

Getting at the source of distinctive encoding effects in the DRM paradigm: evidence from signal-detection measures and source judgments

Glen E. Bodner^a, Mark J. Huff^{b*}, Raymond W. Lamontagne^{a,c} and Tanjeem Azad^d

^aDepartment of Psychology, University of Calgary, Calgary, AB, Canada; ^bDepartment of Psychology, Washington University in St. Louis, St. Louis, MO, USA; ^cAlberta Health Services, Calgary, AB, Canada; ^dDepartment of Psychological Sciences, Kent State University, Kent, OH, USA

ABSTRACT

Studying Deese–Roediger–McDermott (DRM) lists using a distinctive encoding task can reduce the DRM false memory illusion. Reductions for both distinctively encoded lists and non-distinctively encoded lists in a within-group design have been ascribed to use of a distinctiveness heuristic by which participants monitor their memories at test for distinctive-task details. Alternatively, participants might simply set a more conservative response criterion, which would be exceeded by distinctive list items more often than all other test items, including the critical non-studied items. To evaluate these alternatives, we compared a within-group who studied 5 lists by reading, 5 by anagram generation, and 5 by imagery, relative to a control group who studied all 15 lists by reading. Generation and imagery improved recognition accuracy by impairing relational encoding, but the within group did not show greater memory monitoring at test relative to the read control group. Critically, the within group's pattern of list-based source judgments provided new evidence that participants successfully monitored for distinctive-task details at test. Thus, source judgments revealed evidence of qualitative, recollection-based monitoring in the within group, to which our quantitative signal-detection measure of monitoring was blind.

KEYWORDS



False recognition; DRM paradigm; distinctiveness; signal detection; source judgments

The ability to discriminate previous experiences from novel experiences is a fundamental aspect of memory. Successful recognition is influenced by how items were encoded and by how their retrieval from memory is monitored and evaluated at test. In general, recognition is enhanced when manipulations improve the quality of encoding and/or retrieval-based monitoring. For example, recognition accuracy benefits from distinctive encoding via tasks such as self-generation (Slamecka & Graf, 1978) and mental imagery (Foley, 2012), as well as from the use of warnings and penalties at test that help to cultivate strategic memory monitoring (Gallo, Roediger, & McDermott, 2001).

Discrimination is more difficult when studied and non-studied items are similar. In the Deese–Roediger–McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995), each study list consists of associates (e.g., *hill*, *climb*, and *glacier*) of a single non-studied critical item (e.g., *mountain*). The critical items are typically recognised at very high rates, given their association with list items. Importantly, distinctive encoding tasks (e.g., generating the items from anagram cues or pairing items with pictures of their referents) often increase correct recognition of list

items while reducing false recognition of critical items (Dodson & Schacter, 2001; Gunter, Bodner, & Azad, 2007; Huff & Bodner, 2013; Hunt, Smith, & Dunlap, 2011; McCabe & Smith, 2006; Schacter, Israel, & Racine, 1999) – a mirror-effect pattern (Glanzer & Adams, 1990). In the present study, we used signal-detection analyses and source judgments to clarify how distinctive encoding tasks improve recognition in the DRM paradigm.

As reviewed by Huff, Bodner, and Fawcett (2015), the benefits of distinctive encoding in the DRM paradigm could arise at encoding, at retrieval, or both. During encoding, distinctive encoding tasks might emphasise item-specific processing (i.e., of the distinctive features of individual list items) over relational processing (i.e., of thematic or associative information about the list items), resulting in *impoverished relational encoding* that in turn reduces false recognition of critical items (e.g., Arndt & Reder, 2003; Hege & Dodson, 2004). At test, distinctive encoding might also lead participants to adopt a *distinctiveness heuristic* – a global strategy by which positive recognition decisions are made when recognition is accompanied by recollection of distinctive-task details (Dodson & Schacter,

CONTACT Glen E. Bodner  bodner@ucalgary.ca  Department of Psychology, University of Calgary, 2500 University Drive NW, Calgary, AB, Canada T2N 1N4

*Current address: Department of Psychology, University of Southern Mississippi, Hattiesburg, MS, USA

© 2016 Informa UK Limited, trading as Taylor & Francis Group

2001; 2002; Schacter et al., 1999). Use of a distinctiveness heuristic is predicated on the assumption that recollecting distinctive details provides diagnostic evidence that an item was studied. In the DRM paradigm, participants who perform this type of test-based monitoring should be better able to reject the critical items because memory for having performed the distinctive encoding task should be absent.

One method for assessing the contributions of encoding and retrieval processes in the DRM paradigm relies on a *within group* who perform distinctive processing on half the DRM lists and simply read the other half of the lists (i.e., a non-distinctive task). If distinctive tasks reduce relational encoding, then the DRM illusion should be reduced selectively for the lists studied using the distinctive tasks. Indeed, Arndt and Reder (2003) obtained this pattern when half the DRM lists were studied in a unique (versus same) font. In contrast, if distinctive tasks lead to use of a global distinctiveness heuristic at test, then critical items from both list types should be excluded, and hence, false recognition should be low and equivalent for both list types. Schacter et al. (1999) obtained this pattern when half the DRM lists were studied with a picture of each list item's referent, and half were not. Although it remains unclear why the choice of distinctive task can yield different patterns (see Huff et al., 2015), distinctive tasks appear to be able to exert their effects at encoding and/or at retrieval.

