Psychonomic Bulletin & Review
2002, 9 (3), 590-596

The reliability of the DRM paradigm as a measure of individual differences in false memories

IRENE V. BLAIR, ALISON P. LENTON, and REID HASTIE
University of Colorado, Boulder, Colorado

Despite considerable research on the Deese–Roediger–McDermott (DRM) false memory paradigm, little attention has been paid to the reliability of the paradigm as a measure of individual differences. In the present research, we examined the reliability of the DRM paradigm in a 2-week test–retest design. This analysis showed that the false memories produced in the paradigm were quite stable across the 2-week period and that this stability had both global (cross-list) and list-specific components. In contrast, correct memories showed only global stability across the testing period.

Research on memory errors has grown tremendously since Roediger and McDermott (1995) reintroduced Deese's (1959) false memory paradigm (for reviews, see Roediger, McDermott, & Robinson, 1998; Schacter, 1999). Indeed, Bruce and Winograd (1998) argue that the scientific zeitgeist of the late 20th century was such that the procedures originally developed by Deese "had to be discovered" (p. 622). With societal concerns about the validity of recovered memories inevitably came scientific interest in memory errors and illusions.

The Deese–Roediger–McDermott (DRM) paradigm is characterized, most generally, by presenting participants with a list of words, all associated with a particular concept (e.g., *sleep*), and then asking the participants to perform either a free recall or a recognition memory test. The basic finding is that rates of intrusions or false alarms (FAs) are much higher for unstudied words that are associated than for words unassociated with the target concept. Indeed, these associative false memories are often as strong as true memories for the presented words, with participants claiming to remember the associated words, and not just know that they were presented (Roediger & McDermott, 1995).

Since Roediger and McDermott (1995) reported their findings, research on the DRM paradigm has focused on the task conditions that moderate the effect, such as study instructions, presentation format, and testing repetition and delay (e.g., Gallo, Roberts, & Seamon, 1997; McDermott, 1996; McDermott & Roediger, 1998; Payne, Elie, Blackwell, & Neuschatz, 1996), and on the characteristics of the false memories produced in the paradigm, such as participants' "memory" for the speaker, the sound of the words, the feelings they had when they "studied" the associated words, and neural correlates of the false memories (e.g., Mather, Henkel, & Johnson, 1997; Norman & Schacter, 1997; Payne et al., 1996; Roediger & McDermott, 1995; Schacter, Verfaillie, & Pradere, 1996). In comparison with that extensive body of research, there has been little investigation of the reliability of the DRM paradigm as a measure of individual differences in false memories. Such research is important if false memories are to be studied from an individual-difference perspective.

Are some people more likely to experience false memories? If so, how stable is that tendency? The first question has begun to be answered with research showing that some groups of people—such as the elderly, people with Alzheimer's dementia, and those who report more dissociative experiences—are more susceptible to the false memories produced in the DRM paradigm (e.g., Balota et al., 1999; Clancy, Schacter, McNally, & Pitman, 2000; Kensinger & Schacter, 1999; Norman & Schacter, 1997; Winograd, Peluso, & Glover, 1998). The second question, however, has yet to be answered. There is currently little evidence that the DRM paradigm produces false memories in a manner that reveals *stable* individual differences. If individual differences in false memories cannot be reliably measured, individual differences in the predictors, correlates, or consequences of false memories cannot be meaningfully studied. Evidence for the reliability of the DRM paradigm would show that individuals do vary in their propensity for false memories and that individual variance is stable over time. With that knowledge, researchers would be encouraged to study the participant variables that may influence or predict false memories.

Up to this point, we have referred to the reliability of the DRM paradigm and the false memories it generates in relatively global terms. That is, existing individual-difference research has generally assumed that some people are more likely to have false memories for list-related items and

This research was supported by an Implementation of Multicultural Perspectives and Approaches in Research and Teaching (IMPART) grant from the University of Colorado, by NIH Grant MH63372, awarded to the first author, and by NSF Grant SBR9816458, awarded to the third author. We thank Kenneth Norman and researchers at the University of Colorado Stereotyping and Prejudice (CUSP) laboratory for their comments on an earlier draft of this article. Correspondence concerning this article should be addressed to I. V. Blair, University of Colorado, Department of Psychology, Muenzinger Psychology Building, Boulder, CO 80309-0345 (e-mail: irene.blair@colorado.edu).

