Hunt, R. R., & Worthen, J. B (Eds.). (2006). *Distinctiveness and memory*. New York: Oxford University Press.
MacKay, D. G., & Ahmetzanov, M. V. (2005). Emotion, memory, and attention in the taboo Stroop paradigm. *Psychological Science, 16*, 25–32.
MacKay, D. G., Shafto, M., Taylor, J. K., Marian, D. E., Abrams, L., & Dyer, J. (2004). Relations between emotion, memory, and attention: Evidence from taboo Stroop, lexical decision, and immediate memory tasks. *Memory and Cognition, 32*, 474–488.
McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Reviews of Neuroscience, 27*, 820–840.
McIntyre, J. S., & Craik, F. I. M. (1987). Age differences in memory for item and source information. *Canadian Journal of Psychology, 41*, 175–192.
Most, S. B., Chun, M. M., Widders, D. M. & Zald, D. H. (2005). Attentional rubberneck-ing: Cognitive control and personality in emotion-induced blindness. *Psychonomic Bulletin and Review, 12*, 654–661.
Naveh-Benjamin, M. (2000). Adult-age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1170–1187.
Naveh-Benjamin, M. (2002). The effects of divided attention on encoding processes: Underlying mechanisms. In M. Naveh-Benjamin, M. Moscovitch, & H. L. Roediger (Eds.), *Perspectives on human memory and cognitive aging* (pp. 193–207). Philadelphia: Psychology Press.
Reisberg, D., & Heuer, F. (2004). Remembering emotional events. In D. Reisberg & P. Hertel (Eds.), *Memory and emotion* (pp. 3–41). New York: Oxford University Press.
Spencer, W. D., & Raz, N. (1995). Differential effects of aging on memory for content and context: A meta-analysis. *Psychology and Aging, 10*, 527–539.
Talmi, D., & Moscovitch, M. (2004). Can semantic relatedness explain the enhancement of memory for emotional words? *Memory and Cognition, 32*, 742–751.
Winograd, E., & Neisser, U. (Eds.). (1992). *Affect and accuracy in recall*. Cambridge, UK: Cambridge University Press.

11

Putting Context in Context

KRYSTAL A. KLEIN, RICHARD M. SHIFFRIN,
and AMY H. CRISS

Roediger and McDermott's comprehensive review of the implicit memory literature (1993) included a puzzling pair of findings: If a prime item is studied in massed fashion (i.e., longer study or successive presentations), the prime benefit on a later perceptual-implicit memory test (e.g., lexical decision) does not increase beyond that provided by a brief or single presentation. However, if the prime item is repeated in spaced fashion, the prime benefit increases with the number of presentations (see Jacoby & Dallas, 1981; Roediger & Challis, 1992). These findings are puzzling because explicit memory tests show that performance increases with the number of both massed and spaced presentations (albeit more so for spaced presentations). Starting with Shiffrin and Steyvers (1997) we had been developing a theory that included a key role for context to account for implicit and explicit memory and the relation between these. Long-term priming, for example, was explained in large part by the assumption that event-study produced not only an explicit trace (incomplete and noisy), but also additional context storage in that event's knowledge trace (if one existed; see Schooler, Shiffrin, & Raaijmakers, 2001). However, the findings highlighted by Roediger and McDermott did not fit that developing model, and led us to look deeper into the role of context and the mechanisms by which it affected memory. Now a decade and more after Roediger and McDermott put implicit memory in context, we believe it fitting to report our subsequent attempts to place context in context.

WHAT IS CONTEXT?

The idea of "context" is at once the bane and boon of those aiming to understand memory. Although context has been employed almost universally as an explanatory construct, in such areas as classical conditioning (Gantt, 1940), motor learning (Wright & Shea, 1991), recall (Anderson & Bower, 1972; Dulsky, 1935; Raaijmakers & Shiffrin, 1981; Strand, 1970), recognition (Criss & Shiffrin, 2004a; Dennis

& Humphreys, 2001), and directed forgetting (Sahakyan & Kelley, 2002), systematic manipulations of context have shown only moderate and occasional effects (Fernandez & Glenberg, 1985; Smith, 1988, 2001). Today's situation is not so distant from Underwood's (1977) comment: "never in the history of choice of theoretical mechanism has one been chosen that has so little support in direct evidence" (p. 43). Evidentiary support aside, the importance of context in memory theory has continued to grow. The situation is in some ways akin to that of "quarks" in physics: Direct evidence is difficult to come by, but the construct is necessary to build a coherent theory.

Context is a necessary theoretical construct in good part because no event occurs in isolation from the observer's internal states, or in isolation from the environmental surround. In general, context is the term used to describe the joint contribution of all of these factors to the mental state of a person at a given time. Some of the more important of these factors are: (1) information manipulated by the investigator (other than the nominal target of memory testing itself), such as adjacent words in a word list, the computer monitor background, and the room setting in which study and testing might occur; (2) external environmental information that is in principle identifiable and measurable, but not manipulated (such as the temperature of the laboratory, the font of orthographic study materials, the instructions, the illumination conditions); (3) external environmental information that is not manipulated and difficult to identify and measure (such as the changing external verbal and ambient auditory noise, air currents, transitory vibrations, movements of the participant in her or his seat during testing); (4) internal bodily conditions, mental and otherwise, that are measurable and identifiable (e.g., verbalizable strategies, body temperature, heart rate); (5) internal information not readily identifiable and measurable (such as various cognitive states and strategies, transitory perceptions and thoughts, evanescent bodily changes like itches and stomach upsets).

Many theories assume that some of this information is stored in memory with events, and also used to probe memory during retrieval. Researchers: (1) sometimes assume context is unattended, though this does not have to be the case; (2) typically assume context refers to factors not manipulated by the experimenter, though some theorists include nearby events as part of the context for a given event (e.g., nearby words in a study list—see Howard & Kahana, 2002); (3) often define context to exclude content information directly tied to the study event (such as the meaning of a word). It seems clear to us that real progress in understanding context and its role in memory will require both a richer empirical database, and further modeling and testing particular assumptions. In this chapter we lay out some first steps along this path, starting with a brief historical perspective and then a brief overview of context as specified in our "Retrieving Effectively from Memory" model (REM; Criss & Shiffrin, 2004a; Shiffrin & Steyvers, 1997). We then attempt to flesh out an evolving picture of context and its effects by presenting results from a number of studies, some published and others previously withheld, interpreting the findings within the REM approach.