A second method relies on an *inclusion test* during which participants are asked to endorse list items as well as all other test items that are related to the list items (Brainerd, Payne, Wright, & Reyna, 2003; Gunter et al., 2007; Hege & Dodson, 2004; Hunt et al., 2011; Hunt & Smith, 2014). The inclusion test is assumed to lead participants to abandon a distinctiveness heuristic, given that determining whether each item was studied is not as focal. Using an inclusion test, distinctive processing has sometimes reduced endorsement of critical items (e.g., Hege & Dodson, 2004), consistent with impoverished relational encoding, but in other cases it has not (e.g., Gunter et al., 2007), consistent with use of a distinctiveness heuristic. Thus, similar to the within-group method, inclusion tests have yielded equivocal support for an encoding versus retrieval locus for the effects of distinctive encoding.

A third method, which we have advocated, is to examine *signal-detection indices* of encoding (i.e., memory information) and retrieval (i.e., memory monitoring) processes (Gunter et al., 2007; Huff et al., 2015; Huff & Bodner, 2013). This method allows the two loci to be separated and evaluated. In contrast, with the within group and inclusion test methods, evidence for one process denies a potential influence of the other process by default, thus eliminating the possibility of showing that both processes contribute to distinctive reductions. Moreover, use of a distinctiveness heuristic would result in a null effect using each method, which is not ideal.

Signal-detection analysis allows for the calculation of an index of memory information, d' , which is the standardised difference between the mean hit and false alarm rates (also

termed sensitivity; Macmillan & Creelman, 1991; Wickens, 2002). In the DRM paradigm, a d' measure for *critical items* can also be computed treating recognition of critical items from studied lists as hits, and recognition of critical items from non-studied lists as false alarms. We interpret this d' as an estimate of the amount of information encoded about critical items at study. Consistent with an impoverished relational encoding locus, Huff et al.'s (2015) meta-analysis showed that d' for critical items is generally lower for distinctive (versus non-distinctive) lists.

Signal-detection analyses can also be used to derive an index of test-based monitoring, such as use of a distinctiveness heuristic. Specifically, we calculate lambda (λ ; see Wickens, 2002) using only the false alarm rate for non-studied list items. Importantly, λ is computed independently of hit rates, which would also be expected to vary across distinctive and non-distinctive conditions. Higher values of λ are taken as evidence of greater memory monitoring at test. Consistent with use of a distinctiveness heuristic, Huff et al.'s (2015) meta-analysis showed that λ at test is greater when all DRM lists were studied using a distinctive task. Thus, the signal-detection method suggests that the mirror-effect pattern induced by distinctive tasks can be due to both encoding and retrieval processes.

Despite its merits, a limitation of the signal-detection method is that it only provides *quantitative* estimates of encoding and retrieval. It cannot reveal *qualitative* differences in the types of memory information participants encode, or in the types of memory information they monitor for at test. Estimating the overall amount of encoded memory information for critical items using d' may be appropriate, given that impaired relational processing would reduce associative activation (Roediger, Balota, & Watson, 2001) or thematic consistency (Brainerd & Reyna, 2002). However, λ cannot provide direct evidence that participants monitor for distinctive-task details during test.

An alternative possibility is that participants might base their recognition decisions on an evaluation of the overall strength or fluency of their recognition experience for a given item. Having studied some items in a distinctive task, they might simply apply a more conservative response criterion at test. Distinctively studied list items would surpass this criterion more than would either non-distinctively studied list items or all of the non-studied items – including the critical items. Thus, the recognition pattern typically attributed to use of a distinctiveness heuristic might instead be due to participants having adopted a more conservative response criterion. As discussed in Huff et al. (2015), a strength-based account of the effects of distinctive encoding tasks in the DRM paradigm has never been empirically evaluated.

We conducted a DRM recognition experiment designed to evaluate these alternatives. Recognition in a within group who studied 5 lists by reading, 5 by generation, and 5 by imagery was compared to a control group who studied all 15 lists by reading. The purpose of including two types of distinctive task is explained shortly. Based

on Huff et al.'s (2015) meta-analysis, the within group was expected to show higher correct recognition for the distinctive lists, and a reduced DRM illusion relative to the read control group. Our signal-detection indices were then analysed to see whether this mirror effect was driven by impaired relational encoding and/or by increased monitoring at test.

As we have discussed, the lambda index cannot specify the nature of participants' test-based monitoring. Importantly, therefore, we also asked the within group to make a read/generate/imagery source judgment after each "old" recognition judgment. Analyses of these source judgments allowed us to evaluate whether the within group likely monitored their recollections for specific distinctive-task details (e.g., using a distinctiveness heuristic) or simply set a more conservative response criterion. Specifically, if participants monitored for distinctive-task details, then their source judgments should accurately distinguish between the two strong/distinctive list types (i.e., generate versus imagery). In contrast, if they did not monitor or have access to such details, then their source judgments should be at chance for distinguishing between these two list types.

