Memory & Cognition
2006, 34 (1), 148-157

Release from generation failure: The role of study list structure

PHILIP A. HIGHAM and HELEN TAM
University of Southampton, Southampton, England

Three experiments, using the original encoding-specificity paradigm, investigated the role of study list structure in producing Higham and Tam's (2005) generation failure effect. Generation failure occurs when cued recall performance for strong, extralist cues is worse than target production in a control group that is given no study list but is instead required merely to generate responses to the same test cues. In the present study, generation failure was replicated in Experiment 1, and Experiment 2 demonstrated that strong, extralist cues were more likely to elicit targets in pure generation groups when participants had studied a list of strong associates than when they had studied a list of weak ones. In Experiment 3, participants were released from generation failure when a study list of moderate associates was used and the cue-to-target associative strength was equated between the reinstated- and extralist-cue conditions. Together, these results suggest that generation failure is partly attributable to participants' searching inappropriate domains that, though consistent with the study list structure, are unlikely to contain targets.

A number of authors over the years have noted that scientists, inventors, and problem solvers alike have difficulty thinking outside the domain in which they are working. For example, Kuhn (1970) argued that scientists work within "paradigms," venturing into new territory only rarely, when a scientific revolution makes it a necessity. Similarly, problem solvers work within the confines of their "mental sets" (see, e.g., Wertheimer, 1959). Indeed, many so-called insight problems have at their roots implicit, unwarranted assumptions which, if adhered to, make it impossible to reach a solution. In the same vein, new technology often shows clear roots in previous technology. For example, very early automobiles looked like cars of the steam train. Similarly, early, automated floor cleaners resembled modern-day hair dryers more than the vacuum cleaners we know today. Presumably, the intention of the inventors was to automate the action of the broom by driving dust away, rather than drawing it up into the machine itself.

All of these examples have one thing in common: They represent restrictions imposed by memory. Memory clearly benefits us in many ways, but there are also times when it imposes unnecessary restrictions and impairs our performance. In our view, these restrictions exist not just for scientists, problem solvers, and inventors, but can also plague participants in standard memory experiments. One could argue that problem solving and remembering are so different that useful analogies between them cannot be made. However, differences between problem solving and memory tasks are probably more apparent than real. Indeed, Koriat (2000) argued that memory tasks should be treated as problem-solving tasks. Likewise, Schooler, Dougal, and Johnson (1998) suggested that the feeling of recollection bears a close resemblance to the feeling of insight.

Recent experiments in our lab (see, e.g., Higham & Tam, 2005) have focused on a form of memory restriction in cued recall that we have dubbed *generation failure*. We believe an important factor in producing generation failure is that the retrieval context, in conjunction with prior learning, implicitly defines a search set that is inappropriate for the task at hand, in much the same way that insight problems implicitly define an inappropriate domain of possible solutions. In the experiments that we report here, we investigated generation failure in the cued recall paradigm that supported the principle of encoding specificity (see, e.g., Thomson & Tulving, 1970; Tulving & Thomson, 1973). In the classic version of this paradigm, participants first study a list of weakly associated word pairs (e.g., *bats*–BLOOD). At test, target recall performance is compared among three cuing conditions: (1) the reinstated, weak-cue condition (e.g., *bats*–?), in which the same cues shown at study are used to cue recall of the targets at test; (2) the extralist, strong-cue condition, in which new cues that are strongly related to the target (e.g., *donor*–?) are used to cue recall of the targets at test; and (3) the no-cue (free-recall) condition. This paradigm is interesting because recall performance with the strong, extralist cues is often poor despite the fact that these cues

Preparation of this article was supported by a research grant from the British Academy. Portions of this research were presented at the 43rd and 45th annual meetings of the Psychonomic Society (November, 2002, Kansas City, MO, and November 2004, Minneapolis, respectively). We thank Robert Greene, Bill Hockley, and Cathy McEvoy for helpful comments on an earlier draft of this article. We are also grateful to Wendy Kneller and David Brook for research assistance. Correspondence concerning this article should be addressed to P. A. Higham, School of Psychology, University of Southampton, Highfield, Southampton SO17 1BJ, England (e-mail: higham@soton.ac.uk).

are strong associates of the targets to be remembered (see, e.g., Higham, 2002, free-report condition; Higham & Tam, 2005, Experiment 1, free-report condition; Murphy & Wallace, 1974; Santa & Lamwers, 1974, Experiment 1; Thomson & Tulving, 1970, Experiment 2; see also Roediger & Adelson, 1980; Roediger & Payne, 1983).

