognition or other tests with powerful retrieval cues have not produced positive results. As in other cases we have considered, recognition tests, using copies of the to-be-remembered target as a cue, are quite powerful and may overshadow the influence of subtler cues like state. The same seems true in the case of cued recall with strong cues. Pharmacological state serves as an effective retrieval cue only when no explicit cues are provided to subjects.

A study by Eich et al. (1975) provides a good example to illustrate the cue-dependent nature of state-dependent retrieval. Eich et al. used marijuana cigarettes to manipulate the pharmacological state of subjects who were Army recruits. Subjects either smoked a cigarette containing marijuana or a placebo cigarette before learning a list of categorized words presented a half an hour or so after smoking (to permit the drug to take effect). Retrieval took place 4 h later, and subjects either smoked the marijuana or the placebo cigarette before the test. All four possible conditions were used (marijuana at study, no marijuana at test; marijuana at study, marijuana at test, etc.). During retrieval, half of the subjects engaged in cued recall with the corresponding category names provided, and the other half engaged in free recall. Overall, marijuana impaired memory (either during encoding or retrieval) on both tests. However, a state-dependent retrieval effect was observed in free recall, but not in cued recall. That is, in free recall, subjects in the matching state conditions (e.g., marijuana at study–marijuana at test) recalled more words than when in the mismatched condition (e.g., marijuana at study–placebo at test). Surprisingly, the sober subjects at test actually performed worse than those given marijuana at test, but only if they had learned the information under the influence of marijuana at study. On the other hand, when subjects in these two conditions were given category names as powerful cues at test, the effect vanished. As we have seen in other cases of encoding–retrieval interactions, subtle state cues can be overwhelmed by more powerful external cues. Of course, being sober at study and at test led to the best performance overall; marijuana and other depressives drugs depress recall (free or cued).

Similar results were obtained by Petersen (1977), who used alcohol to manipulate subjects' pharmacological state. Prior to encoding and retrieval, subjects drank a beverage either with or without alcohol. He found that imbibing alcohol at retrieval aided free recall if alcohol had also been drunk before encoding (relative to the alcohol at encoding, no alcohol at test condition). The same effect was not obtained with category names as cues. Examining 57 studies, Eich (1980) reported that the type of psychoactive drug used in these studies did not determine the presence or absence of the state-dependent retrieval effect. Instead, the type of retrieval task was a better predictor of whether state dependency will be observed, with cued recall and recognition tests generally failing to produce the effect.

More recently, state dependency was extended beyond pharmacological states through examining the relation between other physiological states and memory (Miles and Hardman, 1998; Schramke and Bauer, 1997). In one study, subjects learned a list of 36 words either during aerobic exercise (i.e., pedaling on a stationary bicycle) or at rest. Five minutes after learning, subjects were asked to recall the words they studied either during aerobic exercise or at rest. Each subject was tested individually; words were presented auditorily, and recall occurred orally. A significant change in heart rate was considered a change in physiological state. Fig. 7 shows recall performance across four conditions. Overall recall did not differ for different encoding or retrieval conditions, but the crossover interaction between them revealed a state-dependent effect: Recall was better when physiological states

![Figure 7](image-url)  
**Figure 7** Results of Miles and Hardman (1998). Not enough information was reported to calculate confidence intervals. Results show mean number of words recalled across four conditions.
of study and test matched (i.e., study at rest–recall at rest and study during exercise–recall during exercise) compared to when they did not match (i.e., study at rest–recall during exercise and study during exercise–recall at rest). Miles and Hardman (1998) concluded that a change in physiological state (i.e., cardiovascular activity) could lead to state dependency.

Another study by Schramke and Bauer (1997) that examined the relationship between physiological states and memory also revealed a state-dependent effect in both younger and older adults. Subjects exercised (i.e., walked in a corridor) or rested for 5–7 min before learning a word list. After a brief delay, subjects engaged in free recall, cued recall with category cues, and a recognition test, respectively, either after 5–7 min of exercise or rest. Changes in heart rate, blood pressure, and self-reported distress were considered a change in physiological state. Critically, state dependency was observed in both younger and older adults in a free recall task, but was not observed during cued recall or recognition. Clearly, aside from pharmacological states, physiological states can serve as an alternative to study encoding and retrieval interactions.