other people are less likely to do so, regardless of the specific concepts involved. Such a global tendency for false memories is obviously an important problem, and determining the reliability of such a tendency is critical for a complete understanding of it. However, individual differences in false memories may arise from specific, as well as global, tendencies. Some people may be more likely to have false memories in response to certain list associates (e.g., the *fruit* list), whereas other people may be more likely to have false memories in response to other list associates (e.g., the *sleep* list). Although this specific tendency for false memories has not been directly studied, related research has shown that people vary in the extent to which they are schematic or chronic for certain concepts (e.g., *kindness*, *honesty*) and that schematicity facilitates memory and judgment of related attributes (Bargh, Bond, Lombardi, & Tota, 1986; Bargh & Thein, 1985; Higgins & King, 1981; Higgins, King, & Mavin, 1982). We are proposing here that people who are schematic for a concept may be more likely to have false memories specific to that concept and that people are likely to differ reliably in terms of the concepts for which they are schematic.

The primary goal of the present research was to assess the reliability of the DRM paradigm in terms of both global and specific tendencies for false memories. This goal necessarily constrained the research method. Specifically, the reliability of a global tendency for false memories can be assessed in several ways, including the test–retest, alternative-forms, and split-halves methods (Carmines & Zeller, 1979). In contrast, only the test–retest method can be used to determine the reliability of specific tendencies for false memories. Thus, in the present research, we investigated the reliability of the DRM paradigm by giving participants the same memory lists and recognition test in two different sessions, separated by 2 weeks. With this design, the reliability of a *global* tendency for false memories would be demonstrated in two ways, across time and within each test. That is, FAs associated with a list at Time 1 ought to predict FAs associated with different lists at Time 2. And, FAs associated with a list ought to predict FAs associated with different lists within each of the tests (i.e., inter-item reliability). The reliability of a *specific* tendency for false memories would be demonstrated if an FA at Time 1 predicted the same FA at Time 2, over and above any global tendency for false memories.

The second goal of the present research was to examine the reliability of false memories relative to true memories. Several studies have shown that false memories sometimes respond to study and test conditions in the same manner as true memories. For example, Toglia, Neuschatz, and Goodwin (1999) showed that semantic processing or blocked list presentations resulted in higher levels of both correct recall and false memories. Prior testing has also been shown to increase both types of memory (e.g., Brainerd, Reyna, & Brandse, 1995; McDermott, 1996). Other conditions, however, appear to affect true and false memories in an inverse manner. For example, memory for studied items has been shown to worsen with delayed testing, whereas false memories are less likely to fade over time (Brainerd et al., 1995; McDermott, 1996; Payne et al., 1996; Toglia et al., 1999), although this difference may depend on whether memory is measured through recall or recognition (Seamon et al., 2001). In addition, multiple study–test trials over a short period of time have been shown to enhance true memories but to decrease the rate of false memories (Kensinger & Schacter, 1999; McDermott, 1996). The present research differs from prior work by examining the test–retest reliability of true and false memories across a 2-week period of time. As such, the primary issue is not whether the overall levels of correct and false recognition are similar or different, but whether they demonstrate similar or different levels of reliability under the present conditions.

OVERVIEW

The participants were exposed to five word lists, including four lists taken from Roediger and McDermott (1995) and one list with social roles associated with males or females. Following the exposure phase, the participants completed a recognition test that included words from the exposure lists and both list-related and list-unrelated lures. Approximately 2 weeks later, the participants returned to the laboratory and completed the same study–test sequence.

The social role list was included in this study to investigate false memories produced by social associations. Nearly all of the research with the DRM paradigm has studied nonsocial associations, such as associations to *fruit* and *chair*. Indeed, there are only two demonstrations that false memories in the DRM paradigm can be produced by social associations (Bihm & Winer, 1983; Lenton, Blair, & Hastie, 2001). The present research extended that work by examining the test–retest reliability of false memories based on social stereotypes.

METHOD

Participants

Fifty-nine students at the University of Colorado participated in partial fulfillment of a course requirement. Approximately 64% of the participants were female, and 86% were White.

Materials

The participants were presented with five lists, each composed of 15 words associated with a specific concept. Four of the lists were taken from those used by Roediger and McDermott (1995), and they contained associates of the concepts *chair*, *fruit*, *window*, and *sleep*.[1] According to norms published by Stadler, Roediger, and McDermott (1999), these lists vary in their ability to produce list-related false memories, with the *window*, *sleep*, and *chair* lists among the top performers rated by Stadler et al. and the *fruit* list one of the worst. Thus, these lists provide some generality to the tests.