Poor strong-cue recall performance is usually explained in terms of recognition failure rather than generation failure (see, e.g., Tulving & Thomson, 1973). That is, participants are assumed to have no difficulty generating targets in response to strong, extralist cues, but because these cues do not reinstate the study context, the targets are not recognized.[1]

However, more recently, Higham and Tam (2005) have demonstrated that this explanation is incomplete. In the context of cued recall, participants experience not just a failure to recognize targets when they are generated, but also a failure to generate targets in the first place. For example, Higham and Tam demonstrated that strong-cue performance in cued recall was worse than when the same cues were given to a group of control participants who were simply instructed to generate responses to the cues. This difference was found despite the fact that the control group was given no study list at all! The strong-cue difference between the cued recall and control groups was not attributable to participants' unwillingness to offer generated, but unrecognized, targets to the cues; the difference held despite the fact that the participants were given an incentive to produce several responses to every cue (up to a maximum of six) and were forced to provide at least one. Furthermore, the better performance in the control group in comparison with that of the cued recall group was also not due simply to the control group's providing more responses per cue; when trials were analyzed separately according the number of responses offered (one to six), the probability of producing the target was higher for the control group than for the cued recall group at all levels of responding. Thus, Higham and Tam demonstrated that the cued recall participants given strong, extralist cues suffered generation failure even though the same cues were quite effective at eliciting (unstudied) targets in the control group; the cued recall participants just could not think of plausible responses.

The purpose of the present experiments is to investigate one possible reason for generation failure in the encoding specificity paradigm: inappropriate study list structure. Higham and Brooks (1997) demonstrated that participants develop sensitivity to the rules experimenters use to homogenize word lists for use in memory experiments, such as those pertaining to word length, grammatical class, and lexical frequency. For example, new items that were consistent with the experimenter's rules were more likely than inconsistent items to be endorsed in both recognition and categorization. Higham and Tam (2005) reasoned that participants may "learn the experimenter's design" in cued recall as well. However, to date there is only weak experimental support for the learning of study list structure as an explanation of generation failure. Although generation failure occurred in Higham and Tam's experiments after participants had studied a list of weak associates, those cases were also demarcated by the use of direct-memory instructions, so it is not clear which variable was responsible for the effect. Indeed, for one experiment in which direct-memory instructions were not used and participants were asked simply to generate responses "like those seen at study" (rather than to "remember" them), no effect of study list structure was observed on target generation; that is, strong-cue target production in forced report, after participants studied a weak-associate study list, did not differ from target production after participants studied no list at all. This was true despite the fact that all targets were removed from the weak-associate study list, making episodic retrieval impossible. Although the failure to find an effect of study list structure in these pure-generation groups does not necessarily mean that such a structure plays no role in the generation failure effect observed in cued recall, it is certainly the case that there is a need for stronger experimental support if Higham and Tam's reasoning is to be considered viable.

In an attempt to marshal this support, in Experiment 1 we replicated generation failure using the same weak-associate study list structure as previously (Higham & Tam, 2005). To demonstrate that the study list structure affected generation processes, in Experiment 2 we compared the items that participants generated after studying lists of different structures (pure list of weak associates vs. pure list of strong associates). Finally, in Experiment 3 we released participants from generation failure by requiring them to study a pure list of moderate associates, and then tested them with both reinstated and extralist cues. Both test cue types were moderately and equally associated with the targets. Unlike the participants in Experiment 1, these cued recall participants were more likely to produce targets to the test cues than was a control group that received no study list but was given the moderate cues at test.

EXPERIMENT 1

In Experiment 1, we sought to replicate Higham and Tam's (2005) generation failure effect. In particular, we compared target production performance to strong, extralist cues between (1) a group given cued recall instructions, which studied targets paired with weak associates during the study phase, and (2) a control group given no study list but asked to generate words that "we have in mind." To minimize the influence of report bias, participants were required to provide responses to all test cues. To determine the phenomenology of the participants in the cued recall group, we applied a version of Higham and Vokey's (2004) independent-scales methodology. Specifically, the participants were required to indicate the degree to which each response was "recollected" and, separately, the degree to which it was "free associated," using two independent 6-point scales.

Method

Participants. Thirty-three students from the University of Southampton participated in return for payment or course credit. Seventeen were assigned to the cued recall group and 16 to the control group.

Design and Materials. The materials, taken from Higham and Tam (2005), were 100 target words, each with a weak associate and a strong associate. The mean probabilities of producing targets were 35% for strong cues and 1% for weak cues. No word was repeated across weak associates, strong associates, or targets.