Another interesting study was conducted by Garczynski and Brown (2013) on the relationship between active self-aspects or self-concepts and memory. By self-aspects, the authors mean the different roles and hence somewhat different identities that people have in different situations (e.g., the way one behaves at work vs. at home). The authors were interested in whether recall would be greater when active self-aspects during encoding and retrieval matched compared to when they did not. They noted that how one defines “the self” is context specific, and thus different self-aspects could affect how individuals think, feel, and behave by changing the way individuals perceive and store stimuli around them. In this sense, self-aspects can be considered as mental states. Garczynski and Brown used two self-aspects in their student subjects: home and school self-aspect. To activate them, half of the subjects wrote about how they think, feel, and behave at home for 5 min, and the other half did the same for school. Subjects were then presented with 40 words and made a semantic judgment (i.e., whether the word is a synonym of another word) for half of the words and a self-referential judgment (i.e., whether the word describes them) for the remaining half of the words. Following a 5-min distractor task, half of the subjects wrote about their thoughts, feelings, and behaviors at home and the other half did the same for school. After activating a matching or a nonmatching self-aspect, subjects recalled as many words as they could. The results showed state dependency, where subjects recalled more words when the active self-aspect during encoding and retrieval matched compared to when they did not. State dependency occurred regardless of the type of judgment made on the words, suggesting that the effect is not only driven by self-relevance. Of course, it may be that activating different self-concepts by writing about thoughts, feelings, and behaviors at a certain place could be context reinstatement or mood dependency instead of a separate phenomenon. To address this argument, the authors conducted a content analysis on what the subjects wrote before encoding and retrieval and controlled for these in a separate analysis. The interaction between self-aspects during encoding and retrieval was still significant, suggesting that activation of self-concepts is a different form of state dependency.

Many studies on state-dependent effects of drugs on memory have used depressive drugs such as alcohol and marijuana. Kelemen and Creeley (2001, 2003) extended such research to examine the effects of caffeine on memory, as well as to people's metamemory judgments (i.e., do people predict that taking caffeine will improve retention?). Thus, they examined how the presence or absence of caffeine during encoding and/or retrieval affected retention, but they were also interested in how it affected the predictive accuracy of metamemory judgments. Kelemen and Creeley (2001) examined state dependency in multiple memory tests (free recall, cued recall, and recognition) and metamemory judgments (judgments of learning). They found a reliable state dependency only in free recall and found a trend for the other memory tests. There was no state dependency in the predictive accuracy of metamemory judgments. The authors conducted another study where they used only one memory test (cued recall) to reduce the interference between tasks. In this study, Kelemen and Creeley (2003) had subjects either drink a sweetened beverage with caffeine or a placebo beverage before studying 40 word pairs. Subjects were informed about a later memory test and were asked to make a prediction for each item about its recall a day later (i.e., judgment of learnings). All subjects came back a day later and either drank a sweetened beverage with caffeine or a placebo beverage before retrieval. All four possible conditions were examined (caffeine at study, caffeine at test, etc.). During both encoding and retrieval, the subjects' self-reported level of arousal was used as a measure of change in state. Because the task was paired-associate learning, subjects were given the left-hand member of the pair as a cue and their task was to type in the second word. A state-dependent effect was observed, because subjects in the same state at encoding and retrieval recalled more compared to those who were in different states. Note that this is among the few studies observing state dependency when the retrieval task is cued recall. Interestingly, caffeine did not lead to better retention overall; the caffeine–caffeine condition did not lead to greater recall than the placebo–placebo condition.

Kelemen and Creeley (2003) also examined the predictive accuracy of judgments of learning, as measured by how accurately a subject's judgment of learning on an item matched his/her recall of that item on the subsequent memory test. Surprisingly, the state-dependent effect did not extend to metamemory judgments: The predictive accuracy of judgments of learning was not higher for subjects in the same state, but it was higher for subjects in different states during encoding and retrieval. Obviously, this outcome deserves further examination. Kelemen and Creeley's study is important in showing state dependency with caffeine, a stimulant, especially with a cued recall test, but it is also important because it examined encoding and retrieval interactions using metamemory judgments.