The participants also received one of two stereotype lists, with one of the lists containing 15 roles associated with the concept *male* (e.g., "brother," "doctor") and the other list containing 15 roles associated with the concept *female* (e.g., "mother," "secretary"). The participants were randomly assigned to receive either the *male* or the *female* list, with all the participants in a session receiving the same stereotype list. Because these two lists produced the same results, we will collapse across them and refer simply to the *stereotype* list.

The recognition test included 10 studied and 31 nonstudied words. Two words from each of the five exposure lists were randomly selected to appear on the test. The 31 nonstudied words (lures) had the following composition: 2 words semantically related to each of the four lists obtained from Roediger and McDermott ("chair" and "arm," "fruit" and "produce," "window" and "light," "sleep" and "pajamas"); 3 roles stereotypically related to the stereotype list ("engineer," "carpenter," and "architect," or "hairdresser," "librarian," and "cashier," depending on list condition), and 20 filler words that were unrelated to any of the lists (e.g., "bulletin," "popular," "cord"). These words were presented in a single pseudorandom order, with the constraint that words from or related to the same list not appear next to each other. The participants were asked to indicate their recognition of each word by checking one of four options: *confident new*, *probably new*, *probably old*, and *confident old*.

Procedure

The participants were presented with the five word lists, one at a time, via audiotape. The 15 words within each list were spoken slowly and clearly by a female speaker, with a 1.5-sec delay between each word. The experimenter stopped the tape after each list had been presented, and the participants were instructed to work on simple multiple-choice arithmetic problems for 2 min before the next list began. After the final list was presented, the participants worked again on the arithmetic problems for 2 min before they received the recognition test. The five lists were presented in the following order: (1) *chair* list, (2) *fruit* list, (3) *stereotype* list, (4) *window* list, and (5) *sleep* list. The participants returned to the laboratory approximately 2 weeks later, at which time they heard the five lists and completed the recognition test a second time. All of the items and the order in which they were presented in both study and test were exactly the same in the two sessions.

RESULTS

Recognition Performance

If a participant correctly identified a studied word as *confident old* or *probably old*, the response was coded as a hit; if a participant incorrectly identified a nonstudied word as *confident old* or *probably old*, the response was coded as an FA. The mean hit and FA rates for each study list are presented in Table 1. Consistent with prior research, the participants produced relatively high hit rates at both Time 1 and Time 2 ($M = .89$ and .86, respectively) and low FA rates to lures unrelated to the study lists ($M = .14$ and .18, respectively). Of greater interest, however, were the rates of list-related FAs. As was expected, the participants were significantly more likely to make FAs to related than to unrelated lures at both Time 1 and Time 2 [$M = .56$ vs. .14 and $M = .57$ vs. .18, respectively[2]; $t(58) = 14.70$, $p < .0001$, and $t(58) = 12.48$, $p < .0001$, respectively]. Indeed, all five lists produced significant levels of list-related FAs in both sessions (all ts > 5.0, $p < .0001$). Unless specified otherwise, all FA rates from this point forward will refer to list-related FAs.

Table 1
Mean Rates of Hits and False Alarms to List-Related Lures at Time 1 and Time 2

	Time 1		Time 2	
List	Hits	False Alarms	Hits	False Alarms
Chair	.89	.61	.88	.63
Fruit	.91	.51	.86	.56
Window	.89	.64	.88	.60
Sleep	.85	.59	.80	.54
Stereotype	.91	.44	.86	.50

Test–Retest Reliability of False Memories

As the first indication that the DRM paradigm produces reliable false memories, the overall level of FAs at Time 1 was highly correlated with the overall level of FAs produced 2 weeks later ($r = .76$, $p < .0001$). To determine whether a global tendency for false memories contributed to that reliability, correlations between FAs at Time 1 and FAs at Time 2 were calculated across the five lists and are presented in the top panel of Table 2. Looking across the rows of the table, one can see that the FA rates for all of the word lists at Time 1 significantly predicted at least one other FA rate at Time 2, sometimes substantially so. Indeed, 50% of the cross-list correlations were statistically significant, as was the average cross-list correlation ($r = .27$, $p < .05$).[3] These correlations indicate that a global tendency to produce false memories was stable across the 2-week testing period. Those participants who were more likely to falsely recognize list-related items at Time 1 were also more likely to falsely recognize list-related items at Time 2, even though the specific items involved were different. Calculations of the interitem reliabilities of the FAs within each session provided additional evidence for the reliability of this global tendency ($\alpha = .61$ and .69 at Time 1 and Time 2, respectively).

