Psychological Review
2000, Vol. 107, No. 2, 368–376

0033-295X/00/$5.00 DOI: 10.1037//0033-295X.107.2.368

The Case Against a Criterion-Shift Account of False Memory

John T. Wixted and Vincent Stretch
University of California, San Diego

M. B. Miller and G. L. Wolford (1999) argued that the high false-alarm rate associated with critical lures in the Roediger–McDermott (H. L. Roediger & K. B. McDermott, 1995) paradigm results from a criterion shift and therefore does not reflect false memory. This conclusion, which is based on new data reported by Miller and Wolford, overlooks the fact that Roediger and McDermott's false-memory account is as compatible with the new findings as the criterion-shift account is. Furthermore, a consideration of prior work concerned with investigating the conditions under which participants are and are not inclined to adjust the decision criterion suggests that the criterion-shift account of false memory is unlikely to be correct.

Roediger and McDermott (1995) resurrected a procedure first used by Deese (1959) that seemed to reliably produce false memories of words. The procedure involves asking people to memorize items that are closely related to a critical item that did not appear on the list. For example, participants might hear 15 words related to the critical item *needle* (e.g., *thread, pin, eye, thimble, sharp,* etc.), but *needle* itself would not be presented. After studying a list like this, participants were very likely to falsely recall or to falsely recognize the nonpresented critical item. Remarkably, Roediger and McDermott (1995) found that the false-alarm rate to these critical lures (.72 in one experiment) generally equaled or exceeded the hit rate to the related items that constituted the list (.65 in that experiment). Such findings appear to suggest that false memories can be created in the laboratory with remarkable ease. However, Miller and Wolford (1999) argued that, in this case, appearances can be misleading.

Miller and Wolford (1999) argued that the high false-alarm rate to critical lures in the Roediger–McDermott (Roediger & McDermott, 1995) paradigm results from a criterion shift and does not represent false memory. The specifics of their argument are presented in more detail below, but the essence of their claim is that participants tend to respond "old" to a critical lure, not because they falsely remember having studied it but because they realize that it is closely related to the theme defined by the list items. That realization, according to this account, leads one to believe that there is a good chance that the critical item appeared on the list. On the basis of that belief, participants are willing to respond "old" to a critical item even though they do not actually remember having encountered it. In other words, the high false-alarm rate to critical lures reflects a criterion shift, not false memory.

John T. Wixted and Vincent Stretch, Department of Psychology, University of California, San Diego.

We thank Henry Roediger, Thomas Wickens, Elliot Hirshman, and George Wolford for their valuable reviews.

Correspondence concerning this article should be addressed to John T. Wixted, Department of Psychology 0109, University of California, San Diego, La Jolla, California 92093-0109. Electronic mail may be sent to jwixted@ucsd.edu.

To test this possibility, Miller and Wolford (1999) modified the Roediger–McDermott (Roediger & McDermott, 1995) procedure by including some lists in which the critical item actually appeared (e.g., in the above example, the word *needle* might have appeared on the list along with the other related items). McDermott (1997) had done this as well and showed that the presented critical items were much better recalled than the other items on the list. However, Miller and Wolford went further in that they tested recognition and used the obtained hit and false-alarm rates for critical items (.97 and .81, respectively, in their Experiment 1) and related items (.88 and .36, respectively) to compute estimates of response bias based on signal-detection theory. Those estimates revealed that the critical items were indeed associated with a more liberal response bias than the related items, a finding they construed as supporting their criterion-shift account.

In a reply to Miller and Wolford's (1999) comment, Roediger and McDermott (1999) argued that a criterion-shift account of false memory is unlikely to be correct because participants continue to falsely recognize critical lures at a high rate even when they are given clear warnings about the existence of those lures and clear instructions to avoid choosing them (information that should prevent them from using a liberal criterion for critical items). In addition, participants typically report clear perceptual details associated with these false memories and often give them "remember" (as opposed to "know") responses. In other words, participants act as if they actually remember hearing the critical items.

Another critical point that has not yet been fully discussed is the fact that the bias parameters reported by Miller and Wolford (1999) are precisely what the most straightforward false-memory model would predict. Miller and Wolford devoted most of their article to an attack on the wrong false-memory model, one that has not been taken seriously by anyone but them. After convincingly refuting that model and arguing in favor of the criterion-shift account, they briefly acknowledged that an alternative signal-detection model that does not involve a criterion shift is, after all, compatible with their reported bias measures. What they do not seem to have recognized is that this alternative model, which they say was suggested by a reviewer, is the false-memory model that

corresponds most closely to the views expressed by Roediger and McDermott (1995). Thus, Roediger and McDermott's view of false memory is not at all incompatible with the bias measures reported by Miller and Wolford (in fact, it predicts them).

How is it possible that a fixed-criterion false-memory model is viable in light of the different estimates of response bias associated with related and critical items? Given the assumptions of signal-detection theory, it is true that a change in a participant's decision strategy entails a change in measured bias, but it does not follow that a change in measured bias (which is what Miller and Wolford, 1999, found) implies a change in the participant's decision strategy. Methods other than the mere computation of signal-detection bias parameters are needed to address that issue. To illustrate why, we turn now to a discussion of three models: the false-memory model that most closely approximates the views expressed by Roediger and McDermott (1995; Model 1), the false-memory model conceived and refuted by Miller and Wolford (Model 2), and the criterion-shift model that Miller and Wolford ultimately determined to be supported by their findings (Model 3). After presenting these three models and evaluating them in light of Miller and Wolford's new data, we consider evidence suggesting that the criterion-shift account offered by Miller and Wolford is unlikely to be correct. Model 1, by contrast, is consistent with all of the available evidence.