Research on state dependency, like that on context dependency, reveals numerous encoding–retrieval interactions. At least on free recall tests, subjects in the same state during encoding and retrieval often perform better than those in alternate states during study and test.
2.02.3.7 Mood Dependency

How far does state dependency extend with regard to internal states? We considered studies of physiological arousal and self-concept in the previous section. This section considers mood states: Do people remember better if they both learn and retrieve something in a happy mood relative to being elated at learning and then in a neutral or depressed mood at test? And if people are depressed when they learn—a state known to harm memory performance—will they be able to retrieve more information in the same unhappy state at test relative to neutral or happy conditions? In general, will a mood match between encoding and retrieval lead to better retention than a mismatch?

Mood-dependent memory is defined as enhanced recollection of events when the mood at the original learning is reinstated (Bower, 1981; Kihlstrom et al., 2000). Unfortunately, the evidence regarding mood-dependent retrieval is even more mixed than evidence in other sections. Failures to replicate abound. Using standard simple word list paradigms, Bower et al. (1978) conducted two experiments and failed to find mood dependency. In Experiment 1, they induced happy or sad mood to subjects through a hypnosis manipulation and had subjects learn 16 unrelated words. They were tested with free recall after 10 min either under the same mood (i.e., match condition) or different mood (i.e., mismatch condition); however, no difference was observed between the match and mismatch conditions. Experiment 2 used a similar design and failed to find an effect with a retention interval of 24 h. In Experiment 3, they used an interference paradigm. As in the previous experiments, subjects studied a list of 16 unrelated words (i.e., List A) under happy or sad mood conditions. After that, they studied a second list of words (i.e., List B) under matching or mismatching mood conditions. Finally, during the test, they were asked to recall words from List A, but not B, again either under happy or sad mood induction. The results showed that subjects who learned lists in mismatching moods (e.g., List A in a happy mood and List B in a sad mood) and then retrieved List A in a matching mood to the original study phase (i.e., a happy mood) performed better than subjects who learned lists in mismatching moods but also retrieved List A in a mismatching mood. Thus the results showed a pattern of mood-dependent retrieval, but only when interference occurred between study and test. Bower et al. concluded that in single-list learning, random words were distinct enough that a mood match was not a strong enough cue to boost retrieval; however, with multitest learning and hence interference, the mood cue helped to make the list more distinctive and aid retrieval.

One important feature of Bower et al.’s (1978) experiments is that they only used subjects who were highly susceptible to hypnosis. Others were able to replicate this pattern using unselected students and manipulating mood a different way (Schare et al., 1984), yet some other results were not so positive. To try to help settle the case, Bower and Mayer (1989) conducted a series of six experiments with multiple lists using interference paradigms. However, five of the experiments failed to find any hint of a mood-dependency effect, and they concluded that their earlier finding (from 1978) was unreliable and spurious.

Other researchers took up the challenge. One challenge in the mood-dependency literature is how to determine that a mood has really been changed. Researchers must rely on self-report rather than have a physiologically objective indicator as in state-dependent retrieval. Eich and Metcalfe (1989) conducted an interesting study using music to induce happy or sad moods. Subjects listened to uplifting or depressing classical music (normed ahead of the experiment to reliably induce the moods) and had subjects rate their mood every few minutes. Once subjects met the researcher’s criterion of being happy or sad on the self-report scale, the learning phase began in which they were to learn 32 words. Of 32 words, 16 were read in a semantic context and 16 were generated from the context with the first letter as a clue. In the reading condition, words were presented in pairs with a category name (e.g., natural earth formations: river–valley) and in the generation condition, the second word had to be generated by the subject (e.g., natural earth formations: river–v______). The reason for this manipulation was to see if internally generated words might be more susceptible to mood congruency than externally presented words.

The test occurred 2 days later under either matching or mismatching mood conditions, with the moods induced the same way. Subjects were first tested by free recall and then by recognition. Fig. 8 shows the results. Of course, the results revealed a main effect of item type, where generated words were better recalled and recognized than read words regardless of mood (the standard generation effect) (Slamecka and Graf, 1978). The interesting results concerned mood-dependent memory: For read words, mood dependency was not observed, though recall in the happy–happy condition was slightly higher compared to that in the happy–sad and sad–happy conditions. For generated words, however, subjects in the match conditions (happy–happy or sad–sad) recalled more words than subjects in the mismatch conditions, revealing mood dependency. However, this pattern did not occur for the recognition test, where memory for neither type of item showed mood dependency. Two other research groups also confirmed a mood-dependency effect for generated items (Balch et al., 1999; Beck and McBee, 1995). Note that these experiments did not use interference designs.