To determine the reliability of a more specific tendency to FA to list-related items, multiple regressions were conducted in which the FAs for each list at Time 2 was regressed on *all five* of the Time 1 FAs simultaneously. These regressions allowed us to examine whether the FA rate for a specific Time 1 word list was a significant predictor of the FA rate for that same list at Time 2, even when the global tendency to make FAs, as revealed through the other Time 1 FAs, was controlled. The results of these multiple regressions are presented in the top panel of Table 3, and they show that a specific Time 1 FA was a significant predictor of the same Time 2 FA for four of the five lists: *chair*, *fruit*, *sleep*, and *stereotype* [all ts$(53) > 3.00$, $p < .01$]. Moreover, in all of those cases, the same Time 1 FA accounted for more than twice the variance of any other Time 1 FA (average partial $r = .48$).

Consistent with prior research on the DRM paradigm, we have used the list-related FA as evidence for "false memories." It is reasonable to wonder, however, whether a general tendency (bias) to make FAs to *all* types of lures might explain the reliability of the related FAs. Is it the case that some people simply have a more liberal response style than do others? Can we show that the tendency to make *related* FAs is stable, independent of other response biases? We can answer those questions through analyses of the FA data for the unrelated lures. If people differ in their general response bias, the unrelated FAs ought to be correlated with the related FAs within a particular testing session. This was indeed the case at Time 1 ($r = .35$, $p < .01$), although less so at Time 2 ($r = .22$, n.s.). In light of these correlations, it is important to test whether the ten-

Table 2
Correlations Between Time 1 and Time 2 False Alarms, and Between Time 1 and Time 2 Hits, by List

Time 1 Lists	Time 2 Lists				
	Chair	Fruit	Window	Sleep	Stereotype
			False Alarms		
Chair	.52***	.10	.24	.14	.29*
Fruit	.03	.62***	.39**	.17	.35**
Window	.15	.42**	.40**	.16	.44***
Sleep	.20	.35**	.31*	.47***	.42***
Stereotype	.17	.27*	.24	.38**	.56***
			Hits		
Chair	.21	.26*	.32*	.07	.28
Fruit	.30*	.52***	.41**	.26*	.18
Window	.34**	.38**	.33*	.05	−.03
Sleep	.30*	.04	.16	.20	.12
Stereotype	.18	.13	.30*	.12	.14

*$p < .05$. **$p < .01$. ***$p < .001$.

dency to make related FAs is stable, independent of the influence of general response biases. One way to do that is to compute *d´*, treating the related FAs as "hits" and the unrelated FAs as "FAs" (Seamon et al., 2001). Overall, this false memory *d´* at Time 1 was significantly correlated with the false memory *d´* at Time 2 ($r = .49, p < .0001$). Moreover, this reliable tendency for false memories had both global and specific components. A global tendency for false memories with this measure was shown through a number of significant cross-list correlations (40%) and adequate inter-item reliabilities ($\alpha = .58$ and .76 at Time 1 and Time 2, respectively). A specific tendency for false memories was observed in the results of multiple regressions with the *d´* measure. That is, a specific Time 1 false memory continued to be a significant predictor of the same Time 2 false memory for the four lists noted earlier [all ts(53) > 2.25, $p < .05$; average partial $r = .45$]. Thus, the reliability in false memories cannot be explained by a general response bias.

Strong Versus Weak Lures

For the four lists obtained from Roediger and McDermott (1995; *chair*, *fruit*, *window*, and *sleep*), a strong and a (relatively) weak lure can be identified, allowing for separate analyses of their reliabilities. For the strong lures, good test–retest reliability was observed in the correlation between the overall levels of FAs at Time 1 and at Time 2 ($r = .62, p < .0001$). Evidence for the reliability of both global and specific tendencies to FAs to these lures was also obtained. Significant cross-list correlations from Time 1 to Time 2 for all four lists provided evidence for the reliability of the global tendency across time. Fifty-eight percent of the cross-list correlations were significant, and the average cross-list correlation was also significant ($r = .30, p < .025$). Moreover, Chronbach's alphas of .67 and .79 at Time 1 and Time 2, respectively, provided evidence for the reliability of this tendency within each session. The reliability of a specific tendency to FAs to strong lures was demonstrated through the multiple regressions. Specifically, the Time 1 FAs were significant predictors of the same Time 2 FAs, while controlling for the other Time 1 FAs, for three of the four lists [*chair*, *fruit*, and *sleep*; all ts > 2.5, $p < .05$; average partial $r = .42$].