Method

Participants

University of Calgary undergraduates participated for course credit. They were randomly assigned to the within group or the read group (36 per group).¹

Materials

We selected 31 DRM lists from the norms of Stadler, Roediger, and McDermott (1999) and divided them into 2 sets of 15 lists, plus 1 practice list. Each list comprised 12 words (excluding the critical item) presented in descending associative strength, and each word consisted of at least 4 letters. Across participants, the set of lists that was studied (versus not studied) was counterbalanced. In the within group, we counterbalanced which 5-list subsets were assigned to the read, generate, and imagery conditions as well as the order the 3 list type blocks at study. A 120-item recognition test was presented in a newly randomised order for each participant. It comprised 45 list items (from list positions 2, 8, and 10), 45 list item controls from the new lists (from list positions 2, 8, and 10), 15 critical items from the studied lists (1 per list), and 15 critical item controls from the new lists (1 per list).

Procedure

Participants were tested individually via computer. They were told they would see several lists of words presented one word at a time. The read group was instructed to read each word aloud; they received an initial 12-item

practice list. The within group received an instruction screen prior to each set of 5 lists, and prior to studying each set they completed a 4-item practice list. For the read set, participants were asked to read each word aloud. For the generate set, they were told that the words would be presented as anagrams ("words with mixed-up letters") and their job was to solve each anagram (e.g., house) by switching letters 1 and 3 (e.g., uohse) or 2 and 4 (e.g., hsuoe), and to then say the solution aloud. Each rule worked for half the words per list, and rule type was counterbalanced across participants. Participants were allowed to ask for a "rule hint", which was sometimes necessary, and they could pass if they still could not solve the anagram, which was rarely the case. For the imagery set, participants were given the following instructions, based on Gunter et al. (2007):

When each word comes up, form as vivid an image as possible of yourself interacting with the word. Read each word out loud to the experimenter when you are done forming an image. For example, if you see the word "snow", vividly imagine yourself playing in the snow and then say "snow". Try forming a unique image for each word. After you have formed the image and said the word, describe the image out loud.

After each practice list, the experimenter provided feedback and answered questions as necessary to ensure the participant understood the task. For the imagery set, participants were told they would no longer need to describe their images aloud, but they were still to form images for each list word. They then studied the 15 critical lists. Each list was separated by the words "next list". The recognition instructions immediately followed the study phase. Participants were told they would see a list of words presented one at a time. They were asked to press a labelled "old" or "new" key on the keyboard to indicate whether the word was studied or new. The within group was further told that if they recognised a word they would be asked to press an additional labelled key to indicate whether the word had been read, solved as an anagram, or self-imaged.

Results

We averaged across the six orders of the read, generate, and imagery lists in the within group, except as noted below. The anagram completion rate in the within group was 96.7%; given this high accuracy, recognition was not conditionalised on correct generation at study. A .05 level of significance was used. Table 1 provides the recognition means, but our analyses focus in turn on the signal-detection indices and the source judgments.

Signal-detection analyses of memory information at encoding and monitoring at test

We used signal-detection indices of memory information (Table 1) to examine whether the distinctive encoding

Table 1. Mean (SE) proportion "old" responses and signal-detection indices.

Item type/index	Read group	Within group		
		Read lists	Generate lists	Imagery lists
List items	.77 (.02)	.66 (.02)	.83 (.02)	.86 (.02)
List item controls	.12 (.02)		.09 (.01)	
List items d'	2.13 (.08)	1.91 (.06)	2.54 (.12)	2.70 (.11)
List items λ	1.35 (.09)		1.48 (.08)	
Critical items	.68 (.03)	.53 (.05)	.43 (.05)	.56 (.04)
Critical item controls	.18 (.03)		.14 (.02)	
Critical items d'	1.62 (.11)	1.17 (.16)	0.94 (.15)	1.41 (.14)
Critical items λ	1.07 (.11)		1.19 (.09)	

Note: The λ indices for the within group are based on false alarms to control items which were not studied and hence could not be ascribed to a list type.

tasks performed by the within group influenced memory information at encoding (as indexed by d') and/or memory monitoring at test (as indexed by λ ; see Gunter et al., 2007; Huff et al., 2015; Huff & Bodner, 2013). False alarm rates of 0 and hit rates of 1 were adjusted using Macmillan and Creelman's (1991) 1/2N correction.

We first examined the effect of list type on correct recognition. For each condition, d' values were calculated by taking the z-score of a participant's hit rate for a given type of list item minus the z-score of the same participant's false alarm rate for the list item controls. A repeated-measures ANOVA showed that his d' index of memory information differed across the read, generate, and imagery list types in the within group (1.91 versus 2.54 versus 2.70), $F(2, 70) = 29.87$, $MSE = .21$. Both distinctive tasks boosted memory information for list items relative to reading ($ps < .001$). Importantly, imagery was not significantly more effective than generation, $F(1, 35) = 1.84$, $MSE = .23$, $p = .18$.

We also compared the amount of memory information encoded for each type of list item in the within group relative to the read control group. As expected, both the generation (2.54) and imagery (2.70) tasks increased the amount of memory information encoded for list items relative to the read group (2.13), $F(1, 70) = 7.95$, $MSE = .39$, and $F(1, 70) = 18.28$, $MSE = .32$. Thus, both of our distinctive tasks increased correct recognition. In contrast, less memory information was encoded for read list items in the within group (1.91) relative to the read group, $F(1, 70) = 4.88$, $MSE = .18$. A cost of distinctive encoding for non-distinctively encoded lists was also found in Huff et al.'s (2015) meta-analysis. In sum, both distinctive tasks increased memory information for studied items to a similar degree, likely by increasing item-specific processing. This equivalence is essential for interpretation of the source-judgment data we report below.