The participants in the cued recall group were presented with all 100 targets, along with their weak associates, during the study phase. For each trial during the test phase, the participants were shown either the weak or the strong associate as a cue to retrieve the studied target. For counterbalancing purposes, approximately half of the participants were cued with the weak associate for the first set of 50 targets and with the strong associate for the second set of 50 targets, whereas this order was reversed for the remaining participants. The order in which the test trials were presented was uniquely randomized for each participant. After randomization, the first six test trials were treated as practice trials and were not included in the analyses. The participants in the control group were not presented with the study phase and were given the test phase only. The method of item counterbalancing in test was identical for this group and for the cued recall group.

Procedure. The cued recall group was first given the study phase, during which each weak cue–target pair was presented individually at the center of a computer screen for 3 sec. The cue words were presented in lowercase and to the left of the targets, which were presented in uppercase. The participants were instructed to study the targets and to attend to the cues, since they could possibly assist in target recall during a memory test that was to follow. The participants in the control group were not presented with this study phase.

In the test phase, both the cued recall and control groups were presented with cues, one at a time, at the center of a computer screen. The same cues were used for the two groups. Each cue was accompanied by a question mark (?) to its immediate right, indicating that a response was to be typed in using the keyboard. The cued recall participants were told that each cue had an uppercase word (i.e., a target) related to it and the cue could therefore assist them in recalling this target. These participants were not informed that half of the cues were reinstated from study whereas the other half were novel, nor were they provided with any information regarding the relationship of the cues to the targets. The control-group participants were told that, for each cue, there was a related word we had in mind, and that they were to respond with what they thought the related word would be. No specific information was provided as to the nature of the relationship between the cue and the word we had in mind (i.e., whether the relationship was semantic, orthographic, etc.). Both groups were also informed that, during test, each trial would begin with a "points stage," in which one point would be awarded for each correct answer but four points would be deducted for each incorrect answer. The participants could bypass this points stage by typing "B" (for blank), thus entering the "guessing stage," during which a response was required but no points would be awarded or deducted.

In addition, regardless of the stage at which the response was given, the cued recall participants were asked to rate, first, the extent to which their responses were recollected and, second, the degree to which they were freely associated. The definition of *recollection* was adapted from instructions used in Higham and Vokey's (2004) independent-scale methodology, which, in turn, were based on remember–know instructions that are standard in the literature (see, e.g., Rajaram, 1993; Tulving, 1985). *Free association* was defined as the process of producing a response by generating candidate words related to the cue word. These recollection and free-association ratings were made on a 6-point scale (1 = *extremely low confident correct*, 6 = *extremely high confident correct*). Following Higham and Vokey, we indicated to the participants that they could use the scales independently. For example, a response that was generated and then recognized could be rated as high on one scale and low on the other, or it might be rated as high on both scales. On the other hand, a response that was a "wild guess" might be rated as low on both scales.

Because the control participants were not engaged in an episodic memory task, it would have been nonsensical to request recollection ratings. Instead, they were asked to rate how confident they were that their response was the word we had in mind. A 6-point scale (1 = *extremely low confident correct*, 6 = *extremely high confident correct*) was used for this rating. The requirement that the control participants provide confidence ratings was intended only to match the procedures between the control and cued recall groups as much as possible, and so the confidence data are not discussed further.

Results

An alpha level of .05 was adopted for all statistical tests. Analyses were conducted on the free- and forced-report target production rates from both the cued recall and control groups (see Table 1). The term *free-report target production rate* refers to the proportion of targets that were offered in the points stage of the experiment, whereas the term *forced-report target production rate* refers to the summed proportion of targets offered in both the points stage and the guessing stage. For both groups, the term *targets* refers to the capitalized items that were shown to the cued recall participants during study, but not to the control participants. As is shown in Table 1, the mean target production rate (and hence the variance) for weak cues was virtually zero in both free and forced report for the control group, so these data were not analyzed further.

Of greatest importance were any group differences in target production rates for strong cues, since such differences in forced report might represent generation failure. Two independent-samples *t* tests, in which these rates were compared separately in free and in forced report, showed that strong-cue target production was significantly better for the control group than for the cued recall group in free report [$t(31) = 4.77$, $SE = 0.045$; control = .30, cued recall = .09] as well as in forced report [$t(31) = 3.24$, $SE = 0.038$; control = .41, cued recall = .29].