Model 1 (Roediger and McDermott's, 1995, False-Memory Model)

Roediger and McDermott (1995) did not provide a mathematical version of their model, but a mathematical statement of their ideas is relatively straightforward. To place their ideas into a signal-detection framework, we make the standard assumption that recognition decisions are made on the basis of a unidimensional "strength of evidence" variable that represents the degree to which one remembers having recently encountered an item. Familiarity is one such variable that is commonly used to illustrate the application of signal-detection theory to recognition memory, but Roediger and McDermott would probably prefer the more neutral term *strength of evidence* because memories (both true and false) are richer in perceptual detail than a concept like familiarity can do justice to (e.g., Roediger, McDermott, & Robinson, 1998). Neglecting preexperimental strength (which may actually differ slightly for critical and related items), the strength of item i (S_i) on a recognition test is, in this model, assumed to be a function of the direct effects deriving from the presentation of the item (P_i) plus indirect effects due to associative activation from the other items on the list (A_i):

$$S_i = P_i + A_i.$$

This simple model is consistent with a wide variety of more specific theoretical assumptions about how strength is actually incremented. For example, the strength represented by P_i could be assumed to derive either from the conscious rehearsal of a presented item or from the enhanced perceptual fluency an item receives by virtue of being perceived through the senses. Either way, P_i represents a quantity of strength that would not exist had the item not appeared on the list. Similarly, the strength represented by A_i could be assumed to derive either from the conscious rehearsal of an associatively activated item or from a more passive (and unconscious) spread of associative activation. Either way, A_i represents a quantity of strength that would not exist had the item not been associatively activated.

The application of Model 1 to the procedure used by Miller and Wolford (1999) is straightforward. Their procedure actually involved three classes of items: critical items (e.g., *needle*), related items (e.g., *thread, pin, eye*), and unrelated items (items that were not related to any of the items that appeared on the list). According to Model 1, strength due to item presentation (P_i) for lures from all three classes must be equal to zero because these items did not appear on the list. For both targets and lures, strength deriving from associative activation (A_i) is, by design, highest for critical items, next highest for the related items, and lowest for unrelated items. In the simplest version of this model, P_i and A_i are assumed to be uncorrelated, which means that the effect of study time is, on average, the same for all three kinds of items. This assumption is also implicit in the analysis offered by Miller and Wolford and is supported by their receiver operating characteristic (ROC) analysis. As they noted, "The equality of sensitivity across item types indicates that the critical lures profit as much as any other item type by being presented" (p. 401). Actually, the linearity of their ROC is debatable, and their data appear to be compromised by ceiling effects for the critical targets. Nevertheless, for the sake of simplicity, we shall accept this aspect of their position at face value. Future research may prove this assumption wrong, but it would not change the essence of the argument presented by us or by them.

Although Roediger and McDermott (1995) could certainly have advanced a more complicated model, Model 1 appears to be the simplest possible mathematical statement of their ideas. They noted in their article, for example, that "the earliest idea about false recognition—the implicit associative response—still seems workable in helping to understand these phenomena" (p. 810). They then cited Underwood's (1965) classic article, an article that presents an account much like Model 1. For example, Underwood stated that there are two kinds of responses to a "verbal unit" (i.e., a word) on list: "There is first the response made to the unit itself as the act of perceiving it" (p. 122), which he termed the *representational response*, and there is, second, the *implicit associative response* that consists of "another word which is associated with the actual word presented" (p. 122). In other words, items acquire evidence of having been seen before both from being presented on the list (P_i) and from associative activation (A_i), which is all that Model 1 assumes. Although Roediger and McDermott have never unequivocally endorsed Underwood's account in their various writings, and although they typically discuss other interpretations of false memory as well, it is hard to find another model that has figured more prominently in their thinking.

To illustrate the behavior of this simple model, assume that the average value of P_i is 1 unit of strength on some arbitrary scale (i.e., briefly presenting an item on a list increases its strength by 1 unit on average). Further assume that the average value of A_i is 2 units for critical items, 1 unit for related items, and 0 units for unrelated items (i.e., the unrelated items acquire no strength