Turning to false recognition, the amount of memory information encoded about critical items was estimated using the d' difference between critical items (treated as hits) and critical item controls. This measure indexes the amount of information encoded for critical items when the corresponding list was studied versus not studied. A repeated-measures ANOVA revealed a difference across list types that brushed the .05 level, $F(2, 70) = 3.12$, MSE

$= .64$, $p = .051$. Previous studies have typically found that false recognition is low and similar for distinctive lists and non-distinctive lists in a within group (e.g., Dodson & Schacter, 2001; McCabe & Smith, 2006; Schacter et al., 1999). Here too, the amount of memory information encoded about critical items was not significantly different across the read and generate lists (1.17 versus 0.94), $F(1, 35) = 1.16$, $MSE = .81$, $p = .29$, or across the read and imagery lists (1.17 versus 1.41), $F(1, 35) = 1.77$, $MSE = .59$, $p = .19$. However, more memory information about critical items was encoded for imagery versus generate lists (1.41 versus 0.94), $F(1, 35) = 7.77$, $MSE = .51$, suggesting that the imagery task may not have been as effective at reducing relational encoding.

We next compared the amount of memory information encoded for critical items from each list type in the within group relative to our read control group. Relative to the read control group, the within group encoded less memory information about critical items from generate lists (1.62 versus 0.94), $F(1, 70) = 12.82$, $MSE = .65$, whereas this difference was not significant for imagery lists (1.62 versus 1.41), $F(1, 70) = 1.36$, $MSE = .57$, $p = .25$. Thus, the same imagery task that reduced the DRM illusion when applied to all DRM lists at study (Gunter et al., 2007, Experiment 3) did not do so in our within group design.

Huff and Bodner (2013) showed that distinctive encoding tasks can be biased to emphasise either item-specific or relational processing. We speculate that our imagery task permitted enough relational processing to undermine a reduction in false recognition, akin to Huff and Bodner's "relational processing" distinctive-task variants, and to other relational imagery variants in the literature (e.g., Burns, Jenkins, & Dean, 2007; Foley, Wozniak, & Gillum, 2006; Robin & Mahé, 2015). However, we contend that this result does not undermine our main goal, namely to test whether participants can distinguish between two similarly strong conditions with their source judgments, as reported below.

Relative to the read control group, the within group encoded significantly less memory information about critical items from read lists (1.62 versus 1.17), $F(1, 70) = 5.20$, $MSE = .70$. Performing distinctive encoding on some lists had a carryover effect on the encoding of the read lists, consistent with an impoverished relational encoding account. Indeed, when the read lists were studied after either the generation and/or imagery lists ($n = 24$), the mean d' index for critical items was 0.99, which was again lower than for the read control group, $F(1, 58) = 8.74$, $MSE = .65$. In contrast, when the read lists were studied first ($n = 12$), the mean d' index for critical items was 1.53, which was very similar to the read control group, $F < 1$. Interestingly, the reduction in relational processing for read lists after performing one or both distinctive tasks was not offset by an increase in item-specific processing, given the cost to recognition of read list items in the within group relative to the read control group that we reported above. Exposure to distinctive

tasks can reduce relational processing of the read lists, but it does not provide participants with a means for increasing item-specific processing of those lists.

The preceding analyses afford an alternate interpretation, however. Namely, the obtained pattern would also arise if the within group responded to the distinctive tasks by adopting a more conservative response criterion at test. Indeed, the d' s for list items studied in the two distinctive tasks were substantially higher than for all other items (see Table 2). This situation would enable the within group to set a more conservative response criterion – one that distinctive items would exceed more often than all other item types.

To test this possibility, we compared memory monitoring in the within group to the read control group. A λ index for list items was computed by taking the z-score of 1 minus the false alarm rate of list item controls, and a λ index for critical items was computed by taking the z-score of 1 minus the false alarm rate of critical item controls. Higher λ values are an indicator of more conservative memory monitoring at test, and Huff et al.'s (2015) meta-analysis showed that λ is typically higher for groups who study all their lists using distinctive tasks relative to read control groups. Contrary to that pattern, here λ was equivalent for the within and read groups both for list items (1.48 versus 1.35), $F(1, 70) = 1.16$, $MSE = .26$, $p = .28$, and critical items (1.19 versus 1.07), $F < 1$. Thus, monitoring at test was not more stringent when distinctive encoding was performed on only some of the lists (i.e., in a within group). Importantly, Huff et al. did not compare the λ s in prior studies for within groups to read control groups, and a review of their Table 1 suggests that monitoring was generally not higher for within groups.

Our signal-detection analyses suggest that distinctive encoding had effects at encoding (impairing relational encoding in the within group), but not at test (no increase in monitoring in the within group). However, rather than monitoring quantitatively more than the read control group, the within group may instead have monitored their recollections, qualitatively, for distinctive-task details. As we have discussed, our signal-detection index, λ , cannot detect qualitative aspects of monitoring. Therefore, we next turn to analyses of participants' source judgments, which provide a new source of evidence regarding how distinctive encoding shapes the monitoring processes used at test.