MODELING CONTEXT

Precursors to REM

In REM's precursor, the Search of Associative Memory model (SAM; Raaijmakers & Shiffrin, 1981), context information was described in general terms; however, in its application to episodic recall, it was essentially information identifying an item (typically a word) as having occurred on the current study list (analogous to list markers in Anderson & Bower, 1972). During retrieval, a context cue derived from the current context, or possibly a reconstruction of some past context, could be used to focus memory search upon episodic traces of the words in the current list.

The SAM model was very successful in predicting a variety of memory effects, but had several theoretical drawbacks. For one, the occasional erroneous recalls of items from previous lists indicated that search could not be perfectly limited to traces from a given list, as SAM assumed. Another simplification, that a single context might apply to an entire list, seemed impractical; this assumption was elaborated and revised in Mensink and Raaijmakers' (1988) adaptation of SAM, which incorporated a drift of active contextual features over time. However, the assumption of one context for an entire list may be close to the truth; we shall provide relevant evidence in the form of studies by Klein, Criss, and Shiffrin (2004, 2006). Finally, the earlier theories did not address the role of context in the interaction between episodic traces and knowledge traces during storage and retrieval. Rectifying these omissions was one of the forces underlying the development of the REM model.

REM in a Nutshell

REM attempts to retain the basic conceptual content that has allowed SAM to deal so effectively with episodic recall and recognition, but makes three important extensions and changes: (1) It assumes a featural representation for event traces and knowledge traces and traces (in the form of a vector of feature values), allowing similarity of traces to be assessed; (2) it assumes that retrieval is based on a conditionally optimal Bayesian decision process, conditioned on incomplete and error prone event storage (see Anderson, 1990); (3) it assumes that study of an item both produces an episodic memory trace (including information from the knowledge trace[1] of the studied item, if one exists) and adds information to the item's knowledge trace or a previous episodic trace (if either exists). Assumption 3 allows one to see how knowledge can grow from repeated event-episodes during development.

The features themselves are represented in a very simple fashion: A feature type lies in a fixed, specified position of the vector (allowing probe to trace comparisons, cf. Criss & Shiffrin, 2004c, 2005). A given trace has just one value for a given feature type; this value is either an integer indicating the content of the feature (e.g., color feature might have different color values, or a size feature might have different size values), or a zero value to represent features that have not been encoded (see Shiffrin & Steyvers, 1997, for a thorough description of memory trace representation). Thus an item, memory trace, or probe is represented as a vector of integers and zeros.

Because knowledge traces of words are formed from an accumulation of episodic traces, each with differing context features, the knowledge trace is associated with too much context rather than too little: More precisely, a knowledge trace contains a generic context, rather than any one identifiable context, and is in this sense decontextualized. In contrast, an episodic trace stores just one context, albeit incompletely and with error.

Memory retrieval begins with a vector of probe features; in a recognition experiment, this would contain the features of the probe to which the participant must respond "old" or "new." In principle this probe vector consists of both content and context features, and is compared in parallel to all traces in memory. Activation of each trace is based on feature similarity, calculated as a likelihood ratio. Traces with both enough information and a large enough likelihood ratio are above activation threshold and take part in subsequent aspects of retrieval. Single-item episodic recognition is based on the average likelihood ratio, usually termed "familiarity." Recall is based on a cycle of sampling and recovery operations: Sampling of a trace is proportional to the trace activation in comparison to all above-threshold traces. Recovery and output then depend on what can be retrieved from that trace, which depends on the amount of information stored. Each cycle of cued recall uses the item and context cue. In free recall, some cycles use as a probe the content related to a recently retrieved item plus context, while others use as a probe context only.

In practice, for a variety of list memory studies in which new items occur on every list, this general approach is typically simplified by separating out the role of context. In particular, it has been assumed that the main role of context is to restrict activated traces to those from the recent list, so that context causes all of these traces and no others to be in the activated set. Then the trace activations within that set of list traces are calculated only on the basis of the matching of the content features. In effect, this approach is justified on the assumption that list context is constant for the list, therefore playing no differential role. Shiffrin and Steyvers (1997) investigated the implications of using the calculations appropriate for the simplified case in more general settings. In many such cases the qualitative pattern of predictions did not change.

Item versus Context Information

Episodic recognition in REM is usually based on familiarity (ignoring applications requiring the supplementary use of recall). The simplified version of REM (Criss & Shiffrin, 2004a) calculates familiarity based on the degree to which the probe and the memory traces share matching features but do not have mismatching feature values. Note that the digit codes in REM convey base rate information, but not magnitude information. Although obviously a simplification, similarity does not depend on the magnitude difference between mismatching integers. A target probe tends to match its own memory trace better than any other and thus the average familiarity for a target is greater than for a foil. However, in both cases, one (or more) of the nontarget memory traces can match well enough to trigger a false recognition, and this occurs more frequently as list length rises. This approach lets performance be determined by confusions with other traces, within the set of traces specified by the context cue. Dennis and Humphreys (2001) proposed an alternative in which the test-word content features activate all episodic traces of the test word in long-term memory (not just traces from the current list), and that traces of other words are not activated at all.[2] A positive recognition decision is made when the context information in these traces is sufficiently similar to the test context. Recognition is facilitated by the fact that old test words tend to have been stored with more recent context than new test words. (In their view, recall tasks, and recognition for items other than words, are handled differently.) Although their model also uses both word content and context information to recognize, there are no confusions with traces of other words.