Eich (1995) provided some ideas as to the necessary circumstances for mood-dependent retrieval to be observed. Not surprisingly, the first was the nature of the to-be-learned material (internally versus externally generated, as just reviewed). The assumption is that internally generated events are more associated to the current mood relative to external events. In addition, Eich pointed out that once again the type of retrieval task also matters. In cases where the test provided strong cues like a recognition test, the mood match (a relatively subtle cue) does not add to the powerful external cue. These interpretations are in line with the encoding specificity principle as well as the prior body of research on context- and state-dependent memory effects.

The idea that internal events produce the mood-dependency effect helped direct the literature toward real life material instead of word lists. Eich et al. (1994) conducted three experiments using subjects’ autobiographical memories, which of course they generated. During the encoding session, subjects listened to music to induce either happy or sad moods and after reaching the required level of mood, they were presented with 16 neutral unrelated words one at a time and were asked to generate an
autobiographical memory for each word. For half of the words, they were asked to generate positive autobiographical events from their lives and for the remaining half, negative autobiographical events. They were asked to provide as many details as possible with the event, so as to produce a rich encoding. Testing occurred after 2 or 3 days and just before testing, subjects were induced either happy or sad moods again through music. Subjects then tried to recall as many autobiographical events as possible from the earlier session (without cues). The results revealed that subjects in the happy–happy and sad–sad conditions recalled more of the autobiographical events than subjects in the happy–sad and sad–happy conditions. This outcome confirmed the previous finding with word lists as materials.

Going one step further, Eich et al. (1997) conducted a study with subjects who suffered from a rapid-cycle bipolar disorder in which they would slip rapidly between mania (great elation) and profound depression. Thus the researchers did not induce the mood state but rather assessed it and superimposed the experimental manipulations onto the naturally occurring mood swings. Following the procedure of Eich et al. (1994), subjects’ moods were assessed and then they were presented with 10 neutral words and were asked to generate an autobiographical memory for each word, with details. During retrieval, subjects’ mood was assessed again (this determined their match or mismatch condition), and they engaged in the autobiographical event recall task. The results confirmed mood dependency for people who frequently experience mood shifts, because retrieval was better when the moods matched than when they mismatched.

The results of Eich and his colleagues stand out as a possible avenue forward in the study of mood-congruent memory, but we have not reviewed other rather dispiriting research here. The topic remains controversial. There is a parallel line of research asking whether events with similar emotional valence are better remembered, but this is outside the purview of our chapter (see Kihlstrom et al., 2000).

2.02.3.8 Physical Operations

In an earlier section, we discussed the match or mismatch between mental operations and their effect on retention (e.g., research on transfer-appropriate processing). Similar experiments have been conducted in which physical operations are manipulated at study and test, with interesting results. Engelkamp et al. (1994) asked subjects either to read phrases verbally (e.g., close the book) or to read them and actually perform them (closing a book in front of them). For the recognition test, subjects were asked to judge whether the phrase was the one they read in the earlier phase. Again, they either just read the phrase at the test or read and performed it. (Of course, there were lure phrases too, never studied or performed during learning). The authors confirmed previous findings that performing actions increases retention of phrases relative to just reading them (the enactment effect; see Roediger and Zaromb, 2010 for a review). For current purposes, however, the interesting finding is that when subjects performed the actions at both study and test, they recognized them better than if they only performed them at study and read them at test. Performing the action both during encoding and retrieval enhanced recognition that the phrase had been read (and performed) previously (Fig. 9). Notice that self-performed tasks during retrieval serve as excellent cues; they even help when sentences have been verbally encoded.

Dijkstra et al. (2007) found that having subjects adopt body postures congruent with the autobiographical memories they were trying to retrieve led to better recollection of memories, again expanding the encoding specificity principle to memories from one’s life. The experiment had two sessions. First, subjects were asked to retrieve six autobiographical memories with prompts (e.g., going to dentist office or playing sports). While they were retrieving the memories, they were instructed to either adopt a congruent body posture (e.g., lying down in a recliner for going to dentist office) or an incongruent body posture (e.g., standing up with the hands on the hips for going to dentist office). Then approximately 2 weeks later, subjects were asked to recall (first free recall, then recall
with prompts) the memories they retrieved at the first session. The results revealed that when the body postures matched the cue, subjects later remembered experiences in richer detail than when the cues mismatched the body postures.