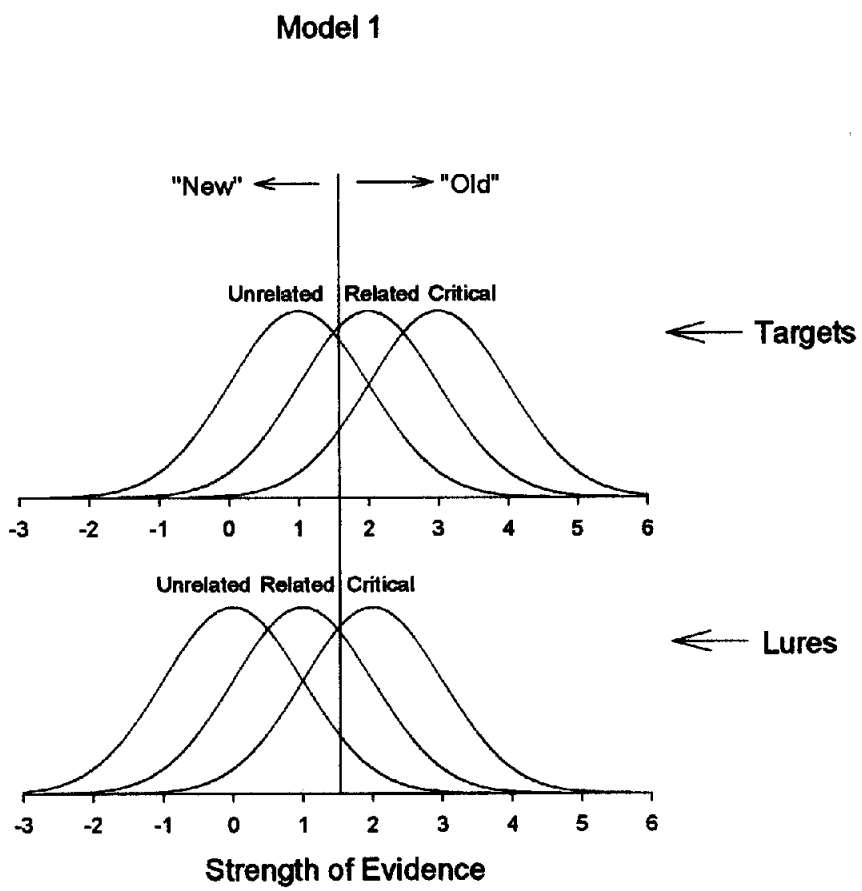

Figure 1. Hypothetical target and lure distributions for related, unrelated, and critical items according to the false-memory account of Roediger and McDermott (1995).

through associative activation).[1] Given these values, the average strength values for the various item types are as follows:

Critical target (CT): $S_{CT} = 1 + 2 = 3$
Related target (RT): $S_{RT} = 1 + 1 = 2$
Unrelated target (UT): $S_{UT} = 1 + 0 = 1$

Critical lure (CL): $S_{CL} = 0 + 2 = 2$
Related lure (RL): $S_{RL} = 0 + 1 = 1$
Unrelated lure (UL): $S_{UL} = 0 + 0 = 0.$

These are mean values; individual items will be distributed normally about these means, as illustrated in Figure 1. The standard deviation of the target and lure distributions is arbitrarily set to 1 in this example (although, in practice, the target distribution is usually somewhat more variable than the lure distribution). To arrive at a recognition decision, participants set a single decision criterion (placed at 1.5 in this example) and evaluate all items on the recognition test against that criterion (i.e., they do not shift their decision criterion on an item-by-item basis).[2] Items that exceed the decision criterion are judged to be "old," whereas items that fall below the criterion are judged to be "new."

Note that the means of the critical lure and related target distributions are equal (i.e., $S_{CL} = S_{RT}$), so the hit rate for related targets will equal the false-alarm rate for critical lures. This is the signal-detection interpretation of what Roediger and McDermott (1995) meant when they said that "subjects were unable to distinguish items actually presented from the critical lures that were not presented" (p. 808), and, more generally, it is the signal-detection interpretation of false memory (i.e., the "strength of evidence" for critical lures rivals that of the targets). Curiously, Miller and Wolford (1999) seem to have interpreted this sentence to mean that Roediger and McDermott were asserting that critical lures are as strong as they would be had they actually appeared on the list (i.e., that $S_{CL} = S_{CT}$) and that this outcome constitutes the essence of false memory. Apparently as a result of this misinterpretation, they designed their experiments to rule out a model that no one previously advocated rather than the model that is most consistent with the views expressed by Roediger and McDermott (1995). We turn now to a more detailed consideration of Miller and Wolford's view of false memory.

[1] These are hypothetical values designed to illustrate Model 1 in a simple way. More realistic values could be obtained by actually fitting the model to empirical data as Wickens and Hirshman (2000) have done.

[2] The criterion presumably shifts somewhat from item to item as a result of sequential dependencies, but this kind of movement would not affect the present arguments so long as those sequential dependencies were similar across item types.

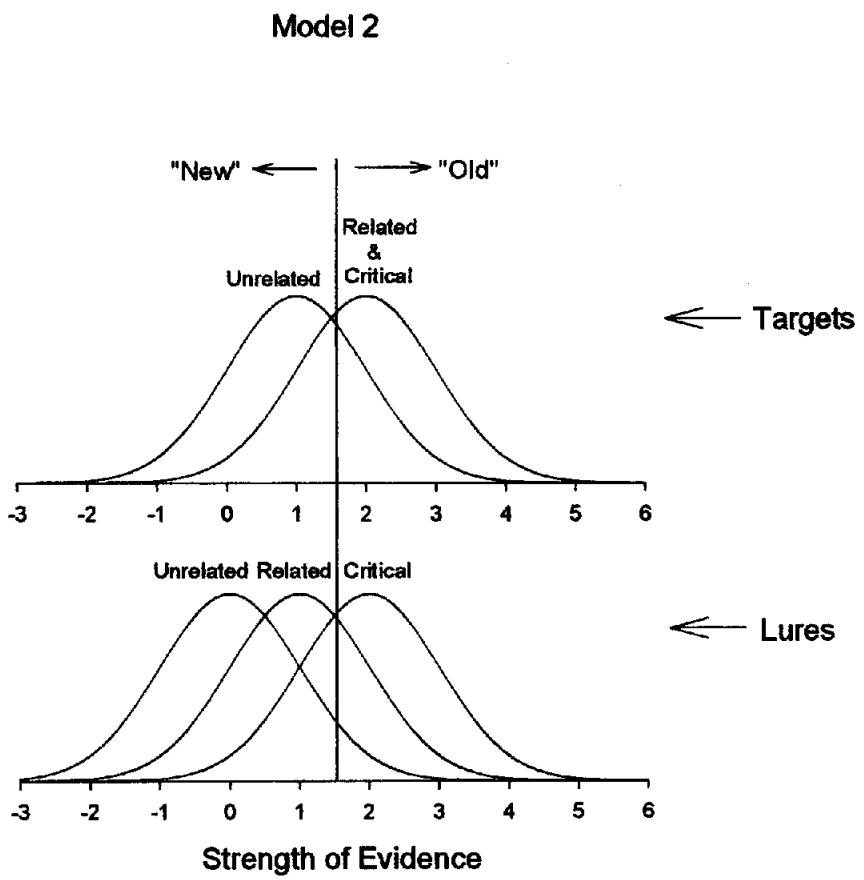

Figure 2. Hypothetical target and lure distributions for related, unrelated, and critical items according to the false-memory account of Miller and Wolford (1999).