Criss and Shiffrin (2004a) used new data (see below) and modeling to defend the view that similarity of traces to both context and content features determine episodic recognition generally and word recognition in particular. Their version of the general REM model allows differential weighting of item features and context features, making it possible to assess the relative contributions of item noise and context noise to performance. For one paradigm that was designed to build in many context confusions (test items could have been presented in any combination of the three most recent lists, or in none), they showed that the effect of context similarity did outweigh the effect of content similarity, though both played a role. In order to fit the REM model to this task, they had to incorporate a (simplified) model of context change across lists. This REM approach, with probes using both context and content features, and context changes between but not within lists, will be the starting point for theoretical explanations in this chapter.

EXPERIMENTAL STUDIES

Over many years, a number of unpublished and published studies by our group have provided evidence concerning the role of context in memory encoding and retrieval. Although few of these studies manipulated context directly, they allowed indirect inferences to be drawn. We will describe several of these, and also present data from new experiments using comparative recency judgments that are aimed to allow more direct inferences about context.

The "One-Shot of Context" Hypothesis

Roediger and McDermott's review (1993) led us to a hypothesis to explain why (1) additional distributed study increases priming, although (2) massed study does not increase priming (although it does improve explicit memory). According to REM, study leads to the addition of "new" current context features to a knowledge trace; when the test context is similar to the study context, the increased context matching will increase speed and accuracy of access to that knowledge trace. Why should this process fail following massed study? We guessed that extra massed study might produce extra storage for content information, but not for context information. However, we guessed further that extra spaced study of a repeated item would cause additional accumulation in one trace of both context and content information. This was termed the "one-shot of context" assumption (Malmberg & Shiffrin, 2005).

Malmberg and Shiffrin (2005) carried out an independent test of this hypothesis by exploring the list-strength effect (LSE) in free recall. Strengthening some list items had been shown to harm free recall of other list items (Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990), a result termed a positive LSE (in contrast with a null or negative LSE, which is obtained in recognition and cued recall). Malmberg and Shiffrin noted that the addition of the "one-shot of context" hypothesis to the REM (or SAM) theory resulted in the following prediction: A positive free-recall LSE should occur only for spaced but not massed study of the strengthened items. Because only spaced study had been used in previous free-recall LSE experiments, Malmberg and Shiffrin conducted a study using both spaced and massed study conditions to test the "one-shot" theory; Figure 11.1 shows the data, which exhibit the predicted interaction.

The basis for the predicted interaction is not only the "one-shot" assumption, but also the REM/SAM assumption that probe cues in free recall are of two types during the course of retrieval: Sometimes the probe uses both content and context

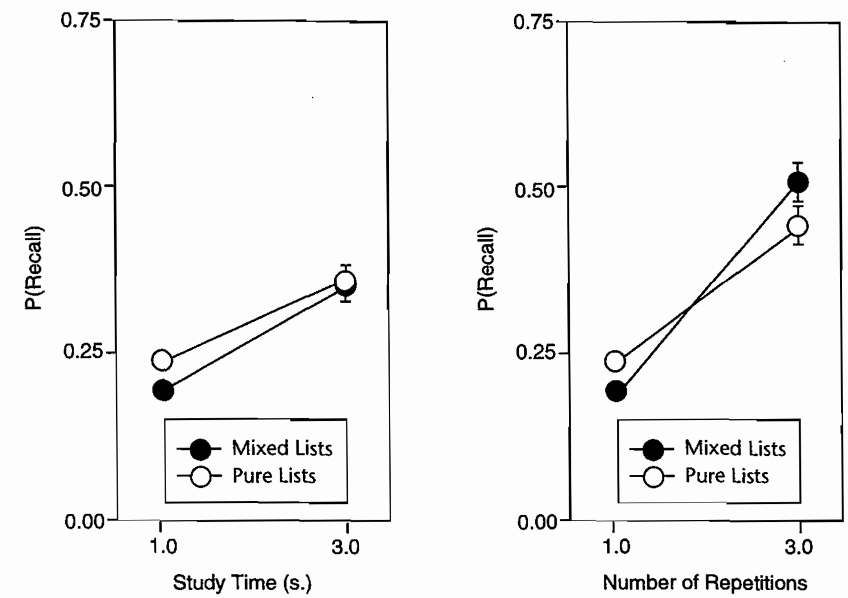

FIGURE 11.1 Results from Malmberg and Shiffrin (2005): Probability of recall for mixed and pure lists. The left panel plots a null list-strength effect observed for lists using massed (study-time) strengthening; the right panel plots a positive list-strength effect obtained for lists using spaced (repetition) strengthening. Error bars reflect standard error of the mean. Figure reproduced from Malmberg and Shiffrin (2005).

features, in which case a null or slightly negative LSE is predicted (the situation is like that holding in cued recall), and other times the probe uses context features only, in which case the list-strength effect will be positive if context storage had increased for strengthened items (as occurs for spaced study), and will be null if context storage had not increased for strengthened items (as occurs for massed study). Thus, the positive LSE is predicted only for spaced repetitions of strengthened items. Malmberg and Shiffrin (2005) therefore obtained converging evidence: Both the priming results and the free-recall LSE findings can be explained in the REM/SAM framework by the same "one-shot of context" assumption. Why should such a hypothesis hold? It may be that some storage of both context and content occurs automatically in the first second or two following event onset, but that additional storage depends on coding and rehearsal processes that are strategically allocated to content information. Whether it is possible to induce reallocation to context, thereby changing the pattern of findings, is presently unknown.

In order to account for the null LSEs found in recognition and cued recall, both SAM and REM incorporated the principle of differentiation—that strengthening an item produces a memory trace that is more dissimilar to a random alternative item. This assumption in turn requires that repetitions of an item within a given experimental list tend to produce accumulation of information in a single episodic memory trace (contrasting with the alternative assumption that repetitions of a given item produce separate list traces for each—see Ratcliff et al., 1990; Shiffrin, Ratcliff, & Clark, 1990). Accumulation of item information in a single trace is easiest to justify, especially for spaced repetitions, if context does not change much during a single list. In this case, each repetition will be as similar as possible to the others. We provide evidence below that supports this hypothesis.