For the weak lures, the correlation between the overall levels of FAs at Time 1 and at Time 2 was somewhat lower ($r = .43, p < .001$). This lower reliability estimate appeared to be due to the unreliability of the global tendency to make FAs to weak lures. Specifically, there was only one significant Time 1 to Time 2 cross-list correlation (8% of all cross-list correlations), and the average cross-list correlation was small and not significant ($r = .07$). Furthermore, the interitem reliabilities within each session were low ($\alpha = .39$ and .50 at Time 1 and Time 2, respectively). In contrast, the reliability of the specific tendency to make FAs to weak lures was relatively high. The Time 1 FAs were significant predictors of the same Time 2 FAs, while controlling for the other Time 1 FAs, for all four of the lists (all ts > 2.5, $p < .05$; average partial $r = .44$).

Table 3
Multiple Regression Analyses in Which Each Time 2 False Alarm or Hit Rate Was Regressed on All of the Time 1 False Alarm or Hit Rates Simultaneously

Time 1 Predictors	Time 2 Outcomes									
	Chair		Fruit		Window		Sleep		Stereotype	
	B	r^2	*B*	r^2	*B*	r^2	*B*	r^2	*B*	r^2
					False Alarms					
Chair	**.54*****	**.26**	−.02	.00	.16	.03	.13	.01	.24	.05
Fruit	−.02	.00	**.57*****	**.32**	.22*	.09	.07	.01	.17	.04
Window	−.09	.01	.25	.04	**.20**	**.04**	−.18	.02	.15	.02
Sleep	.15	.04	.26	.07	.16	.04	**.43****	**.17**	.26*	.07
Stereotype	.05	.00	.00	.00	.00	.00	.24*	.08	**.38****	**.18**
					Hits					
Chair	**.03**	**.00**	.07	.00	.15	.02	−.09	.00	.30	.05
Fruit	.28	.06	**.65*****	**.25**	.36*	.09	.38	.06	.12	.01
Window	.35*	.10	.46**	.15	**.29**	**.07**	.02	.00	−.14	.01
Sleep	.29*	.08	.02	.00	.10	.01	**.27**	**.04**	.09	.00
Stereotype	−.04	.00	−.18	.02	.11	.00	.00	.00	**.06**	**.00**

Notes—*B* = unstandardized regression coefficient, r^2 = squared partial correlation coefficient. List matches from Time 1 to Time 2 are in boldface. *$p < .05$. **$p < .01$. ***$p < .001$.

Relative Test–Retest Reliability of True and False Memories

Similar to false memories, individual differences for true memories were relatively stable across the 2-week time period, as was indicated by a substantial, albeit smaller, correlation between the overall level of hits at Time 1 and Time 2 ($r = .51$, $p < .0001$). Correlations among the hits for the five lists showed that the hits for all of the word lists at Time 1 significantly predicted at least one other hit rate at Time 2 (see the bottom panel of Table 2). These multiple cross-list correlations suggest a reliable global tendency to make hits. However, this tendency was slightly weaker than what was observed for the FAs. Fewer (45%) of the cross-list correlations were significant, and the average cross-list correlation was somewhat lower ($r = .21$, n.s.). In addition, the interitem reliability of hits at Time 1 ($\alpha = .56$) was somewhat lower, although the Time 2 interitem reliability was somewhat higher ($\alpha = .77$).

In stark contrast to false memories, however, there was very little evidence for reliability in the recognition of specific studied items. Multiple regressions predicting each Time 2 hit rate with all five Time 1 hit rates showed that only one of the five Time 1 hits significantly predicted the same Time 2 hits, when the general tendency to make hits on other Time 1 items was controlled (see the bottom panel of Table 3). In all four of the other lists, other Time 1 hits accounted for more of the Time 2 variance than did the same Time 1 hits. Thus, although false and true memories both demonstrate reliability in global tendencies, only false memories show reliability in a specific tendency.[4]

Comparisons between FAs and hits both across and within the testing sessions showed that the tendency to make FAs was unrelated to the tendency to make hits. Overall, the participants' FAs at Time 1 were uncorrelated with their hits at Time 2, and similarly, their hits at Time 1 were uncorrelated with their FAs at Time 2 ($r = .20$ and .15, respectively, both n.s.). Correlations between FAs and hits across the five lists showed that only 4 of the 60 (7%) Time 1 to Time 2 correlations were significant (average $r = .07$, n.s.). Moreover, none of the correlations between FAs and hits within the Time 1 session was significant (average $r = .07$, n.s.), and only 2 of the 25 (8%) correlations between FAs and hits within the Time 2 session were significant (average $r = .12$, n.s.). These null findings confirm other research that has shown that hit and FA rates produced in the DRM paradigm are generally unrelated (e.g., Stadler et al., 1999).[5]

DISCUSSION

Research with the DRM paradigm has shown repeatedly that semantic associations produce robust false memories. The present work provides the first test of the reliability of the paradigm as an individual-difference measure. The results of this test showed that participants' rates of FAs to list-related items remained remarkably stable over a 2-week period of time, and this reliability was based on both global and specific tendencies for false memories.