Even more surprisingly, metaphoric moves during retrieval that have positive or negative connotations seem to also affect the memory performance. In one experiment, Casasanto and Dijkstra (2010) asked subjects to retrieve 24 autobiographical memories to neutral prompts (e.g., Tell me about an event that happened yesterday). The experiment consisted of retrieving and retelling phases. During the retrieval phase, while subjects were asked to retrieve personal memories silently, they were also asked to move a series of marbles either up or down. The idea is that up indicates a positive movement and down indicates a negative experience in almost all cultures (Lakoff and Johnson, 1980). During the retelling phase, subjects saw the prompts again and retrieved them aloud. After these phases, they were asked to indicate the valence of their memories. The results revealed that when subjects were moving the marbles up during retrieval, they retold more positive memories whereas when they were moving the marbles down, they retold more negative memories given the neutral prompts. This outcome would seem to indicate that even the metaphoric match in valence during retrieval can affect memory retrieval. In general, all these studies provide results that are broadly consistent with the encoding specificity principle in that matching conditions (congruent ones) produce better performance than those that mismatch.

2.02.3.9 A Final Case of Encoding–Retrieval Interactions

Nearly every experiment summarized in this chapter has employed a 2 × 2 design: A manipulation of two conditions during study or encoding is crossed with two retrieval conditions, usually with an attempt at matching or mismatching the encoding conditions (see Fig. 1). We do not mean to leave the impression that more complicated designs cannot and have not been used.

The most ambitious attempt to understand encoding and retrieval interactions is provided in a study by Challis et al. (1996) that examined the effects of standard levels-of-processing manipulations using a large number and range of memory tests, both explicit and implicit. Overall, the study crossed five encoding conditions with 10 retrieval assessments. Here we only focus on their first experiment, involving a 5 (encoding conditions) × 6 (retrieval conditions) design.

Subjects studied different sets of words studied under five (within subjects) conditions: (1) intentional learning with no specific instructions on how to remember (intentional learning); (2) counting and reporting the number of letters while they studied the words (letter counting); (3) counting the number of syllables in the word while studying them (syllable counting); (4) making judgments about whether the word named a living or a nonliving thing (living/nonliving judgment); and (5) judging how relevant the words were to themselves (self-reference judgment). Tasks 2 and 3 are typical “shallow” levels-of-processing tasks, whereas 1, 4, and 5 are usually considered deep tasks (and often foster better performance). Later, different groups of subjects were given one of six memory tests, four explicit tests and two implicit tests. The explicit tests were free recall, graphemic cued recall (cha___ as a cue for chair), semantic cued recall (table as a cue for chair); or recognition (chair as a cue for chair). In the implicit tests, subjects were told simply to respond with the first word that came to mind, and performance was measured as priming from previously studied words above a nonstudied baseline. The two tests were word stem completion (cha__) and general knowledge questions (e.g., an object sometimes thrown through a window in western movies—__).

Findings from prior research (and the levels of processing and transfer-appropriate processing frameworks) predict that semantic orienting tasks should lead to better performance than shallow orienting tasks on all the explicit memory tests. The results in Fig. 10 show that this is indeed the case (the effect of orienting tasks is measured in terms of least significant difference units in the figure, to put the measures on a common scale): Recognition, free recall, semantic cued recall, and graphemic cued recall tests all revealed benefits from the deep orienting tasks relative to the shallow tasks. The same general pattern occurred for the implicit test of answering general knowledge questions. However, the implicit test with degraded perceptual cues showed equivalent priming from all orienting tasks, a finding that also replicates other studies.

Figure 9 Results of Engelkamp et al. (1994). Confidence intervals could not be calculated because Engelkamp et al. (1994) did not report the necessary information.
In general, the results in Fig. 10 show a complex pattern of encoding–retrieval interactions, many of which are in line with transfer-appropriate processing ideas outlined above and spelled out more thoroughly elsewhere (e.g., Roediger and McDermott, 1993). But not all results can be readily accommodated by any theory, because the various semantic orienting tasks themselves produce interactions on the three explicit tests assumed to require and rely on semantic processing (recognition, free recall, and semantic cued recall). To convince yourself of this point, consider what encoding condition produces the best performance on each of these three tests: For recognition and semantic cued recall, the answer is self-reference encoding, but for free recall it is intentional learning. Why should these three differ? Challis et al. (1996) do a thorough job of discussing various puzzles raised by their complex series of interactions, but they are hard to explain. It may be for this reason that few experiments like this have been performed. As we have seen in our review, interpreting even $2 \times 2$ interactions (or lack thereof) can be complicated enough. Our point is that such encoding–retrieval designs should be the beginning of an analysis and not its end point. We need more complex experiments like those of Challis et al. if we are to comprehend encoding–retrieval interactions more fully.