Context Changes Within and Between Lists

Context change within list (for one summary of the REM approach see Malmberg & Shiffrin, 2005) and context change across lists are each important in their own right. Criss and Shiffrin (2004a, Exp. 2) contrasted within and between list manipulations. They presented three long lists separated by arithmetic periods; participants carried out a different incidental task for each list (in lieu of study for a memory test). Some words were repeated across various combinations of the three lists. At the end of the experiment, participants received an unexpected recognition task in which they were asked to respond "old" only to items appearing on the most recent list (targets), and "new" to items from previous lists or to items that had not appeared in the experiment (foils). "Old" responses to foils from recent lists give evidence concerning context confusions. Word confusions were examined through the use of a semantic manipulation: Only for List 3 (the most recent one), participants studied differing numbers of words from different categories. Thus "old" response probabilities that varied with List 3 category size provided evidence pointing to confusions among words based on similarity. Evidence was found for both context- and word-based interference, albeit the context effect was larger (see Table 11.1). For example, the hit rate for items appearing on list three

TABLE 11.1 Probability of "Old" Response to Probes Appearing on Various Combinations of Three Experimental Lists

	Previous list appearances of probe			
Probe type	**List 1,2**	**List 2**	**List 1**	**None**
Target (on List 3)	.92 (.02)	.92 (.02)	.91 (.02)	80 (.02)
Lure (not on List 3)	.74 (.02)	.61 (.03)	.56 (.03)	.10 (.02)

Note: Results are reproduced from Criss and Shiffrin (2004a, Exp. 2). Participants were asked to respond "old" only to items on the current list (List 3). Means are listed followed by standard error in parentheses.

alone was scarcely higher than the false alarm rate for items that had been studied on both Lists 1 and 2 but not 3 (a similar false alarm effect was observed in the associative recognition domain by Criss & Shiffrin, 2005). This finding may indicate any or all of the following: (1) that context does not change much between lists with different incidental tasks separated by arithmetic; (2) that retrieval uses context features appropriate for all three lists instead of List 3 specifically; (3) that context and task-type features are not well integrated with word-specific features in storage; and (4) that context and task-type features are too little emphasized when constructing a retrieval probe. The studies to follow help narrow the possibilities.

The One-List-Back Paradigm and List Discrimination

Shiffrin (1970) had participants study twenty successive lists of words, each containing 5 or 20 unrelated items. A free recall period followed each list (except the first), during which participants were asked to recall items not from the most recent list, but from the list prior to that. As shown in Figure 11.2, the level of recall was determined by the length of the to-be-recalled list, without regard to the length of the intervening list. Shiffrin suggested that participants could reconstruct a context cue enabling them not only to access the older list, but also to focus sufficiently on that prior list to prevent interference by items on the intervening list.

According to REM and similar theories, elimination of interference from the intervening list could not have occurred unless context changed significantly between lists. Attempting to obtain converging evidence, Shiffrin and Rosenthal (1973) had subjects study groups of three successive study lists of varying lengths (all combinations of 20 and 5), separated by brief periods of arithmetic. No testing was administered until the end of each list triad, when participants tried to free recall each of the three lists; instructions at the start of recall indicated whether the order of list recall should be from the first list to the last, or vice versa. Figure 11.3 shows data from this study, which indicates that the level of recall was determined by the sum of all three list lengths rather than the target list length. In this study, it appeared, therefore, that the probe context pointed to the entire group of three lists—either context did not change much between lists, or individual list contexts could not be reconstructed. This model is a bit too simplistic, however, because it

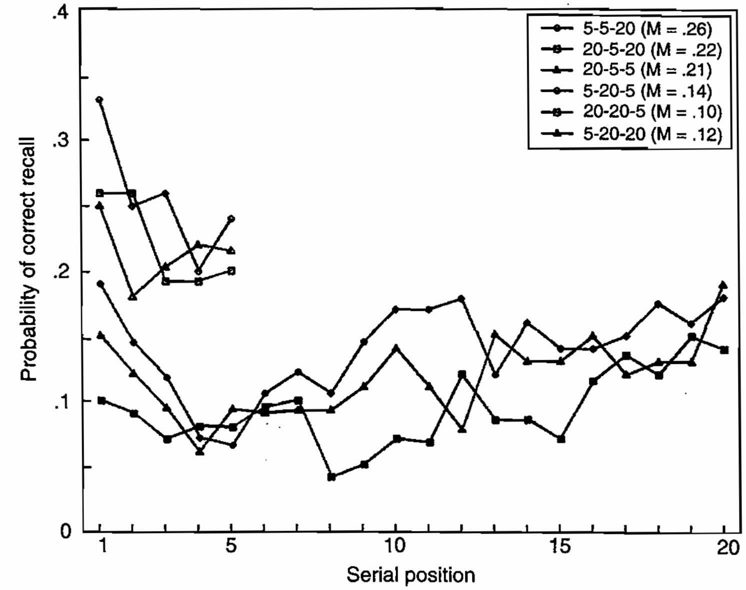

FIGURE 11.2 Results from Shiffrin (1970): Recall of the list before the last in various target and intervening list-length conditions. List-length effects were determined by the length of the target list, regardless of intervening list length. Figure based on Shiffrin (1970).