Source-judgment accuracy and patterns of misattributions

Table 2 provides the source judgment means for the within group. Source judgments for list items were generally quite accurate, and indeed, were well above a pure-chance level of accuracy (.33) for each list type ($ps < .001$).² Thus, the within group possessed memory for source information that would enable them to monitor their recollections for distinctive-task details when making their recognition

judgments. At chance memory for list source, in contrast, would challenge the claim that participants had relied on a distinctiveness heuristic when making their recognition judgments.

We next examined whether the within group's source judgments were able to distinguish between list items from generate and imagery lists. Using overall memory strength to infer list type would allow participants to accurately attribute "weak" items to the read lists, resulting in above-chance accuracy for read list items. However, given that the within group had similar amounts of memory information for generate and imagery list items, this group should have been at chance at attributing "strong" list items (i.e., those from the generate or imagery lists) to the generate versus imagery lists. Contrary to that possibility, source-judgment accuracy was well above .5 for both generate list items (.84), $t(35) = 13.60$, $SEM = 2.53$, and imagery list items (.68), $t(35) = 3.64$, $SEM = 4.88$.

However, most participants showed at least a small difference in d' for generate and imagery list items, which they could have used to infer the list source for a given item, rather than recollecting it. As a strong test, we examined source-judgment accuracy for participants who had *identical* d' scores for list items for generate and imagery lists (and thus no detectable memory strength difference). Despite the small sample size ($n = 7$), source-judgment accuracy remained well above .5 for both generate list items (.83), $t(6) = 6.73$, $SEM = 4.98$, and imagery list items (.80), $t(6) = 3.26$, $SEM = 9.23$. Thus, the within group could typically recollect which distinctive task they had performed on a given item, or at least on a given DRM list.

As shown in Table 2, falsely recognised critical items were typically attributed to the encoding task used to process the corresponding list (e.g., 75% of critical items from read lists were attributed to the read condition), as others have reported (e.g., Hicks & Hancock, 2002; Hicks & Marsh, 2001). Critical items were attributed to their corresponding list type above both a pure-chance level (.33; all $ps < .001$) and indeed, above the .5 level (all $ps < .05$), for

Table 2. Mean (SE) proportions of source judgments for the within group.

Item type/attribution	Read lists	Generate lists	Imagery lists
List items			
Attributed to correct list type	.73 (.04)	.84 (.03)	.68 (.05)
Attributed to read list type	–	.13 (.02)	.28 (.05)
Attributed to generation list type	.16 (.03)	–	.04 (.01)
Attributed to imagery list type	.11 (.03)	.03 (.01)	–
Critical items			
Attributed to "correct" list type	.75 (.06)	.75 (.06)	.63 (.06)
Attributed to read list type	–	.23 (.07)	.34 (.06)
Attributed to generation list type	.10 (.03)	–	.03 (.02)
Attributed to imagery list type	.15 (.05)	.02 (.01)	–
List item controls			
List attributions	.48 (.06)	.39 (.05)	.13 (.04)
Critical item controls			
List attributions	.51 (.07)	.38 (.08)	.11 (.04)

Note: For list and critical items each row sums to 1. For control items the columns sum to 1.

each list type. These source-judgment patterns could be indicative of phantom recollection of distinctive-task details (e.g., Brainerd, Wright, Reyna, & Mojardin, 2001), but they could instead reflect inferences drawn from recollecting the encoding task used to study the corresponding list. Our data cannot adjudicate between these possibilities. Regardless, the source patterns for list items and critical items support our contention that participants were able to monitor their recollections at test for distinctive-task details. In contrast, these patterns are not expected if the within group had merely adopted a more conservative response criterion at test – thus corroborating our finding that the monitoring was not higher in the within group than in the read control group.

Finally, the pattern of source-judgment misattributions further elucidates how participants evaluate their memories in the DRM paradigm. Two important patterns emerged. First, participants showed no systematic bias in whether they misattributed read list items to the generate versus imagery task (.16 versus .11), $F(1, 35) = 3.18$, $MSE = 146$, $p = .08$, or in their misattributions of recognised critical items from read lists to the generate versus imagery task (.10 versus .15), $F < 1$. These similarities further support our claim that memory strength was similar for the two distinctive list types.

Second, and in stark contrast, participants showed highly systematic biases in their misattributions for recognised items from generate lists and imagery lists. When participants failed to accurately report the list source for the distinctive list types, they systematically misattributed both list items and critical items to the read list type over the distinctive list type (all $ps < .01$). We dub this an “it-had-to-be-read” bias, akin to the “it-had-to-be-you” bias in reality-monitoring judgments (e.g., Johnson, Raye, Foley, & Foley, 1981). This is the first demonstration of this misattribution pattern; given our study is the first to examine source judgments among three encoding tasks. This pattern suggests that when participants fail to recollect distinctive-task details, they may use expected differences in memory strength for each list type to guide their source judgments – indeed, the same may be true of their recognition judgments. Once again, this systematicity suggests that memory judgments in the within group were influenced by consideration of list-based criteria (e.g., recollection of the distinctive task and/or of how other items from the same list theme were studied).