To investigate the effect of cue strength, a paired-samples *t* test was conducted to compare target production rates for weak cues versus those for strong cues in the cued recall group, first in free report and then in forced report. In free

Table 1
Mean Free- and Forced-Report Ratings of Target Production in Experiments 1 and 2 as a Function of Cue Type and Experimental Group

	Report Option			
	Free Report		Forced Report	
Experimental Group	*M*	*SD*	*M*	*SD*
Weak Cues				
Cued recall (Experiment 1)	.23	.13	.25	.14
Control (Experiment 1)	.00	.00	.01	.01
Weak study (Experiment 2)	.00	.01	.00	.01
Strong study (Experiment 2)	.00	.01	.00	.01
Strong Cues				
Cued recall (Experiment 1)	.09	.15	.29	.13
Control (Experiment 1)	.30	.10	.41	.09
Weak study (Experiment 2)	.30	.09	.36	.08
Strong study (Experiment 2)	.40	.10	.44	.11

report, a significantly greater proportion of targets was retrieved for weak cues than for strong cues [$t(16) = 2.41$, $SE = 0.059$; weak cues = .23, strong cues = .09], but this difference in retrieval between weak and strong cues was eliminated in forced report [$t(16) = 0.76$, $SE = 0.049$; weak cues = .25, strong cues = .29]. These data replicate similar patterns found by Higham (2002) and Higham and Tam (2005, Experiment 1).

The mean recollection and free-association ratings in the cued recall group are shown in Table 2. Fewer than half of all the participants correctly produced targets to weak cues in the guessing stage, and cell counts were low for the few cases in which targets were produced in this condition. In the same vein, only 65% of the participants were able to produce even a single target to strong cues in the points stage. Because of the number of empty cells, it was not feasible to statistically analyze the rating data, so we will limit our discussion to a brief description of some trends in the means. The inability to perform inferential statistics on the rating data was not of great concern to us because no major inferences were based on these data.

Generally speaking, the rating data revealed few surprises. The participants' recollection ratings were numerically higher for responses given to weak cues than for those given to strong cues, and were numerically higher for responses given in the points stage than for those given in the guessing stage. On the other hand, the participants' free-association ratings were numerically higher for responses given to strong cues than for those given to weak cues, particularly if the responses were targets.

In addition to these fairly predictable results, however, one unexpected and interesting finding was observed in the free-association ratings. It might be expected that the participants faced with strong, extralist cues first attempted to recollect the corresponding targets, but if recollection failed they passed to the guessing stage and offered a guess based on preexperimental associations. Because targets are strong associates of these extralist cues, such a strategic shift would explain the large improvement in performance on strong, extralist cues in forced report. However, the data in Table 2 suggest that this strategic shift did not occur. The participants' free-association ratings for extralist cues in the guessing stage were no higher, and in fact were numerically *lower*, than their ratings for responses given in the points stage. This pattern held for both target responses (points stage = 3.64, guessing stage = 3.21) and incorrect responses (points stage = 2.99, guessing stage = 2.95).

Table 2
Mean Recollection and Free-Association Ratings in Experiment 1 as a Function of Cue Type, Experimental Stage, and Response Accuracy

	Experimental Stage					
	Points Stage			Guessing Stage		
Rating	No.	*M*	*SD*	No.	*M*	*SD*
	Weak Cues					
Recollection						
Correct	17	4.87	1.09	8	2.73	1.61
Incorrect	17	3.22	1.28	16	1.41	0.47
Free association						
Correct	17	2.88	1.74	8	2.69	1.48
Incorrect	17	2.58	1.35	16	2.99	1.86
	Strong Cues					
Recollection						
Correct	11	3.27	1.42	16	1.51	0.65
Incorrect	12	2.72	1.36	16	1.29	0.38
Free association						
Correct	11	3.64	0.83	16	3.21	1.89
Incorrect	12	2.99	1.03	16	2.95	1.81

Note—No., number of participants contributing data to the mean.

Discussion

Experiment 1 replicated the generation-failure result of Higham and Tam (2005, Experiment 3). The cued recall participants, who studied a pure list of weak associates, produced fewer targets to strong, extralist cues at test than did the control participants, who were given the same test cues but *no study list*. This difference was apparent at both free report and forced report, the latter result suggesting that poor extralist-cue performance in the cued recall group was not due to a failure to recognize successfully generated targets. That is, the poor performance was obtained even after the participants had provided some kind of (perhaps unrecognized) response to all test cues.

Although counterintuitive, the fact that the free-association ratings to extralist cues were no higher in the guessing stage than in the points stage is consistent with generation failure. Had the participants fully switched to a reliance on preexperimental associations in the guessing stage, they would not have evinced the difficulty in producing targets that the comparison with the control group revealed. Instead, it appears that the cued recall participants avoided free-associating to the cues to some degree, even in the guessing stage, in which it was impossible to advance to the next trial until some response was provided. Such a pattern is consistent with Higham and Tam's (2005) suggestion that their cued recall participants searched a set of weak associates when presented with extralist cues because such a set was consistent with the study list structure. We will test this possibility more directly in Experiments 2 and 3.