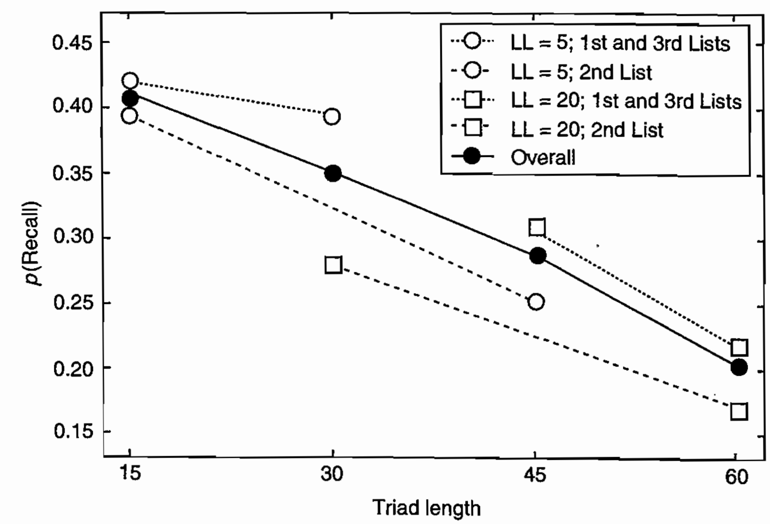

FIGURE 11.3 Results from Shiffrin and Rosenthal (1973): Recall as a function of total length of a triad of lists separated by arithmetic. Results are shown for list lengths of 10 and 20, collapsed across conditions when the lists were recalled proceeding from the first to last and last to first.

would suggest a complete inability to discriminate list origin within the triad. In fact, although the participants did make many intrusions, the level of target list recall exceeded the summed level of intrusions from the other two lists. Thus sampling of traces may have been independent of the target list, but some list origin information may have been stored with some traces, allowing editing of recalls during the recovery or output phases of retrieval.

This account of the results does not explain the differences between the present results and those of Shiffrin (1970). In the earlier study, recalls occurred between lists, but the present study used only arithmetic between lists. Could context change be induced by recalls but not by arithmetic? We will re-examine this possibility when we present results from a study of temporal order judgments.

Samuelson (1993) revisited the Shiffrin (1970) design using recognition testing: Participants studied successive lists of 9 or 36 items, each followed by a recognition test, with a permuted order of lengths so that various combinations of three successive lengths would occur. The control group was asked to respond "old" to items from the most recent list, while the experimental group responded "old" to items from the second most recent list. Study items did not repeat across lists, but test items included not only previously unstudied items, but items from the three most recent lists.

Using *d'* as a measure, the control condition had reasonable levels of performance that was higher for the shorter list than a longer list, a standard finding. The experimental condition showed generally poor performance, only slightly above chance, but with some indications of better performance for shorter target lists. Figure 11.4 gives a summary of the temporal results, summing across list lengths, in terms of a pseudo-*d'* measure: For each condition, and each pair of lists, *d'* is calculated from a *p*(hit) defined as the *p*(old) for the list closer to the target list or more recent in the case of ties, and a *p*(false alarm) defined as the *p*(old) for the list farther from the target list, or less recent in the case of ties. For these purposes, new words were considered "most distant." In the control condition, the performance pattern was normal, with *d'* falling off with temporal distance of the foil, probably indicating that participants used the present context as a cue, with moderate success. In the experimental condition, the pattern of *p*(old) responses among lists *n*–1, *n*–2, and new items was quite similar to that in the control condition, perhaps suggesting that context changes from list to list were similar in the two conditions. However, ability to discriminate list *n* items from list *n*–1 items was at chance. If participants had used the normal list *n* context cue, and said "old" to items *below* a familiarity criterion, they would have performed better than chance. Thus the participants may have tried to reconstruct a list *n*–1 context cue, but failed to do so effectively. Note that this account differs from those used to explain the two free recall studies. Perhaps in recognition the availability of word features induces an over-reliance on such features, to the detriment of retrieval emphasizing context.[3]

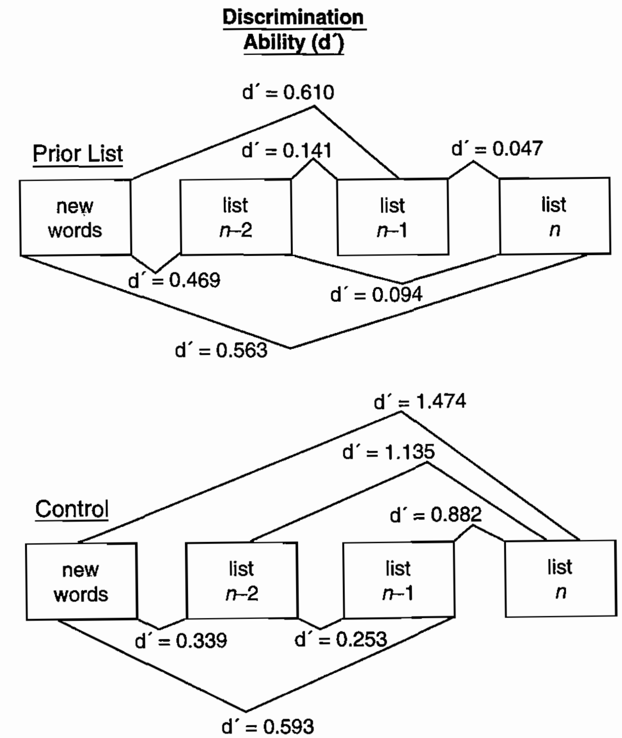

FIGURE 11.4 Results from Samuelson (1993): Recognition performance following study of a triad of lists. Diagrams show *d'* between targets and items appearing on various nontarget lists. The top panel illustrates performance when participants were asked to say "old" to items appearing on the penultimate list. The bottom panel illustrates performance when participants were asked to say "old" to items appearing on the final list.

Judgments of Recency

The theme of the studies discussed thus far is the idea that the benefit gained by using context features in the retrieval probe depends positively on its similarity to context features stored with the to-be-retrieved information, and negatively on its similarity to contexts stored with other traces. Our studies suggest that context changes little when a simple math task separates the study lists (supported by Criss & Shiffrin, 2004a, when the lists are studied using different semantic tasks; Shiffrin & Rosenthal, 1973), but changes to a greater degree when memory tests separate successive study lists (e.g., Samuelson, 1993; Shiffrin, 1970). However, a benefit to using context as a cue depends on its being appropriate for the task, and reconstructing an appropriate context for a list other than the most recent is sometimes very difficult (Samuelson, 1993). These observations are based on indirect tests of context change because the actual task presented to the participant was a test for item information. We turn next to studies that attempt to assess context changes more directly.