As has been suggested by prior research on individual differences (e.g., Balota et al., 1999; Clancy et al., 2000; Kensinger & Schacter, 1999; Norman & Schacter, 1997; Winograd et al., 1998), a reliable global tendency for false memories was evident in significant cross-list correlations between the two testing sessions: Those participants who made more FAs during the first session were more likely to make FAs during the second session, even when the items were different. In addition, calculations of interitem reliability within a session showed that those participants who made more FAs for one list were more likely to make FAs for other lists in that session. What is behind this global tendency for false memories? One possibility is general strategy (i.e., response bias). If some participants tended to use a more liberal response criterion than did others, false memories would appear to be reliable both across time and within session. The present results argue against that explanation. Specifically, even when the participants' tendency to FAs to related lures was corrected for response bias through a d' measure, the tendency for false memories continued to be stable both across and within sessions. In addition, the global tendency to make hits did not correspond to the global tendency to make list-related FAs, either within or between sessions. Thus, a general response bias does not appear to provide a good account for the present findings (for related evidence, see Winograd et al., 1998).

There are at least two potential explanations for the reliability of a global tendency for false memories. First, fuzzy trace theory (FTT) suggests that false memories are the result of extracting *gist* along with verbatim information from the studied material (Reyna & Brainerd, 1995). To the extent that individuals vary in the ease or difficulty with which they can recall specific details of past experiences and their dependence on gist information varies accordingly, reliable individual differences in false memories would be observed. Second, a global tendency for false memories may also arise from a problem in source monitoring—specifically, difficulty in distinguishing between sources of activation that are directly versus indirectly related to past experiences (Johnson, Hashtroudi, & Lindsay, 1993). To the extent that individuals vary in their ability to distinguish among sources of activation, reliable individual differences in false memories would again be observed. Both of these accounts are supported by the finding that a global tendency for false alarms was much more reliable for strong than for weak lures.

Prior research has suggested that older adults and people with Alzheimer's dementia are more likely to rely on gist or to have problems in source monitoring, resulting in higher levels of false memories (e.g., Balota et al., 1999; Kensinger & Schacter, 1999; Norman & Schacter, 1997). The present research suggests that even younger, healthy adults may differ in their reliance on gist or ability to monitor sources of activation and that those differences can be reliably measured.

In addition to providing evidence for the reliability of a global tendency for false memories, the present research

suggests that there may also be a reliable specific tendency for false memories. Incorrectly recognizing "chair" at Time 1, for example, was a far better predictor of making the same error at Time 2, as compared with other Time 1 FAs. One explanation for this finding can be found in the study's methodology. Because the same lists and test were used in both sessions, the participants' memory for specific items on the first test may have influenced their performance on the second test, thereby enhancing the apparent stability of their (specific) memory performance. Although such a testing effect may have had some influence on the results, it does not appear to provide a complete account for them. In particular, specific true memories did not show the same reliability as specific false memories. It is not clear why memory for specific items on the prior test would have increased the reliability of false, but not of true, memories, especially in light of research suggesting that testing effects are more robust for *true* memories than for false ones (see McDermott, 1996). In addition, it is uncertain how much influence a prior test would have in the face of newer, more relevant information. Reexposure to the memory lists just before the second test should have reinstated both true and false memories, which ought to have had a more powerful influence on performance, as compared with the test taken 2 weeks earlier (see, also, Kensinger & Schacter, 1999; McDermott, 1996).

FTT provides another account for the reliability of a specific tendency for false memories, based on the stability of gist. In particular, FTT predicts that specific false memories ought to be more stable across time than are specific true memories, because gist is more stable than verbatim information (Brainerd et al., 1995; Toglia et al., 1999). We note, however, that the research on the persistence of true versus false memories has not revealed consistent differences, especially for recognition memory, which was measured in the present study. Whereas some studies have shown that FAs decline more slowly than hits (Payne et al., 1996; Thapar & McDermott, 2001), other studies have found no difference between FAs and hits (Lampinen & Schwartz, 2000; Neuschatz, Payne, Lampinen, & Toglia, 2001; Seamon et al., 2001), and still others have found that hits are forgotten more slowly than FAs (Brainerd, Wright, Reyna, & Mojardin, 2001). Seamon et al. suggest that the difference in the persistence of false and true memories may be diminished with recognition tests because both the studied and the related words are presented, providing equivalent access to the verbatim and the gist traces. In the present research, this may have been even more likely, because the participants were re-presented with the study words in the second session, which ought to have reinstated both gist and verbatim traces.