Some readers may be concerned that the independent scale ratings that were required of the participants following each response in the cued recall group may have somehow distorted their recall performance. However, it is worth noting that strong-cue cued recall performance in Experiment 1 (free report = .09, forced report = .29) was virtually identical to that of the analogous group in both Higham (2002; free report = .07, forced report = .26) and Higham and Tam (2005, Experiment 1; free report = .09, forced report = .25), where no such ratings were required. Furthermore, strong-cue performance in all these cued recall groups was less than the target production rate in the corresponding report option conditions of the control group (free report = .30, forced report = .41). In the same vein, cued recall performance for weak cues in Experiment 1 (free report = .23, forced report = .25) was similar to performance in comparable conditions in Higham (2002) and Higham and Tam (2005; in both studies, free report = .26 and forced report = .28). Given the similarities of these means across experiments, we believe it is safe to conclude that the independent scale ratings did not influence cued recall performance.

EXPERIMENT 2

In Experiment 2, we compared the likelihood of generating targets between groups of participants given sham study lists (see Higham & Tam, 2005, Experiment 3). The weak-study group studied a list of weakly associated word pairs, much like that of the cued recall group of Experiment 1. The strong-study group, on the other hand, studied a list of strongly associated word pairs. Importantly, neither group had actually encountered during study any of the targets that were scored as correct during the test phase. Nonetheless, the same test cues used in Experiment 1 were presented during the test phase, and responses were scored as correct if the same targets as those defined in Experiment 1 were produced to these cues. The aim of Experiment 2 was to determine whether or not the study list structure had any effect on the kinds of items that participants generated. If it did, it would provide some support for the notion that the generation failure effect observed in Experiment 1 was partially attributable to the study list's being composed solely of weakly associated word pairs.

Method

Participants. Thirty-two undergraduate students participated either without compensation or in return for course credit. Sixteen were assigned to the weak-study group and 16 to the strong-study group.

Design and Materials. The design and materials were the same as those used for the cued recall group in Experiment 1, except that the study list was replaced with a new list of paired associates that contained none of the cues presented at test and none of the targets associated with those cues. The participants in the weak-study group studied 100 weakly associated word pairs (mean probability of target production from cue word = 1%), whereas those in the strong-study group studied 100 strongly associated word pairs (mean probability of target production from cue word = 33%). Although the study list changed, the same cues used in Experiment 1 were presented at test, and performance was scored according to the participants' tendency to respond with the targets as they were defined in Experiment 1.

Procedure. The study phase procedure was identical to that of the cued recall group in Experiment 1, except that in the study phase the weak-study group was given the weak-study list and the strong-study group was given the strong-study list. In the test phase, both groups were told that for each cue word we had "another word in mind" that was related to it, and they were to respond with what they thought that word might be. The participants were informed that the relationship between the cue word and the word we had in mind was similar to the relationship between the word pairs seen in study, but that none of the words seen in study was the same as the word we had in mind. As in Experiment 1, no specific information was provided regarding the nature of the relationship between the cues and the words we had in mind.

Results

As was expected, virtually no targets were generated for weak cues by either the weak-study (.00) or strong-study (.00) group in either free or forced report, so these data were not analyzed further (see Table 1). However, a number of targets were generated for strong cues. Two independent-samples *t* tests revealed that the strong-study group produced a greater proportion of targets for strong cues than did the weak-study group, both in free report [$t(30) = 2.85$, $SE = 0.033$; weak study = .30, strong study = .40] and in forced report [$t(30) = 2.29$, $SE = 0.034$; weak study = .36, strong study = .44].

Discussion

The results of Experiment 2 clearly demonstrate that the participants given generation instructions were sensitive to the constitution of the study list. That is, when these participants were asked to produce responses to the test cues to make pairs that were like those observed in the study phase, the strength of the association between the words of the pairs seen at study affected their performance. This effect occurred despite the fact that there were no instructions to attend specifically to the associative strength between the paired words shown at study. It seems that these participants *spontaneously* attended to this associative information and used it when asked to generate candidates at test.

These results contrast with those produced in a comparable experiment recently reported by Higham and Tam (2005). In that experiment, participants studied either a weak-associate study list or no study list at all, and then, like the participants in the present experiment, they were asked to generate words like those "we had in mind." However, a multiple-response methodology was used such that as many as six responses to each cue were permitted. Contrary to the present results, Higham and Tam found that there was no forced-report difference between these groups in the target production rate to strong cues, suggesting that study list structure had little effect on performance. One possible cause of the discrepancy between the results is that the high demands of the multiple-response methodology caused participants to start searching preexperimentally defined associates for candidate responses, rather than limit their search to domains defined by the study list structure. Alternatively, it may be necessary to compare specifically a generation group given a weak-associate list with a group given a strong-associate list (rather than no list at all) to reveal the difference. Regardless of the particulars, the important message from Experiment 2 is that the associative structure of the study list spontaneously affects the likelihood that the targets will be generated. These results add support to our hypothesis that the study list structure of weak-associate pairs in Experiment 1, and in the classic version of the encoding specificity paradigm more generally, plays a role in producing generation failure in the context of strong, extralist test cues.