Model 2 (Miller and Wolford's, 1999, False-Memory Model)

The false memory model targeted by Miller and Wolford (1999) is illustrated in Figure 2, and its essential feature is that sensitivity for critical items is equal to zero. This is a curious assumption without a clear theoretical rationale, but it is an assumption they made nonetheless. In all other respects, Model 2 is like Model 1. For example, Model 2 is also a true false-memory model because it assumes that the high false-alarm rate to critical lures occurs because of their high strength.

Model 2 can be illustrated by setting P_i (strength due to presentation on a list) to 1.0 for both related and unrelated items (as before) and to 0 for critical items. The same A_i values that were used to illustrate Model 1 apply here as well, such that critical lures acquire the most associative strength (2 units, on average), related lures the next most (1 unit), and unrelated lures the least (0 units). Because item strength derives from the sum of direct (P_i) and indirect (A_i) effects, the mean strength values for each item type are equal to the following:

Critical target: $S_{CT} = 0 + 2 = 2$
Related target: $S_{RT} = 1 + 1 = 2$
Unrelated target: $S_{UT} = 1 + 0 = 1$

Critical lure: $S_{CL} = 0 + 2 = 2$
Related lure: $S_{RL} = 0 + 1 = 1$
Unrelated lure: $S_{UL} = 0 + 0 = 0.$

Except for the value of S_{CT} (the strength of critical targets), which now equals S_{CL} (the strength of critical lures), this model is identical to Model 1. Why critical items would not benefit from being presented on a list, and why such a result would be regarded as the hallmark of false memory, was not specified. None of the many theories considered by Roediger and McDermott in any of their writings seem to us to correspond to Model 2, and the theory they always consider first and foremost (namely, Underwood's, 1965) certainly does not. Still, this is an idealized version of the false-memory model that Miller and Wolford (1999) set out to disprove. That model was rejected in favor of a criterion-shift model, which we describe next.

Model 3 (Miller and Wolford's, 1999, Criterion-Shift Model)

In the model ultimately favored by Miller and Wolford (1999; Model 3), associative activation is not assumed to influence item strength per se but instead influences the placement of the decision

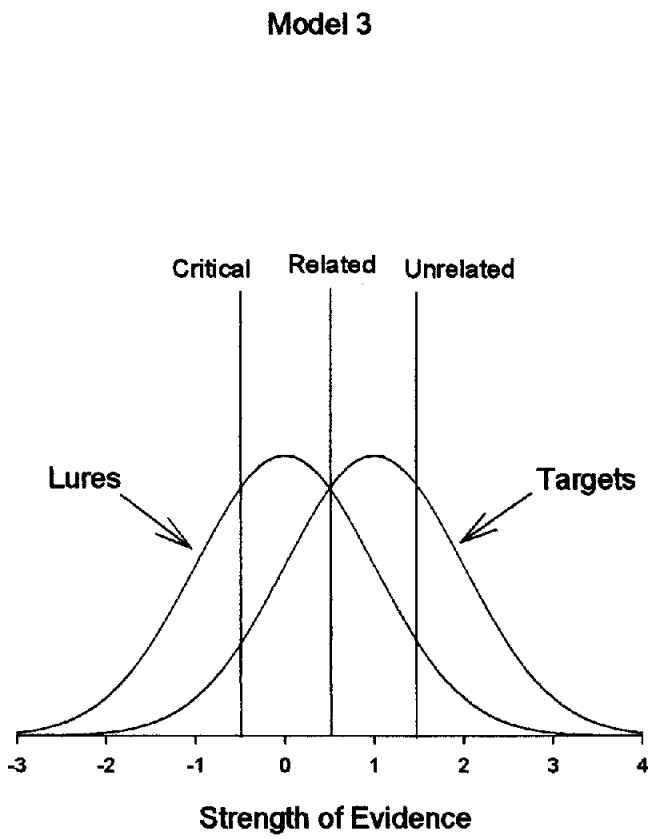

Figure 3. Hypothetical target and lure distributions and three decision criteria (one each for related, unrelated, and critical items) according to the criterion-shift account of Miller and Wolford (1999).

criterion. Strength is derived solely from the effects of study and is assumed to be equal for all item types. Thus, in this model,

$$S_i = P_i,$$

and, for purposes of illustration, P_i is set to 1.0 for all three item types. Thus:

Critical target:	$S_{CT} = 1$
Related target:	$S_{RT} = 1$
Unrelated target:	$S_{UT} = 1$
Critical lure:	$S_{CL} = 0$
Related lure:	$S_{RL} = 0$
Unrelated lure:	$S_{UL} = 0.$

In other words, the strength of the critical lures is no higher than that of the related and unrelated lures. Why, then, are participants so likely to declare these items as being old? Because, according to this model, different decision criteria are used for the different item types. The locations of these criteria are theoretically determined by the associative properties of the test item. Figure 3 provides an illustration of this model. Critical items are recognized as being strongly associated with the list items, so participants assume there is a high probability that these items appeared on the list. This assumption is reflected in the model by the liberal placement of the decision criterion for critical items (placed at −0.5 in this example); unrelated items are not associated with the list items and are therefore generally assumed not to have appeared on the list. In other words, a conservative criterion is used for these items (placed at 1.5). An intermediate criterion is used for the related items (placed at 0.5). Thus, as Miller and Wolford (1999) saw it, the high false-alarm rate to critical lures does not represent false memory. Instead, they said, "The similarity of the values of false alarms on critical lures to hits on related items does not mean that equivalent memories were created. It means that critical lures yielded a lower criterion than related items" (p. 401).