Klein et al. (2004, 2006) asked participants to make relative judgments of recency (JOR). In their paradigms, participants study a list of words, and tests consist of two choice words (2AFC) with instructions to select the word that occurred more recently. Particularly in recent years, most JOR studies have used a continuous paradigm: A long list of words, with repetitions, is presented, and for each word presented participants indicate how many items back they last saw the currently presented word (Hintzman, 2000, 2004). We chose instead to use the 2AFC procedure because it is better suited to distinguish effects due to changes in accuracy from those due to changes in response bias. In previous studies using this 2AFC approach, it has been found that people can make these judgments with above-chance accuracy (Lockhart, 1969; Yntema & Trask, 1963), with performance increasing as a function of interitem lag (i.e., the number of items studied between the two test words during the study phase) and decreasing as a function of study–test lag (i.e., the number of items studied between the most recent of the test words and the time of judgment/test). Other 2AFC JOR tasks have used interspersed study and test trials (e.g., Yntema & Trask, 1963) but some of our results suggest that the interspersed activity itself makes a critical difference in context changes. We therefore presented a long list followed by a 2AFC test.

We began our line of research using 2AFC tests of items at six lags ranging from 2 and 24, studied in a 90-item list, in a study using 30 participants. We were surprised to find that this manipulation resulted in performance that was not statistically different from chance, for any of our lag lengths (Klein et al., 2004; recent data from additional participants did show performance at the longest lags rising just barely above chance, but it is safe to conclude that performance is very close to chance). In subsequent studies we tested longer lags—lags of 36 in 90-item lists—and found Ss could make 2AFC judgments statistically above chance, but still only at about 60% correct. The difficulties in making such judgments suggest that the contextual features stored with list items change very slowly during the course of a single random word list (in cases when the list is presented without interspersed testing). Thus the results are consistent with an REM model that assumes small context changes within lists and larger changes between lists. Why should our studies have revealed such difficulty, when previous studies revealed somewhat better performance? The answer may lie in our use of longer lists and longer study–test lags.

The picture that is emerging from these studies is one in which context changes slowly within list, and more rapidly between lists. But what are the conditions that promote context change between lists? Recall that the Shiffrin's (1970) free recall results suggested when lists were separated by a recall task, enough context change occurred between lists to allow a "one-back" context to be reconstructed, and eliminate most interference from the intervening list. The Shiffrin and Rosenthal (1973) study with arithmetic separating a triad of lists suggested context did not change much. The unpublished recognition results of Samuelson (1993) suggested that either context did not change much, or the recognition task degraded the ability to reconstruct a prior list context. We decided to investigate the conditions that promote context change across lists using the paradigm of comparative recency judgments.

Two hundred participants studied a 160-item word list that was split in half by one of four 90-s intervening tasks (each participant received only one list and thus the intervening task was manipulated between subjects). The degree of recency discrimination when the test pair includes items on each side of the intervening task should provide evidence about the degree of context change promoted by that task. The four tasks separating the two list-halves are denoted math, faces, imagine, and recognition. In the math condition, participants added a series of numbers together. Based on Shiffrin and Rosenthal (1973), we hypothesized that math would promote little or no context change. In the faces condition, participants studied a list of faces, allegedly for later test. This condition was intended as a control condition where items continue to arrive, perhaps promoting context change, but without affecting list length effects due to numbers of words (Criss & Shiffrin, 2004c, have shown that words and faces appearing on the same list do not interfere with one another). In the imagine condition, participants were asked to write a paragraph answering the question, "What would you do if you were invisible and would not be responsible for your actions?" Sahakyan and Kelley (2002) found that answering this question following a list that participants have been instructed to forget leads to higher recall of the following list and lower recall in a surprise test of the "forgotten" list than occurs when participants take a break with no intervening task between lists. They take this as evidence that the task changes a participant's internal context. For this reason, we expected this condition to promote context change between list-halves in our JOR task. Finally, in the recognition condition, participants completed an old/new recognition test of target words taken from some of the words in the first half of the list (none of which would be used in the later 2AFC JOR recency tests) and foil words not from the list. Based on Shiffrin (1970), we hypothesized that such testing might promote context change.

Following presentation of the 160-item list with the embedded intervening task, participants gave 2AFC JOR judgments. There were three types of test pairs: first half pairs, second half pairs, and pairs that crossed the intervening task. On the average, the lag between pair items, calculated as number of list items (ignoring intervening tasks) was constant across the three types of pair test, and was the same for the four types of intervening task. The pairs that crossed the intervening tasks were equated both in terms of number of intervening study items and total time.

Table 11.2 displays 2AFC JOR performance. Consistent with many of our results for long lists, performance was extremely low. List 1 discrimination was poor in all conditions—at or below chance. List 2 discrimination was generally above chance, but more so for math and faces. Of more interest, cross-list performance was best for imagine and recognition, and near chance for math and faces, exactly as expected. Our expectations were met for imagine and recognition: Context change between list-halves, but not within, would have produced above chance performance only for the between-halves tests (L1–L2). The results for math and faces were puzzling: If above chance within the second half-list, why

TABLE 11.2 Recency Discrimination for 160-Item Lists Broken Up with Four Different Distractor Tasks

	Comparison condition		
Distractor condition	**Within first half**	**Across both halves**	**Within second half**
Math problems (math)	.49 (.04)	.53 (.03)	.57 (.04)
Study faces (face)	.49 (.04)	.53 (.05)	.55 (.04)
"Imagine" essay (imagine)	.49 (.04)	.55 (.04)	.53 (.05)
Recognition test (recognition)	.53 (.04)	.58 (.04)	.56 (.04)

Note: Results from Klein et al. (2006). Means are listed, followed by 95% confidence intervals on the mean in parentheses.

should they have been at chance for the cross-list tests? The following hypothesis is one way to account for all the results. Suppose for math and faces there is no context shift due to the intervening task but a general serial position strength-of-storage gradient across the entire list (both halves together), with primacy due to extra storage in long-term memory, and recency due to residual strength in short-term memory. If strength were used by participants as a stand-in for recency, then the strong primacy items would tend to be chosen as more recent than weaker later items, dropping performance below chance for L1–L1 tests. The lack of strength differences in the middle of the list would produce chance performance for L1–L2 tests. The stronger recency items would tend to cause these to be selected as more recent, pushing performance above chance for L2–L2 tests. For the imagine and recognition conditions, context changed between list halves, allowing current context to be used for judgments, producing the results as described above. As an aside, it should be noted that because our cross-task results were obtained in conditions in which time between test items was held constant across conditions, a more complex model is needed than one based strictly on time-related information (e.g., Hintzman, 2004).