EXPERIMENT 3

The results of Experiment 2 were obtained with groups of participants who were not engaged in a memory task. A potential criticism, therefore, is that the generation processes observed in Experiment 2 do not necessarily generalize to cued recall. To alleviate this concern, we conducted an experiment analogous to Experiment 1, except that the study list structure of the cued recall group was changed from a list of weak associates to a list of moderate associates. As in Experiment 1, half of the test cues were reinstated and half were not, but cue strength was equated between these two conditions. A second control group of participants generated responses to the test cues without exposure to any study list.

If our reasoning is correct, replacing the study list of weak associates used in Experiment 1 with one of moderate associates in Experiment 3, coupled with equating cue-to-target associative strength between the reinstated and extralist-cue conditions, should release participants from generation failure. For both test cue conditions, searching and generating candidates from sets of items consistent with the study list structure (i.e., items moderately associated with the test cues) is conducive to producing targets. Because inappropriate search sets no longer undermine performance with extralist cues in the cued recall group, some advantage of presentation of the targets in the study phase should become apparent. In other words, both free- and forced-report target production should be higher with both reinstated and extralist cues in the cued recall group than in the control group.

A second reason for conducting Experiment 3 was to determine what effect, if any, cue reinstatement has on forced report recall performance. Higham (2002) demonstrated that strong-cue recall performance was equivalent to weak-cue recall performance in forced report if the study list consisted of weakly associated word pairs—an effect that was replicated in Experiment 1 of the present research and in Higham and Tam's (2005) first experiment. A perfunctory interpretation of this pattern of results might be that, generally speaking, context reinstatement has no effect on recall once report bias is controlled. However, as both Higham (2002) and Vokey and Higham (2005) pointed out, this interpretation is almost certainly incorrect (see also Zeelenberg, 2005). The problem is that in Thomson and Tulving's (1970) classic paradigm, not only was report bias uncontrolled but there was also a confounding of test cue strength with context reinstatement. The result is that any effect of context in forced report (i.e., weak > strong) was probably *offset* by test cue strength (i.e., strong > weak). Thus, equivalence of weak- and strong-cue performance in forced report is probably more a result of two opposing influences canceling each other out than of a lack of any influence of context at all.

If this interpretation is correct, then balancing the associative strength between the reinstated and extralist-cue conditions *and* controlling report bias will determine whether or not cue reinstatement has any effect on recall. Higham and Tam (2005, Experiment 2) found that varying test cue strength and cue reinstatement independently revealed effects of both variables in forced report. In the same vein, Vokey and Higham (2005) found that once cue-to-target associative strength was controlled in the semantic specificity paradigm (see, e.g., Light & Carter-Sobell, 1970; Roediger & Adelson, 1980; Roediger & Payne, 1983), context reinstatement had reliable effects on cued recall performance in forced report. Thus, we expected that once associative strength was controlled in Experiment 3, cue reinstatement would have a reliable effect on recall.

Method

Participants. Thirty-two University of Southampton students participated in return for course credit or payment. Sixteen were assigned to the cued recall group and 16 to the control group.

Design and Materials. The materials were 100 target words, each with two cues (e.g., SHEET–*blanket–linen*), taken from the Edinburgh Associative Thesaurus (EAT). The mean cue-to-target associate strength was .21 for the first cue set, whereas it was .22 for the second set. These mean values did not differ [$t(98) = 0.95$, $SE = 0.006$], indicating that associative strength was balanced between the sets.[2]

The design was as in Experiment 1, except that the reinstated- and extralist-cue conditions no longer corresponded to (i.e., were redundant with) the weak- and strong-cue conditions, respectively. As described above, all cues were moderate associates of their corresponding targets. As in Experiment 1, no word was repeated across targets or cues.

Procedure. The procedure for the cued recall and control groups in this experiment was identical to that for their corresponding groups in Experiment 1, with one exception. Instead of the cued recall group's making ratings on independent recollection and free-association scales as in Experiment 1, both the cued recall group and the control group were required to rate their confidence in the correctness of their responses on a 6-point scale. Since none of our conclusions hinges on these data, they are not reported.