Discussion

Trying to pinpoint how people arrive at their memory decisions is no simple task. Recognition judgments can be based on familiarity and/or recollection (e.g., Yonelinas, 2002), and can also be influenced by response biases and random guessing. It is tempting to attribute recognition decisions to a recollection process whenever participants performed a nominally distinctive task at encoding. However, just because researchers create a potential for

using recollection does not guarantee that their participants will do so. In the DRM paradigm, the benefits of distinctive encoding tasks are often attributed to participants adopting a recollection-based distinctiveness heuristic (see Huff et al., 2015, for a review; see also MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010). Alternatively, distinctive encoding might lead participants to set a more conservative response criterion – and thus to continue to rely on a simple evaluation of familiarity or memory strength at test. As detailed next, our findings suggest that distinctive encoding reduces relational encoding and also equips participants with the ability to recollect distinctive-task details, thereby facilitating their recognition accuracy.

There are at least two mechanisms by which distinctive tasks can improve recognition accuracy in the DRM paradigm (for a review, see Huff et al., 2015). The first mechanism operates at encoding. Specifically, distinctive tasks can reduce relational processing at study and thus reduce the DRM illusion – likely by focusing participants on item-specific processing that improves correct recognition – thereby netting a mirror-effect pattern (Huff & Bodner, 2013). We evaluated this possibility using our d' index of the amount of information participants had in memory for list items and critical items. Relative to the read control group, the within group had more memory information for list items from generate and imagery lists, and less memory information for critical lures from generate and read lists. In the within group, reduced memory information for the read lists occurred only for participants who first encoded generate and/or imagery lists, providing novel support for the impoverished relational encoding account (e.g., Arndt & Reder, 2003; Hege & Dodson, 2004).

In contrast, our other distinctive encoding task – imagery – appears to not have successfully reduced relational processing. As Huff and Bodner (2013) showed, how a distinctive encoding task is structured will influence its effect on relational processing, in turn modulating its effectiveness at reducing the DRM illusion. Our imagery task/instructions may not have sufficiently emphasised the importance of generating unique, distinctive images for each list item (cf. Gunter et al., 2007, Experiment 3). Consistent with this possibility, integrative imagery tasks (i.e., those that relate study items together) inflate the DRM illusion relative to imagery tasks that emphasise separate individual images of each list item (Foley et al., 2006; see Burns et al., 2007). Despite our request to form unique images, some of our participants may have created integrative images.

The second mechanism operates at test. Specifically, distinctive encoding can result in increased memory monitoring at test, facilitating recognition of list items and reducing endorsements of critical items (Huff et al., 2015; Huff & Bodner, 2013). We evaluated this locus using our λ index of memory monitoring at test. This evaluation did not reveal higher λ values for either list item controls or critical item controls in our within group relative to the read control group. Prior demonstrations of increased monitoring

have always used a between-group design in which the distinctive group studied all of their lists using the same distinctive task. In contrast, monitoring in within groups has not been compared to read control groups to date (Huff et al., 2015, footnote 4). Our findings suggest that the presence of pure read lists in the within group may reduce memory monitoring. Relative to a group who studies all their lists distinctively, a within group might set a more lenient response criterion to ensure they do not miss too many of the items from the read lists.

Although our within group did not show a quantitative increase in monitoring, we were nonetheless able to use source judgments to provide evidence that these participants considered qualitative aspects of their encoding when monitoring their memories at test. The within group's source judgments were able to distinguish between the similarly "strong" generate and imagery conditions, suggesting they could recollect the task used to study a given item, or at least the task used to study items from a given DRM list theme. In contrast, had the within group merely evaluated the memory strength of each item at test then their source judgments should not have distinguished between the two strong/distinctive list types.

The addition of source judgments in the within group did not affect recognition rates (see note 1). Thus, the requirement to make source judgments is unlikely to have increased the reliance on recollection (as opposed to familiarity) for making recognition judgments. Had that been the case, then presumably their recognition rates would have differed in our within group (if they based their responses on recollection) and our read control group (if they based their responses on familiarity). However, this point raises three other potential limitations regarding our source-judgment evidence, which we identify and address in turn.

First, we cannot rule out the possibility that our within group based at least some of their recognition decisions on memory strength or familiarity. Indeed, this could explain why recognition rates were not influenced by the requirement to make source judgments. After making a strength-based recognition decision, the within group may have shifted to using a distinctiveness heuristic. Because participants were able to recollect distinctive-task details enabling accurate source judgments, we assume they would likely have based their recognition judgments on the retrieval of the same information. It seems unlikely that participants would set aside their detailed recollections for test items, and to instead rely on overall familiarity/strength as the basis for recognition (i.e., "cognitive economising"). Still, our data cannot rule out this possibility and therefore it warrants further study.

Second, as we have acknowledged, the above-chance source-judgment accuracy we report does not provide direct evidence that participants recollected distinctive-task details. Participants may have inferred the correct

source for at least some items by recollecting how other items from the same list theme were studied. For example, a participant might fail to recollect the task used to study "hill", but might recollect solving anagrams for "climb" and "glacier". This participant might accurately infer that "hill" was studied in the generate task. Consistent with this possibility, critical items were predominantly attributed to the corresponding task used to study the corresponding list items (Foley & Foy, 2008; Hicks & Hancock, 2002; Hicks & Marsh, 2001). In our example, the critical item "mountain" would likely be attributed to the generate list. Importantly, this possibility does not temper our claim that the within group was able to monitor for distinctive-task details. In order to use the inference-based strategy noted here, participants must have recollected distinctive-task details to connect at least some of the list items to their corresponding list.