Note that this model assumes that many criterion shifts occur during the course of a recognition test because the different item types are randomly intermixed. In fact, on their 72-item recognition test, which probably lasted no more than a few minutes, participants would, on average, have had to adjust their decision criterion more than 40 times.

Model Comparison

Sensitivity and Bias Parameters

Miller and Wolford (1999) built their case against a false memory interpretation of data from the Roediger–McDermott (Roediger & McDermott, 1995) paradigm primarily on the basis of signal-detection sensitivity and bias measures (both of which are computed using obtained hit and false-alarm rates). Sensitivity (e.g., d') refers to the distance in standard deviation units between the target and lure distributions, whereas bias refers to the location of the decision criterion relative to the point of intersection between the target and lure distributions. If the criterion falls exactly at the point of intersection, responding is said to be unbiased (and the relevant bias measure is equal to zero). If it falls to the left of the intersection, the bias measure is negative, and responding is said to be liberal, and if it falls to the right, the bias measure is positive, and responding is said to be conservative. Signal-detection bias measures like c, c2, and $\log(\beta)$ reflect these relative placements of the criterion.

Using signal-detection measures like these, Miller and Wolford (1999) convincingly demonstrated that Model 2 is untenable because sensitivity for critical items was *not* close to zero. Instead, the obtained d' values were greater than 1.0 even for the critical items. Having ruled out Model 2 on that basis, Miller and Wolford then noted that the obtained measures of bias changed across conditions in the manner predicted by Model 3. More specifically, the bias measures suggested that, for critical items, the criterion was placed far to the left of the point of intersection between the target and lure distributions (c2 = −1.19). For related items, the criterion was also placed to the left of the intersection, but to a less extreme degree (c2 = −0.35). Finally, for unrelated items, the criterion was placed to the right of the intersection between the target and lure distributions (c2 = 0.42). For the idealized version of Model 3 depicted in Figure 3, bias for the critical, related, and unrelated items is −1.0, 0, and 1.0, respectively (and d' = 1.0 for all three item types). Thus, both the d' and bias parameters predicted by Model 3 correspond to the pattern observed in the actual data. On that basis, Miller and Wolford arrived at the following conclusion:

> We have shown that the nonpresented critical lures in the Roediger and McDermott [1995] paradigm do not behave as if they had been

> presented. Performance is significantly higher on every measure when the critical lures are presented compared to when they are not. Further, we have shown that the performance on critical lures is more consistent with a criterion shift than with a change in sensitivity. (p. 402)

However, the findings reported by Miller and Wolford (1999) actually have no bearing on whether the high false-alarm rate to critical lures reflects a criterion shift or not for several reasons. First, and most importantly, the appropriate comparison is between Model 1 and Model 3, not between Model 2 (an implausible model that no one advocates) and Model 3. For both Model 1 and Model 3, d' for all three item types is 1.0, and the bias measures are −1.0, 0, and 1.0 for the critical, related, and unrelated items, respectively. Thus, the predictions of both models correspond to the pattern observed in the data. In Model 3, the different bias measures arise because of a criterion shift. In Model 1, they arise because the distributions themselves shift relative to a fixed criterion. The fact that the signal-detection parameters are identical for these fundamentally different models shows that the computation of those parameters will not differentiate between a criterion-shift account (Model 3) and a false-memory account (Model 1).

A second consideration is that one of the main pieces of evidence presented in support of the criterion-shift account, namely, the changes in measured bias, would have been observed even if d' for the critical items had turned out to be zero (as Model 2 requires). Miller and Wolford (1999) found, for example, that the hit rate to critical targets was .97 while the false-alarm rate was .81. However, imagine that the hit rate for critical targets had turned out to be only .81 (such that d' for critical items equaled 0). Under these conditions, the measured bias (c2) would be −.76, which still seems to suggest a substantial change in the location of the critical-item criterion relative to the related and unrelated items (−0.35 and 0.42, respectively). Thus, the reported bias measures could have been taken to support a criterion-shift account of false memory no matter how the experiment had turned out.

Although they were mainly focused on refuting Model 2, Miller and Wolford (1999) acknowledged that another fixed-criterion model was consistent with their results, and their entire description of this alternative model consisted of these two sentences:

> The reviewer proposed that the presence of the related list context could shift both the signal and the noise distributions for critical lures and shift them more than the corresponding distributions for the related items. The signal and noise distributions for the related items also could be shifted relative to the unrelated items, but not as much as the critical lures. (p. 403)

This, of course, is a brief description of Model 1.

Although Miller and Wolford (1999) did not notice that Model 1 is the false-memory model that best reflects the views of Roediger and McDermott (1995), they did argue that the bulk of evidence weighs against this model as well. How is that possible when even they agree that the only experiments they performed do not provide a differential test? They offered four post hoc arguments in the General Discussion section of their article (each of which is briefly considered below), but they did not consider research that directly investigates whether participants rely on metaknowledge to adjust the decision criterion on an item-by-item basis during the course of a recognition test (research that is also considered below). Much of that research was in press when Miller and Wolford wrote their comment, so their failure to consider it is not surprising. Still, once that literature is reviewed, Model 1 emerges as being the much stronger candidate.