![Figure 10](image)

**Figure 10** Results of Challis et al. (1996). Results are reported in units of least significant differences (LSDs) to show the effect of each learning condition on a common scale over a nonstudied baseline. If a bar is missing from a test, performance was not one LSD above zero or baseline.

2.02.4 Conclusion

This chapter is long, but at this point our field has really only skinned the surface of the vexing issue of encoding–retrieval interactions. The general point of such interactions is that explicit tests of memory vary in many ways and reveal different effects of prior experience, and adding implicit tests to the mix makes the challenge even greater. If we had included perceptual and conceptual implicit memory tests, we would have shown many more interactions between priming on perceptual implicit tests and most explicit tests, as in the Challis et al.'s (1996) research (see Roediger and McDermott, 1993 for a partial review). In addition, dissociations between different types of implicit tests are also complex (see Roediger et al., 1989). Of course, many variables can also have parallel effects on tests (e.g., the levels-of-processing effect is routinely obtained in recall and recognition tests). We have generally chosen to emphasize differences across tests in this chapter, ones that (broadly) support the principles of encoding specificity and transfer-appropriate processing.

Encoding–retrieval interactions are important, but readers should also be mindful that these sorts of experiments are embedded within other possible factors that can modify effects in memory experiments. That is, the types of subjects used (young, old, children, depressed individuals, bilinguals, etc.) and types of materials can matter. Jenkins (1979) discusses how these factors, as well as encoding and retrieval factors, may create interactions. And to this list we need to add the type of experimental design: Often effects obtained in within-subjects designs, even large effects, are not obtained in between-subjects designs (see McDaniel and Bugg, 2008) even on the same test.

To return to our theme from the beginning of the chapter, most variables that have been claimed to affect “memory” on some particular task often have no effect or even a reversed effect depending on whether other conditions are held constant or varied. The science of memory is a field in which “laws” transcending the situation of learning and the type of test are vanishingly small, if they exist at all (Roediger, 2008). This is not meant to end on a depressing note, but simply to feature the complex truth of the situation, one we are barely beginning to understand.

What theoretical lesson can we take from this mass of findings? As we have discussed, the great bulk of the findings can be interpreted within the context of encoding specificity and transfer-appropriate processing principles. However, there is one
important caveat: Simply matching a contextual feature at study with another feature similar to this during test may not improve performance because the feature may not have been encoded (at study or at test). Many years ago, Underwood (1963) distinguished between a nominal stimulus (what is presented to the learner) and the functional stimulus. A learner might be asked to remember a seemingly long meaningless of letters (BYOBCLAIBFBIWAOLTMAFAQ – the nominal stimulus) and yet chunk them via recoding to remember the string with ease (BYOB CIA IBM FBI AWOL ATM FAQ – the functional stimulus) because it is composed of seven chunks. Similarly, the experimenter might present a context that is intended to influence the learner, but it is not noticed or not encoded. To obtain context effects, the learner must, at a minimum, encode the context both during learning and when retrieval is required. The context must be part of the encoded bundle of features (the engram or memory trace) to be useful at test. During the test, the context must become active (it can even be activated via imagination) to have an effect. Conditions that tend to exclude context from being encoded as part of the trace are those in which other potential retrieval cues are prominent (e.g., the first word in paired-associate learning, which later becomes the primary cue rather than some background context). In addition, if the context itself is not especially distinctive or prominent, then it might not be encoded. The Godden and Baddeley (1975) experiment may have produced such striking results because the contexts—learning on land or under water—are highly distinctive. And even when the context is part of the encoding, a rival retrieval cue may be much more effective than the context cue; recall that context- and state-dependent effects are often found in free recall but usually not in recognition with its more powerful cues. A final proviso is that the more events that are subsumed under a context, the less effective the context cue will be in provoking recall of any one event, the cue overload principle (Tulving and Pearlstone, 1966; Watkins and Watkins, 1975; see too Nairne, 2002). The body of evidence in this chapter supports all these claims, with an exception here and there.

References