Results

The target production rates in Experiment 3 are shown in Table 3. The terms *Reinstated* and *Extralist* in Table 3 describe only the nature of the cues in the cued recall group and do not apply to the cues in the control group, to which no study list was shown. However, because different cues served in the reinstated and extralist conditions, the mean target production rate in the control group was calculated separately for these two cue sets. As can be seen in Table 3, our attempt to balance the strength of association between the different cue sets and their corresponding targets was successful. In the control group, the rate of target production to the cues that served as reinstated cues in the cued recall group did not differ from that of the cues that served as extralist cues in the cued recall group, in either free report [$t(15) = 1.81$, $SE = 0.018$; reinstated = .16, extralist = .20] or forced report [$t(15) = 1.39$, $SE = 0.017$; reinstated = .21, extralist = .23]. As was expected, the target production rates in the control group were also comparable to those reported in EAT (overall mean = .21).

Two 2 (group: cued recall vs. control) × 2 (cue: reinstated vs. extralist) mixed ANOVAs were performed on the target production rates, first in free report and then in forced report. In both analyses, group was the between-subjects factor and cue was the within-subjects factor. In free report, the cued recall group retrieved a significantly greater proportion of targets (.27) than the control group (.18) [$F(1,30) = 14.27$, $MS_e = 0.004$], and reinstated cues elicited a significantly greater proportion of targets

Table 3
Mean Free- and Forced-Report Ratings of Target Production in Experiment 3 as a Function of Cue Type and Experimental Group

	Report Option			
	Free Report		Forced Report	
Experimental Group	*M*	*SD*	*M*	*SD*
Reinstated Cues				
Cued recall	.41	.11	.52	.09
Control	.16	.08	.21	.07
Extralist Cues				
Cued recall	.13	.08	.29	.08
Control	.20	.06	.23	.07

(.29) than extralist cues (.16) [$F(1,30) = 55.02$, $MS_e = 0.005$]. The group × cue interaction was also significant [$F(1,30) = 86.42$, $MS_e = 0.005$]. The interaction occurred because for reinstated cues, a significantly greater proportion of targets was produced in the cued recall group (.41) than in the control group (.16) [$t(30) = 7.57$, $SE = 0.033$], whereas for extralist cues the difference was significantly reversed [$t(30) = 2.80$, $SE = 0.025$; cued recall = .13, control = .20]. The fact that a greater proportion of targets was produced to extralist cues in the control group than in the cued recall group might be interpreted to mean that generation failure occurred in this experiment just as it did in Experiment 1. However, the difference between the groups could be due to an effect of report bias, which was not controlled in free report. To determine whether the participants were released from generation failure, it was necessary to examine forced-report performance.

In forced report, the cued recall group again retrieved a significantly greater proportion of targets (.40) than did the control group (.22) [$F(1,30) = 71.59$, $MS_e = 0.004$], and reinstated cues (.36) elicited a significantly greater proportion of targets than did extralist cues (.26) [$F(1,30) = 39.95$, $MS_e = 0.004$]. The group × cue interaction was also significant [$F(1,30) = 60.99$, $MS_e = 0.004$]. The interaction occurred because, for reinstated cues, the difference in the rate of target production between the cued recall group (.52) and the control group (.21) was greater than that for extralist cues (cued recall = .29, control = .23). Importantly, although the latter difference was smaller than the former, leading to the interaction, both differences were significant [for reinstated cues, $t(30) = 11.03$, $SE = 0.028$; for extralist cues, $t(30) = 2.23$, $SE = 0.026$].

Discussion

The results of Experiment 3 confirmed both of our hypotheses. First, having the participants study a list of moderate associates and balancing the cue-to-target associative strength between the reinstated- and extralist-cue conditions released the participants from generation failure. The rate of forced-report target production to extralist cues was *higher* in the cued recall group than in the control group, an effect that was opposite to that observed in Experiment 1, in which a study list of weak associates was used. This release was likely due to the fact that the participants in Experiment 3 were not at a disadvantage in cued recall when faced with extralist cues, as those in Experiment 1 had been. The moderate associative relationship between the words of the pairs shown at study, and the fact that targets were moderate associates of the extralist cues used at test, meant that the participants were searching appropriate domains (i.e., domains likely to contain the target). As a result, encountering the targets during study gave the cued recall group an advantage over the generation group for both reinstated and extralist cues.

Second, clear effects of cue reinstatement were shown in both free and forced report. This result supports Vokey and Higham's (2005) argument that equivalent reinstated and extralist cued recall performance in forced report (see, e.g., Higham, 2002; Higham & Tam, 2005, Experiment 1; Experiment 1 of the present series) should not be considered evidence that, generally speaking, context reinstatement has no effect on recall once report bias is controlled (see also Zeelenberg, 2005). Instead, it seems more likely that context effects on recall are ubiquitous, even in those cases of forced report in which reinstated and extralist recall performance is equal. In such cases, context effects are offset by the cue strength advantage of extralist versus reinstated cues.