To test the idea that participants might use strength as a stand-in for recency, particularly when context information is not easily usable for recency judgments, we carried out a 2AFC JOR study varying item strength. Although a number of studies have used multiple presentations of items to examine effects of strength on JOR performance (e.g., Peterson, Johnson, & Coatney, 1969; Wells, 1974), interpretation of such paradigms is complicated by potentially different mechanisms for accessing frequency and recency information, so we varied strength by varying study time. We tested comparative JOR for all combinations of long-presented (strong), short-presented (weak), and unstudied (new) items. Fifty participants received two 90-item lists consisting of half weak items (800 ms stimulus onset asynchrony; SOA), and half strong items (2500 ms SOA) randomly intermixed. Tests following each list were 2AFC for studied items with lag 16, or between an old item and a new item. The results for the six key conditions are given in Figure 11.5. The second letter indicates the more recent presently item. When items were of equal strength (i.e., W–W and S–S), performance was at chance (indicated

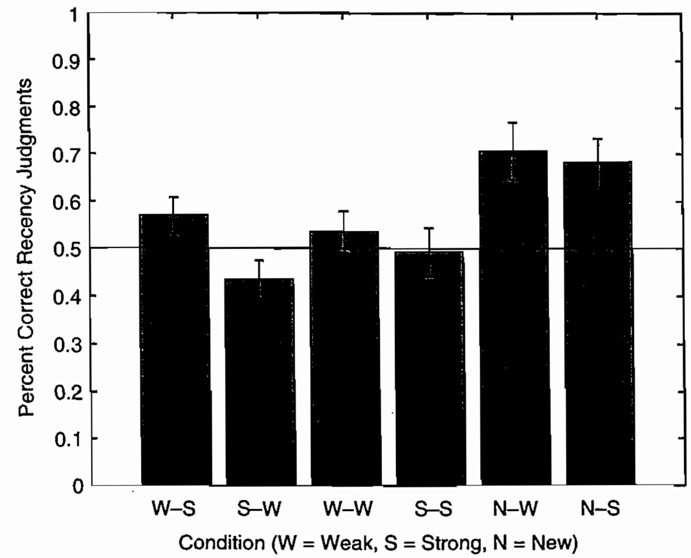

FIGURE 11.5 Results from Klein et al. (2006): Recency discriminations for pairs including weak (W) items (800 ms study), strong (S) items (2400 ms study), and new (N) items (0 ms study). Error bars represent 95% confidence intervals on the standard error of the mean. Condition names represent the status of the nonrecent item (i.e., S, W, or N) followed by the status of the recent item. The first two columns therefore represent a bias to pick the S item as being more recent.

by the horizontal line). However, when the two test items differed in strength, participants chose the stronger of the items regardless of which item was actually presented more recently. Thus performance in the W–S condition is "above chance" because the strong item was more recent but performance in the S–W was "below chance" because the strong item was less recent. The N–S and N–W pairs simply indicate that the participants actually do know (to some extent) which items were on the list but apparently do not know the relative order of presentation for the items. In summary, the data indicate that in conditions when context-based judgments do not work well, participants substitute a strategy, or bias, of choosing the stronger item as more recent, confirming the speculation derived from the prior study.

TOWARD A MODEL OF CONTEXT

"Context" is hardly the only useful concept in the field with an amorphous definition, a large domain of application, and different usages by different investigators in different settings (cf. "attention"). The studies and results we have described refine the concept and suggest some tentative conclusions and avenues for further exploration.

We advise against thinking of context as "one thing"; as described in the introduction, there are many different types of context information, having different properties. It is potentially very misleading to lump all these forms together into a unitary concept termed "context". However, we must emphasize that the variability in types of context does not preclude our assumption that the general rules of storage and retrieval apply to both explicitly and experimentally manipulated information. These include any and all types of context as well as the "content" features of the studied words themselves. A multitude of empirical studies addressing individual context types have shown context effects. Given the number of factors that comprise context, and the usual procedures that vary just a few of these experimentally, it probably should not come as a surprise that the effect sizes are usually quite modest. Further, one should not downgrade the importance of context as a memory component because large effects are difficult to produce.

It is important to recognize the critical roles played by (1) the differential attention that different forms of context information might receive, (2) the natural and sometimes cyclic changes that different forms of context undergo (e.g., changes with time of day, with different levels of sunlight, with day of week, with biological cycles), (3) the contributions to these changes that are controlled by the participant, consciously or otherwise; and (4) the degree of integration of such context information with the task-relevant content information.

The form of context studied in most of our tasks is that involved in list membership specification, a form probably including all of the kinds of context information mentioned in this chapter. We note that typical experimental control reduces external environmental changes to a small factor. In this case, internal environmental changes become extremely important. Our results, from various free recall, recognition, and temporal judgment tasks, suggest this internal context changes slowly within list, and more strongly between list, when lists are separated by appropriate markers and tasks. These markers and tasks include testing of words from lists, and devising stories. On the other hand mental arithmetic and face judgments between lists change internal context very little. These results lead us to the following speculation: At least in the absence of large external environmental changes, people may utilize a hold-until-shift approach to internal context change. That is, the internal states that contribute to list context may be held somewhat unchanged until some trigger event causes that held context to be dropped, and causes the production of changed internal context or perhaps the gathering of a new sample of internal context.