Four Post Hoc Arguments Advanced Against Model 1

The first argument was based on an analysis of recall data. In the procedure used by Miller and Wolford (1999) and by Roediger and McDermott (1995), participants studied several lists and were given a test of free recall following half of those lists (recognition was actually tested after all of the lists had been presented).

Like McDermott (1997), but unlike Roediger and McDermott (1995), Miller and Wolford (1999) also presented some lists that contained the critical item. This allowed them to compute the probability of falsely recalling a critical item both before and after the participant had experienced another list in which a critical item had appeared. The probability of falsely recalling a critical item before seeing a list in which a critical item appeared was .68. The probability of falsely recalling a critical item after experiencing another list in which a critical item appeared was only .32. Why the difference?[3] Miller and Wolford argued that this occurred because participants realized just how "strong" a critical item can be once they saw it on a list. Subsequent to that realization, they adopted a more conservative criterion for critical items, which would reduce the probability of false recall if participants were using a generate–recognize recall strategy. In their words, "we are not suggesting that they recognize a critical lure as a critical lure but once they experience the very high strength of a presented critical lure, their standard or criterion changes for evaluating other items" (Miller & Wolford, 1999, p. 403).

Presumably, the "strength" referred to in this sentence refers to strength of association, not memory strength, because Model 3 (their preferred model) does not assume that critical items produce stronger memories than the other items. In any case, even if the observed effect on false recall is real, which is questionable, and even if it reflects a criterion shift, which is also questionable, those facts would not appear to offer much support for Model 3. Model 3 assumes that, under ordinary circumstances, participants believe that critical lures appeared on the list because those lures are so closely related to the list items. Given that, it seems odd to argue that once a critical item actually does appear on the list participants become *less* likely to harbor that belief. If anything, the appearance of a critical item on a list should reinforce the original belief and lead participants to become even more liberal with respect to critical items than they already are. Instead, in the face of direct evidence suggesting that their belief is true, participants are assumed to become less confident that a critical item appeared on the list (which, in turn, leads them to respond in a more conservative way with respect to those items).

The second post hoc argument advanced in support of Model 3 is similar to the first. Specifically, in the experiments performed by Miller and Wolford (1999), the false-alarm rate to critical lures did not quite equal the hit rate for related items, whereas it did in the Roediger and McDermott (1995) study, possibly because the par-

[3] Actually, no post hoc statistical analyses of this effect were reported, and the relevant data were not presented separately for the two experiments, so it is not clear how seriously to take this finding in the first place.

ticipants were being a bit more conservative when responding to critical items after having experienced an item like that on a previous list. Because this argument is similar to the first, the counterargument is similar as well: The effect may not even be real, and, even if it is, it may not reflect a criterion shift. Even if it does, it is a criterion shift in a direction opposite to what one would expect given the assumptions on which Model 3 is based.

The third argument advanced in favor of Model 3 is that, if Model 1 is correct, it would be a remarkable coincidence that the d' (sensitivity) values for three item types turned out to be identical. This is a curious reason to favor Model 3 over Model 1, because both models make the identical assumption in this regard. That is, the simplest versions of both models assume that d' is the same for all three item types, and this is captured by setting P_i equal to 1 for all item types in the previous examples above (the two models differ only in what the effects of A_i are assumed to be). Thus, the equality of d' across item types obviously cannot be used to discriminate between the two models. Ironically, Miller and Wolford (1999; but not necessarily Roediger and McDermott, 1995) assumed that participants notice the unique characteristics of critical items when they appear on a list. If anything is remarkable, it is that these items, as conspicuous as they are, would nevertheless be processed exactly like all of the other items (thereby yielding the same d').

Finally, Miller and Wolford (1999) argued that it is standard policy to interpret a change in measured bias as a change in the participant's decision strategy. This, of course, is not a reason for preferring Model 3. Further, as indicated earlier, it probably reflects a common logical error. Although it is true that, given the assumptions of signal-detection theory, a change in decision strategy entails a change in measured bias, it does not follow that a change in measured bias entails a change in the participant's decision strategy. Models 1 and 3 are equally compatible with the observed changes in measured bias, and only one assumes an item-by-item criterion shift.

Research on Criterion Shifts Within a Recognition Test

Miller and Wolford (1999) assumed that participants use metaknowledge of an item's associative characteristics to adjust the criterion repeatedly throughout the course of the recognition test. Stretch and Wixted (1998) conducted several experiments that were specifically designed to induce participants to adjust the criterion in that way based on another kind of metaknowledge (specifically, knowledge about item strength). In a preliminary experiment, strength was manipulated between lists. Words on the study list were presented three times each in the strong condition and once each in the weak condition. Obviously, overall recognition performance was better following the strong list than it was following the weak list. Moreover, as is typically the case, a mirror effect was observed. That is, not only was the hit rate significantly higher in the strong condition, but the false-alarm rate was significantly lower as well.

The lures in the weak and strong conditions were physically identical (on average) because, in both cases, they consisted of words drawn randomly from the word pool. Thus, the lower false-alarm rate associated with the strong condition presumably arose because participants, quite reasonably, used a high criterion following a strong list and a low criterion following a weaker list. More specifically, following the strong list, participants presumably knew that any item that appeared on the list would generate a strong sense of prior occurrence, so a high criterion could be used to avoid making false alarms without missing many targets. Following the weak list, by contrast, participants presumably realized that even items that appeared on the list might not seem terribly familiar, so a lower criterion would be called for. By adjusting the criterion in this way (as depicted in the upper panel of Figure 4), participants could maximize the probability of giving a correct answer. The results of this experiment merely suggest that participants can indeed rely on metaknowledge (in this case, metaknowledge of list strength) to set the location of the decision criterion.