Third, we acknowledge that at least some of the within group's source judgments may have been based on an evaluation of memory strength rather than on recollection of distinctive-task details. For example, if a participant experienced greater memory strength for generate items than imagery items, then that participant might attribute stronger recognition experiences to the generate task (and thus make "generate" source judgments) and weaker recognition experiences to the imagery task (and thus make "imagery" source judgments) – yielding above-chance accuracy. Although it is difficult to rule out this possibility, above-chance source-judgment accuracy persisted for the small set of participants who showed equivalent recognition accuracy for generate list items and imagery list items. It remains possible that the rest of our participants used strength differences to guide their source judgments. However, it seems unlikely to us that participants would cognitively economise to the point of avoiding using their recollections to inform their source judgments. After all, the memories we typically hold with the most confidence are those associated with recollection. For example, confidence is greater for recognition experiences classified as "remember" versus "know" (e.g., Wixted & Stretch, 2004).

Distinguishing between recollection- and strength-based changes in recognition accuracy is not a challenge unique to the DRM paradigm. Here we highlight one other example, and show how analysis of source judgments wedged these possibilities apart. The production effect in recognition memory refers to improved recognition of items read aloud over items read silently (e.g., MacLeod et al., 2010). This effect is typically attributed to the production task leading to the encoding of distinctive-task details. Memory for these details facilitates recognition at test, particularly if participants adopt a distinctiveness heuristic where they try to recollect these details. Alternatively, production tasks might simply boost memory strength. Bodner and Taikh (2012) argued that the evidence for a distinctiveness account of the production effect over a strength account was equivocal.

Ozubko, Major, and MacLeod (2014) reported an important test of the distinctiveness versus strength accounts of the production effect. At study, participants studied some items aloud, and other silently, and the strength of a subset of the silent items was increased to the level of the aloud items via repetition. Participants then made source judgments at test (i.e., aloud versus silent versus new). Ozubko et al.'s participants were able to discriminate between these twice-silent items and the aloud items, despite overall recognition rates being similar for both item types. Ozubko et al. argued that participants' ability to remember the study mode for produced items provided new support for a distinctiveness account over a strength account.

Our tack was different from Ozubko et al.'s (2014), but complementary. In our key conditions recognition was similar for two distinctive tasks (generate versus imagery), whereas in their key conditions recognition was similar for non-distinctive items (twice-silent) and distinctive items (aloud items). Importantly, the findings from the two studies dovetail. By the same token, their findings are open to some of the same criticisms as ours. For example, participants might use recollection to perform their study-mode task, but might rely on strength in a standard recognition task. Moreover, study-mode judgments could be based on strength differences between twice-silent items and aloud items for individual participants, even though strength of these two item types was similar at the group level. For now, on the grounds of parsimony, both studies provide evidence that participants have the ability to monitor their memory for distinctive-task details – and thereby can reap considerable memory benefits.

On this issue, our study provided additional evidence regarding the benefits, as well as potential costs, of performing distinctive encoding on a subset of items. Specifically, relative to the read control group, the within group showed a reduction in encoded memory information for critical items from the read lists (a benefit), but they showed a concomitant reduction in encoded memory information for the read list items (a cost). These findings replicate patterns reported in Huff et al. (2015), and suggest that performing distinctive encoding on a subset of items can have a mixture of positive and negative effects on the non-distinctively encoded items.

Finally, we evaluated the claim that distinctive encoding tasks operate at the "back end" by enhancing memory monitoring at test. Recent evidence suggests that distinctive tasks might also help constrain the retrieval of potential memory candidates at the "front end" (see Halamish, Goldsmith, & Jacoby, 2012; Hunt & Smith, 2014; Jacoby, Shimizu, Daniels, & Rhodes, 2005). Both mechanisms should increase monitoring in lambda (by reducing recognition claims for control items), and both should increase the accuracy of source judgments. Indeed, the two processes might sometimes operate in tandem, such that distinctive encoding constrains retrieval and then enhances

memory monitoring at test. Isolating the contributions of front end and back end processes is an important direction for future research – one that may provide valuable new insights into how distinctive encoding shapes memory accuracy.

Notes

1. Another within group ($n = 36$) who did not make source judgments was also tested. Analyses confirmed that the requirement to make source judgments did not affect recognition rates. Means and analyses involving this within control group are available from the first author.
2. We did not compare source-judgment accuracy as a function of list type because participants were likely biased toward making some list types more than others (e.g., as a function of what they could recollect about how they had studied an item and/or other items from a given list). As evidence of this bias, source judgments for control items were not evenly distributed across the three list types (see bottom two rows of Table 2). More informative, therefore, are our analyses focusing on whether source-judgment accuracy was above chance and whether there were systematic biases in the misattributions made for studied and critical items for each list type.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Sciences and Engineering Research Council of Canada [grant number RGPIN-2015-04131] to GEB and by the National Institutes of Health [grant number TG 32AG00000-38] to MJH.