A third potential account for the reliability of specific false memories is based on the reliability of implicit associative responses (IARs). In their reintroduction of the false memory paradigm, Roediger and McDermott (1995) suggested that IARs provide a good account for the false memories obtained in the paradigm. The basic idea behind IARs is that nonstudied words can become activated as a consequence of their associations to studied words. During the memory test, that activation results in false memories for those words. Because the strength of an associative network is likely to vary across individuals in a reliable manner (or alternatively, people are schematic for different concepts), an IAR account provides a plausible account for the specific test–retest reliability observed for false memories. Because true memories generally rely less on the strength of associative networks (i.e., the verbatim trace is not dependent on the strength of existing associations), they would not be expected to exhibit the same specificity in test–retest reliability.

Although we cannot say for certain why specific false memories were more stable than specific true memories in the present study, the finding has important practical implications. Others have commented on the problem of using response stability, over repeated questioning, as a criterion for truth in eyewitness testimony (e.g., Brainerd et al., 1995). The present results suggest that the problem may be even more serious, since specific false memories appear to be more stable than specific true memories, even after reexposure to the facts. Additional research is needed to determine whether such a pattern leads to greater confidence in the "reality" of false memories.

In summation, the results of the present study provide strong support for an individual-difference approach to false memories. By showing that the DRM paradigm is a reliable individual-difference measure, researchers ought to be encouraged to use the paradigm to study participant variables in the antecedents, correlates, and consequences of false memories.

REFERENCES

Balota, D. A., Cortese, M. J., Duchek, J. M., Adams, D., Roediger, H. L., III, McDermott, K. B., & Yerys, B. E. (1999). Veridical and false memories in healthy older adults and in dementia of the Alzheimer's type. *Cognitive Neuropsychology*, **16**, 361-384.

Bargh, J. A., Bond, R. N., Lombardi, W. J., & Tota, M. E. (1986). The additive nature of chronic and temporary sources of construct accessibility. *Journal of Personality & Social Psychology*, **50**, 869-878.

Bargh, J. A., & Thein, R. D. (1985). Individual construct accessibility, person memory, and the recall–judgment link: The case of information overload. *Journal of Personality & Social Psychology*, **49**, 1129-1146.

Bihm, E. M., & Winer, J. L. (1983). The distortion of memory for careers: The influence of the thematic organization of occupational information. *Journal of Vocational Behavior*, **23**, 356-366.

Brainerd, C. J., Reyna, V. F., & Brandse, E. (1995). Are children's false memories more persistent than their true memories? *Psychological Science*, **6**, 359-364.

Brainerd, C. J., Wright, R., Reyna, V. F., & Mojardin, A. H. (2001). Conjoint recognition and phantom recollection. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 307-327.

Bruce, D., & Winograd, E. (1998). Remembering Deese's 1959 articles: The Zeitgeist, the sociology of science, and false memories. *Psychonomic Bulletin & Review*, **5**, 615-624.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment* (Sage university paper series on quantitative applications in the social sciences, 07-017). Newbury Park, CA: Sage.

Clancy, S. A., Schacter, D. L., McNally, R. J., & Pitman, R. K. (2000). False recognition in women reporting recovered memories of sexual abuse. *Psychological Science*, **11**, 26-31.

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

GALLO, D. A., ROBERTS, M. J., & SEAMON, J. G. (1997). Remembering words not presented in lists: Can we avoid creating false memories? *Psychonomic Bulletin & Review*, **4**, 271-276.

HIGGINS, E. T., & KING, G. A. (1981). Accessibility of social constructs: Information-processing consequences of individual and contextual variability. In N. Cantor & J. F. Kihlstrom (Eds.), *Personality, cognition, and social interaction* (pp. 69-122). Hillsdale, NJ: Erlbaum.

HIGGINS, E. T., KING, G. A., & MAVIN, G. H. (1982). Individual construct accessibility and subjective impressions and recall. *Journal of Personality & Social Psychology*, **43**, 35-47.

JOHNSON, M. K., HASHTROUDI, S., & LINDSAY, D. S. (1993). Source monitoring. *Psychological Bulletin*, **114**, 3-28.