Our results also suggest that in cases when context does change sufficiently it may be possible to reconstruct a context cue for an older list, but this is difficult and may only be possible in rare circumstances. Furthermore, repeating items at study and test across lists causes great context confusion and seems to make it almost impossible to do anything but use the present context at test—attempts in such circumstances to reconstruct old context seem to harm performance without producing a counteracting gain. When context is a poor cue for a given task, subjects seem to substitute alternative forms of information as a means for making contextual judgments; e.g., they may use item strength as a stand-in for recency. In the cases when early items are stronger due to primacy effects, such strategies can even reduce performance below chance.

We have discussed the results described in this chapter with respect to the REM model. Context plays a critical role in REM, as it did in SAM, and indeed as it does in most memory models. It is important to note that the role of context in these models is not primarily to explain the effects that are found in studies that directly manipulate context, although this is one of the justifications. The requirement for the context construct is based primarily on the need to explain a host of other memory effects involving learning, forgetting and priming, just a few of which we have touched on in this chapter (such as the priming results when study time is varied, and the "one-shot" study that was motivated by those results). At this time of writing, models that include context provide by far the most broad and coherent account of memory phenomena.

Even a casual reader of this chapter will quickly reach the conclusion that context is difficult to study empirically and model theoretically. Attempts so far to do so are in the earliest and far too simplistic stages (e.g., Mensink & Raaijmakers' 1988 extension to SAM, or Howard & Kahana's 2002 model). We nonetheless expect progress to continue, and hope to continue to contribute to empirical and theoretical advances in this domain.

NOTES

1. When referring to the accumulation of information over a lifetime of experiences, lexical memory and/or semantic memory are often used to describe the resultant information for words. We now prefer to use the more general term knowledge (knowledge memory, or knowledge traces), to emphasize the application to every kind of knowledge we learn, store, and can retrieve.
2. The traces in BCDMEM are composite rather than separate, but this is not critical for the present discussion.
3. Huber, Jang, and Overschelde (2005) carried out one-back studies using both recognition and recall testing. They found that the type and duration of activity separating lists produced critical differences in the results, and built a context model to explain their findings, but the findings are too recent to be discussed in this chapter.

REFERENCES

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, 79, 97–123.

Criss, A. H., & Shiffrin, R. M. (2004a). Context noise and item noise jointly determine recognition memory: A comment on Dennis and Humphreys (2001). *Psychological Review*, *111*, 800–807.

Criss, A. H., & Shiffrin, R. M. (2004c). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory and Cognition*, *32*, 1284–1297.

Criss, A. H., & Shiffrin, R. M. (2005). List discrimination and representation in associative recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1199–1212.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*, 452–478.

Dulsky, S. G. (1935). The effect of a change of background on recall and relearning. *Journal of Experimental Psychology, 18*, 725–740.

Fernandez, A., & Glenberg, A. M. (1985). Changing environmental context does not reliably affect memory. *Memory and Cognition, 13*, 333–345.

Gantt, W. H. (1940). The role of the isolated conditioned stimulus in the integrated response pattern, and the relation of the pattern changes to psychopathology. *Journal of General Psychology, 23*, 3–16.

Hintzman, D. L. (2000). Judgments of recency and their relation to recognition memory. *Memory and Cognition, 31*, 26–34.

Hintzman, D. L. (2004). Time versus items in judgment of recency. *Memory and Cognition, 32*, 1298–1304.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*, 269–299.

Huber, D. E., Jang, Y., & Overschelde, J. P. V. (2005, November). *Using sampling and recovery to estimate context and item effects in memory*. Paper presented at the annual meeting of the Psychonomic Society, Toronto, Ontario, Canada.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*, 306–340.

Klein, K. A., Criss, A. H., & Shiffrin, R. M. (2004). Recency judgments and list context [Abstract]. In *Proceedings of the twenty-sixth annual meeting of the Cognitive Science Society* (p. 1578). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Klein, K. A., Criss, A. H., & Shiffrin, R. M. (2006). Recency revisited. *Manuscript in preparation*.

Lockhart, R. S. (1969). Recency discrimination predicted from absolute lag judgments. *Perception and Psychophysics, 6*, 42–44.

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 322–336.

Mensink, G., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review, 95*, 434–455.

Peterson, L. R., Johnson, S. T., & Coatney, R. (1969). The effect of repeated occurrences on judgments of recency. *Journal of Verbal Learning and Verbal Behavior, 8*, 591–596.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88*, 93–134.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 163–178.

Roediger, H. L., & Challis, B. H. (1992). Effects of exact repetition and conceptual repetition on free recall and primed word-fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 3–14.

Roediger, H. L., & McDermott, K. B. (1993). Implicit memory in normal human subjects. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 8, pp. 63–131). Amsterdam: Elsevier.

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 1064–1072.

Samuelson, L. (1993). *Context cuing in recognition memory*. Unpublished undergraduate honors thesis, Indiana University, Bloomington, IN.

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review, 108*, 257–272.

Shiffrin, R. M. (1970). Forgetting: Trace erosion or retrieval failure? *Science, 168*, 1601–1603.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 179–195.

Shiffrin, R. M., & Rosenthal, J. (1973). [Free recall of three successive lists]. Unpublished raw data.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin and Review, 4*, 145–166.

Smith, S. M. (1988). Environmental context-dependent memory. In G. M. Davies & D. M. Thomson (Eds.), *Memory in context: Context in memory* (pp. 13–34). Oxford, UK: Wiley.

Smith, S. M. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin and Review, 8*, 203–220.

Strand, B. Z. (1970). Change of context and retroactive inhibition. *Journal of Verbal Learning and Verbal Behavior, 9*, 202–206.

Underwood, B. J. (1977). *Temporal codes for memories: Issues and problems*. Oxford, UK: Lawrence Erlbaum Associates, Inc.

Wells, J. E. (1974). Strength theory and judgments of recency and frequency. *Journal of Verbal Learning and Verbal Behavior, 13*, 378–392.

Wright, D. L., & Shea, C. H. (1991). Contextual dependencies in motor skills. *Memory and Cognition, 19*, 361–370.

Yntema, D. B., & Trask, F. P. (1963). Recall as a search process. *Journal of Verbal Learning and Verbal Behavior, 2*, 65–74.