In two additional experiments reported by Stretch and Wixted (1998) that are especially relevant to the issue of item-by-item criterion shifts (the kind of criterion shift envisioned by Miller and Wolford's, 1999, criterion-shift model), item strength was conspicuously manipulated within lists rather than between lists. In each list, half the words were colored red, and they appeared five times each. The other half were colored blue and appeared only once each. On the subsequent recognition test, the red and blue targets were randomly intermixed with red and blue lures. Except for color, the red and blue lures were physically identical (in both cases, they were simply words randomly drawn from the word pool that had not appeared on the study list). Any difference in false-alarm rates to the red and blue lures would therefore be strong evidence of a criterion shift. When faced with a red item on the recognition test, participants could easily use their knowledge that if this red item had appeared on the list it would generate a strong sense of prior occurrence (having appeared five times). Thus, a strict criterion could be used to avoid making false alarms without missing very many of the red targets. When faced with a blue item, participants could similarly use their knowledge that a blue target might not seem very familiar because it would have appeared only once. Thus, a lower placement of the criterion would be necessary to avoid missing too many of the blue targets. If participants shifted their criterion on an item-by-item basis in this way (strict if the item is red, back to liberal if the item is blue), then the blue items would be associated with a higher false-alarm rate than the red items.

Note that it would be very easy for participants to adjust the criterion in this way were they so inclined. The participants were fully aware of the fact that the red items were better encoded than the blue items (every single participant was able to report that this was true), and there is no doubt they could use a color-specific criterion if explicitly instructed to do so. However, when strength was manipulated within list, no evidence for item-by-item criterion shifts emerged. Instead, as shown in Table 1, the false-alarm rates for the red and blue items were nearly identical in both experiments (the slight false-alarm rate differences in the two conditions did not approach significance in either case). By contrast, the hit rate for the strong red items far exceeded the hit rate for the weak blue items. The lower panel of Figure 4 illustrates the underlying theoretics of the within-list situation.

Table 1 also shows that measured bias (c in this case) differed for the strong and weak items just as it did for the different item classes in the Roediger–McDermott (Roediger & McDermott, 1995) paradigm. However, a criterion-shift interpretation of these findings, although technically possible, would be quite convoluted. Specifically, one would have to argue that the lure distributions

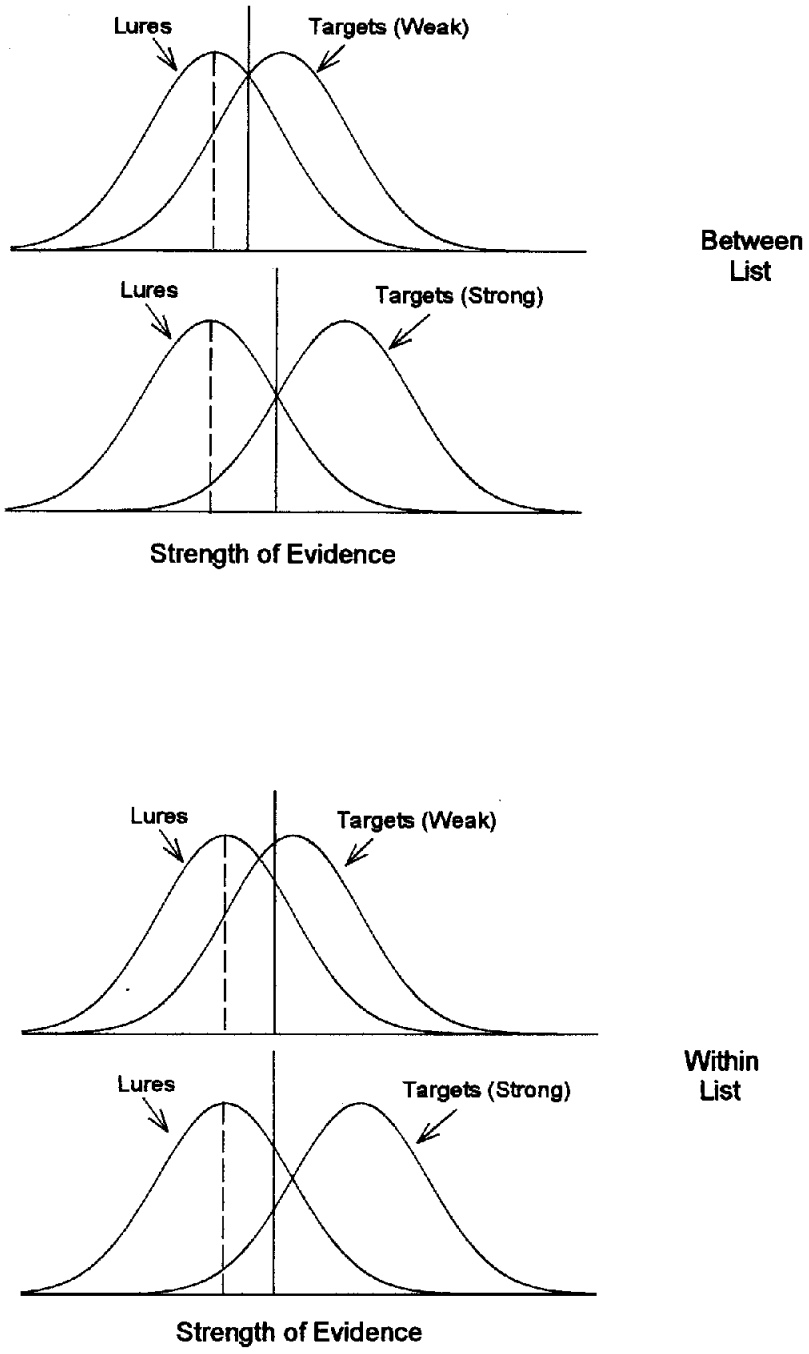

Figure 4. Upper panel: Hypothetical target and lure distributions for the strong and weak conditions of a between-list strength manipulation. *Lower panel:* Hypothetical target and lure distributions for the strong and weak conditions of a within-list strength manipulation. The dashed vertical line shows the mean of the lure distributions in both panels.