KENSINGER, E. A., & SCHACTER, D. L. (1999). When true memories suppress false memories: Effects of aging. *Cognitive Neuropsychology*, **16**, 399-415.

LAMPINEN, J. M., & SCHWARTZ, R. M. (2000). The impersistence of false memory persistence. *Memory*, **8**, 393-400.

LENTON, A. P., BLAIR, I. V., & HASTIE, R. (2001). Illusions of gender: Stereotypes evoke false memories. *Journal of Experimental Social Psychology*, **37**, 3-14.

MATHER, M., HENKEL, L. A., & JOHNSON, M. K. (1997). Evaluating characteristics of false memories: Remember/know judgments and memory characteristics questionnaire compared. *Memory & Cognition*, **25**, 826-837.

MCDERMOTT, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory & Language*, **35**, 212-230.

MCDERMOTT, K. B., & ROEDIGER, H. L., III (1998). Attempting to avoid illusory memories: Robust false recognition of associates persists under conditions of explicit warnings and immediate testing. *Journal of Memory & Language*, **39**, 508-520.

NEUSCHATZ, J. S., PAYNE, D. G., LAMPINEN, J. M., & TOGLIA, M. P. (2001). Assessing the effectiveness of warnings and the phenomenological characteristics of false memories. *Memory*, **9**, 53-71.

NORMAN, K. A., & SCHACTER, D. L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition*, **25**, 838-848.

PAYNE, D. G., ELIE, C. J., BLACKWELL, J. M., & NEUSCHATZ, J. S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory & Language*, **35**, 261-285.

REYNA, V. F., & BRAINERD, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning & Individual Differences*, **7**, 1-75.

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

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. Cornolde (Eds.), *Theories of memory* (Vol. 2, pp. 187-245). New York: Psychology Press.

ROEDIGER, H. L., III, WATSON, J. M., MCDERMOTT, K. B., & GALLO, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, **8**, 385-407.

SCHACTER, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist*, **54**, 182-203.

SCHACTER, D. L., VERFAELLIE, M., & PRADERE, D. (1996). The neuropsychology of memory illusions: False recall and recognition in amnesiac patients. *Journal of Memory & Language*, **35**, 319-334.

SEAMON, J. G., LUO, C. R., KOPECKY, J. J., PRICE, C. A., ROTHSCHILD, L., FUNG, N. S., & SCHWARTZ, M. A. (2001). *Are false memories more difficult to forget than accurate memories?* Manuscript submitted for publication.

STADLER, M. A., ROEDIGER, H. L., III, & MCDERMOTT, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition*, **27**, 494-500.

THAPAR, A., & MCDERMOTT, K. B. (2001). False recall and false recognition induced by presentation of associated words: Effects of retention interval and level of processing. *Memory & Cognition*, **29**, 424-432.

TOGLIA, M. P., NEUSCHATZ, J. S., & GOODWIN, K. A. (1999). Recall accuracy and illusory memories: When more is less. *Memory*, **7**, 233-256.

WINOGRAD, E., PELUSO, J. P., & GLOVER, T. A. (1998). Individual differences in susceptibility to memory illusions. *Applied Cognitive Psychology*, **12**, S5-S27.

NOTES

1. To distinguish concepts or word categories from words that appeared in the study, the former are italicized, and the latter are in quotes.

2. When only the strongest or critical lures were examined for the word lists obtained from Roediger and McDermott (1995; i.e., "chair," "fruit," "window," and "sleep"), the high FA rates of .82 and .78 at Time 1 and Time 2, respectively, were comparable to what has been shown in previous research.

3. Because the sampling distribution of nonzero correlations is skewed, the individual correlations were first transformed by Fisher's z and averaged, and then that average z was transformed back to provide an estimate of the average r. All the reports in this paper of an average r took this approach.

4. These analyses were also conducted using d' to measure correct memories. The only difference was that the global reliability component was stronger with this measure: 65% of the Time 1 to Time 2 cross-list correlations were significant, as was the average cross-list correlation ($r = .30$), and the interitem reliabilities were much higher ($\alpha = .79$ and .87 at Time 1 and Time 2, respectively).

5. Roediger, Watson, McDermott, and Gallo (2001) have recently shown that, across memory lists, higher levels of false recall were associated with lower levels of correct recall. Because in the present research, we examined the association between FAs and hits across *participants*, the present results should not be viewed as contradictory to Roediger et al.'s (2001) findings.

(Manuscript received June 28, 2000;
revision accepted for publication August 20, 2001.)