References

- Arndt, J., & Reder, L. M. (2003). The effect of distinctive visual information on false recognition. *Journal of Memory and Language, 48*, 1–15.
- Bodner, G. E., & Taikh, A. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1711–1719.
- Brainerd, C. J., Payne, D. G., Wright, R., & Reyna, V. F. (2003). Phantom recall. *Journal of Memory and Language, 48*, 445–467.
- Brainerd, C. J., & Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science, 11*, 164–169.
- Brainerd, C. J., Wright, R., Reyna, V. F., & Mojardin, A. H. (2001). Conjoint recognition in phantom recollection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 307–327.
- Burns, D. J., Jenkins, C. C., & Dean, E. E. (2007). Falsely recalled items are rich in item-specific information. *Memory & Cognition, 35*, 1630–1640.
- Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17–22.
- Dodson, C. S., & Schacter, D. L. (2001). 'If I had said it I would have remembered it': Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review, 8*, 155–161.
- Dodson, C. S., & Schacter, D. L. (2002). When false recognition meets metacognition: The distinctiveness heuristic. *Journal of Memory and Language, 46*, 782–803.
- Foley, M. A. (2012). Imagery encoding and false recognition errors: Exploring boundary conditions of imagery's enhancing effects. *Memory, 20*, 700–716.

- Foley, M. A., & Foy, J. (2008). Pictorial encoding effects on memory confusions in the Deese-Roediger-McDermott paradigm: Evidence for the activation of spontaneous imagery. *Memory, 16*, 712–727.
- Foley, M. A., Wozniak, K. H., & Gillum, A. (2006). Imagination and false memory inductions: Investigation the role of process, content, and source of imaginations. *Applied Cognitive Psychology, 20*, 1119–1141.
- Gallo, D. A., Roediger, H. L. III, & McDermott, K. B. (2001). Associative false recognition occurs without strategic criterion shifts. *Psychonomic Bulletin & Review, 8*, 579–586.
- Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 5–16.
- Gunter, R. W., Bodner, G. E., & Azad, T. (2007). Generation and mnemonic encoding induce a mirror effect in the DRM paradigm. *Memory & Cognition, 35*, 1083–1092.
- Halamish, V., Goldsmith, M., & Jacoby, L. L. (2012). Source-constrained recall: Front-end and back-end control of retrieval quality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1–15.
- Hege, A. C. G., & Dodson, C. S. (2004). Why distinctive information reduces false memories: Evidence for both impoverished relational-encoding and distinctiveness heuristic accounts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 787–795.
- Hicks, J. L., & Hancock, T. (2002). Backward associative strength determines source attributions given to false memories. *Psychonomic Bulletin & Review, 9*, 807–815.
- Hicks, J. L., & Marsh, R. L. (2001). False recognition occurs more frequently during source identification than during old-new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 375–383.
- Huff, M. J., & Bodner, G. E. (2013). When does memory monitoring succeed versus fail? Comparing item-specific and relational encoding in the DRM paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1246–1256.
- Huff, M. J., Bodner, G. E., & Fawcett, J. M. (2015). Effects of distinctive encoding on correct and false memory: A meta-analytic review of costs and benefits and their origins in the DRM paradigm. *Psychonomic Bulletin & Review, 22*, 349–365.
- Hunt, R. R., & Smith, R. E. (2014). How distinctive processing enhances hits and reduces false alarms. *Journal of Memory and Language, 75*, 45–57.
- Hunt, R. R., Smith, R. E., & Dunlap, K. R. (2011). How does distinctive processing reduce false recall? *Journal of Memory and Language, 65*, 378–389.
- Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin and Review, 12*, 852–857.
- Johnson, M. K., Raye, C. L., Foley, H. J., & Foley, M. A. (1981). Cognitive operations and decision bias in reality monitoring. *American Journal of Psychology, 94*, 37–64.
- MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 671–685.
- Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York, NY: Cambridge University Press.
- McCabe, D. P., & Smith, A. D. (2006). The distinctiveness heuristic in false recognition and false recall. *Memory, 14*, 570–583.
- Ozubko, J. D., Major, J., & MacLeod, C. M. (2014). Remembered study mode: Support for the distinctiveness account of the production effect. *Memory, 22*, 509–524.
- Robin, F., & Mahé, A. (2015). Effects of image and verbal generation on false memory. *Imagination, Cognition and Personality: Consciousness in Theory, Research, and Clinical Practice, 35*, 26–46.
- Roediger, H. L. III, Balota, D. A., & Watson, J. M. (2001). Spreading activation and arousal of false memories. In H. L. Roediger, III, J. S. Nairne, I. Neath, & A. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95–115). Washington, DC: American Psychological Association.
- 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.
- Schacter, D. L., Israel, L., & Racine, C. (1999). Suppressing false recognition in younger and older adults: The distinctiveness heuristic. *Journal of Memory and Language, 40*, 1–24.
- Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592–604.
- Stadler, M. A., Roediger, H. L. III, & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition, 27*, 494–500.
- Wickens, T. D. (2002). *Elementary signal detection theory*. New York, NY: Oxford University Press.
- Wixted, J. T., & Stretch, V. (2004). In defense of the signal-detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review, 11*, 616–641.
- Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517.