were associated with different average levels of strength (in spite of the lures being physically identical except for color) and that participants adjusted their decision criteria nonoptimally and in just such a way as to maintain equal false-alarm rates. A much more straightforward interpretation holds that participants were simply disinclined to adjust the decision criterion on an item-by-item basis (cf. Wixted, 1992).

Why is evidence suggesting that participants do not adjust the decision criterion item by item on the basis of strength at all relevant to Miller and Wolford's (1999) notion that participants *do* adjust the decision criterion item by item on the basis of category membership? Because the evidence suggests that participants are reluctant to shift the decision criterion item by item, not that they are reluctant to shift the criterion on the basis of strength per se.

Table 1
Hit Rates (Hits), False-Alarm Rates (FA), Sensitivity (d'), and Bias (c) From Experiments 4 and 5 of Stretch and Wixted (1998)

Condition	Hits	FA	*d'*	c
		Experiment 4		
Weak (blue)	.58	.18	1.24	0.360
Strong (red)	.82	.20	2.01	−0.030
		Experiment 5		
Weak (blue)	.74	.15	1.88	0.205
Strong (red)	.92	.13	2.68	0.025

Without the between-lists data, the within-lists data could be taken to suggest that strength is simply not the kind of metaknowledge variable that participants rely on to adjust the location of the decision criterion. However, participants understand perfectly well that a criterion shift based on strength makes sense, and they are quite willing to adjust the criterion on that basis so long as they do not have to do so repeatedly throughout the course of the recognition test. Thus, the overall pattern of results suggests a simple principle that may be relevant to the false-memory debate: Participants appear to readily rely on metaknowledge to adjust the criterion between but not within lists.

Why are participants willing to adjust the decision criterion between lists but not within lists? The key difference may be that, in the between-lists case, participants need only set the decision criterion twice, once following the strong list and once again following the weak list. In the within-lists case, by contrast, dozens of criterion shifts would be required throughout the course of a single recognition test. That kind of mental effort, which involves assessing the status of each item and adjusting the criterion accordingly, may be something participants are generally unwilling to exert even when they understand that they should and even when the different conditions are extremely easy to discriminate (e.g., strong red items vs. weak blue items). In spite of this, Miller and Wolford (1999) assumed that participants make a much more subtle discrimination (e.g., between a related item like *thimble* and a critical item like *needle*) and adjust the criterion to an enormous degree (increasing the false-alarm rate from .36 for related items to .81 for critical items). That participants would put forth the effort to make such a fine discrimination and then adjust the criterion in such a dramatic way would be surprising in light of how reluctant participants are to adjust the criterion even a little under much more obvious conditions.

Admittedly, it is within the realm of possibility that participants are willing to exert the mental effort required to adjust the criterion item by item on the basis of category membership even though they are not inclined to use strength in that way (in spite of knowing that they should). Why that would be is not clear, but our data cannot be taken to completely rule out the possibility. Moreover, there are other experimental arrangements that have been interpreted by some as inducing criterion shifts on an item-by-item basis (e.g., experiments on the "Revelation Effect"; Hicks & Marsh, 1998). Thus, although Miller and Wolford (1999) offered no convincing evidence in favor of a criterion-shift account, the question of whether a criterion shift plays any role in the Roediger–McDermott (Roediger & McDermott, 1995) procedure will require more research before it can be definitively answered.

Conclusion

Our point is not that Miller and Wolford (1999) were definitely wrong. Instead, our point is that the new evidence they presented in their article provides absolutely no support for their criterion-shift account and that, on balance, the prevailing evidence weighs against it. Model 1, the false-memory model that corresponds most closely to the views expressed by Roediger and McDermott (1995), is perfectly compatible with the data reported by Miller and Wolford and with other evidence suggesting that participants are not inclined to engage in item-by-item criterion shifts. Thus, in the absence of compelling evidence to the contrary, Model 1 should probably be regarded as the most viable account.

References

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58,* 17–22.

Hicks, J. L., & Marsh, R. L. (1998). A decrement-to-familiarity interpretation of the revelation effect from forced-choice tests of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1105–1120.

McDermott, K. B. (1997). Priming on perceptual implicit memory tests can be achieved through presentation of associates. *Psychonomic Bulletin and Review, 4,* 582–586.

Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review, 106,* 398–405.

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814.

Roediger, H. L., III, & McDermott, K. B. (1999). False alarms about false memories. *Psychological Review, 106,* 406–410.

Roediger, H. L., III, McDermott, K. B., & Robinson, K. J. (1998). The role of associative processes in creating false memories. In M. A. Conway, S. E. Gathercole, & C. Cornoldi (Eds.), *Theories of memory* (Vol. II, pp. 187–245). East Sussex, United Kingdom: Psychology Press.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1379–1396.

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology, 70,* 122–129.

Wickens, T. D., & Hirshman, E. (2000). False memories and statistical decision theory: Comment on Miller and Wolford (1999) and Roediger and McDermott (1999). *Psychological Review, 107,* 377–383.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 681–690.

Received December 31, 1998
Revision received May 24, 1999
Accepted June 1, 1999 ■