GENERAL DISCUSSION

In the present experiments, we investigated the source of generation failure. In Experiment 1, in which a study list of weak associates was used, cued recall participants were at a disadvantage with extralist strong cues relative to a control group provided with the same cues but no study list. This finding replicates an analogous generation failure effect reported by Higham and Tam (2005), who used a multiple-response methodology. However, the present results further suggest that the structure of the study list contributes to generation failure by causing participants to search for targets in inappropriate domains. In Experiment 2, participants given a study list of strong associates and required at test to make pairs like those seen at study were more likely to generate targets than were participants given a study list of weak associates. Finally, Experiment 3 demonstrated that requiring participants to study a list of moderate associates, and equating the cue-to-target associative strength between the reinstated and extralist-cue conditions, led to a release from generation failure. Unlike the cued recall participants in Experiment 1, analogous participants in Experiment 3 were more likely to produce targets to extralist test cues than were control participants who were given the same test cues but no study list and were then asked to generate responses to the cues.

The conclusion that release from generation failure was attributable to differences in the domains that were searched at test could potentially be bolstered with an analysis of nontarget responses. That is, if it is the case that the study list used in Experiment 1 led participants to search a domain of weak associates, whereas the study list used in Experiment 3 led participants to search a domain of moderate associates, then differences in the associative relationship between cues and nontarget responses should be observed across the two experiments. In particular, if our reasoning is correct, nontarget responses in the cued recall group of Experiment 1 should be less well associated to their respective cues than nontarget responses in the no-study-list control. Conversely, no such difference should be observed in Experiment 3.

To test this hypothesis, we conducted an item analysis. First, all nontarget responses for both weak and strong test cues were tabulated in both the cued recall and the control groups of Experiments 1 and 3. Second, for each cue, EAT was used to determine the probability that each nontarget response would be produced, and a mean associative strength for that cue was calculated. Third, two 2 (cue: reinstated vs. extralist) × 2 (group: cued recall vs. control) mixed ANOVAs were conducted on the mean associative strengths, the first for Experiment 1 and the second for Experiment 3.[3] The ANOVA for Experiment 1 revealed main effects of cue [$F(1,192) = 18.14$, $MS_e = 0.006$ (reinstated = .09, extralist = .04)] and group [$F(1,192) = 28.21$, $MS_e = 0.002$ (cued recall = .05, control = .08)]. The interaction was also significant [$F(1,192) = 6.93$, $MS_e = 0.002$], reflecting the fact that the difference between the cued recall and control groups was larger for weak, reinstated cues (cued recall = .07, control = .11) than for strong, extralist cues (cued recall = .04, control = .05). The main effect of group in this analysis supports our main hypothesis that the kinds of (nontarget) responses that the cued recall participants were producing on the test after studying weakly associated word pairs were less strongly related to the cues than were those produced by the control group. In contrast, the analogous ANOVA on the associative strength data from Experiment 3 revealed no significant effects [largest $F(1,187) = 2.01$, $MS_e = 0.002$, $p = .158$, for the interaction]. Thus, after the cued recall group studied a list of moderate associates, the mean associative strength between cues and respective nontarget responses did not differ from that in the control group (for reinstated cues, cued recall = .08 and control = .09; for extralist cues, cued recall = .11 and control = .10).[4]

The present research also demonstrated that context reinstatement has reliable effects on cued recall even when report bias was controlled. One possible interpretation of the equivalence of weak- and strong-cue performance in Higham's (2002) forced-report condition (an effect that was replicated in Experiment 1 of the present article) and in Higham and Tam's (2005) research is that previous demonstrations of context effects in the encoding specificity paradigm were, in fact, due to a failure to force output from participants. To the extent that participants withheld target responses, free-report recall performance—the measure of recall performance typically reported in this paradigm—will underestimate actual recall. Instead, Experiment 3 demonstrated that cue reinstatement had a substantial effect on recall even in forced report once cue strength was balanced between the reinstated- and extralist-cue conditions (see also Zeelenberg, 2005). Thus, an interesting situation seems to have arisen: By failing to control cue strength while considering only free-report performance (i.e., by having two problems with their experiments), Thomson and Tulving (1970) seem to have achieved an accurate depiction of the role of context, because cue strength exerts itself only at forced report. Thus, the bottom line seems to be that Thomson and Tulving were right in their conclusions and that the encoding specificity effect is alive and well. However, as Higham argued, " . . . the Thomson and Tulving (1970) experiments, by themselves, provide, at best, weak evidence for the encoding specificity principle, despite the fact that they are considered classic and cited in textbooks throughout the world as providing its very foundation" (p. 77).

One reaction to the fact that Thomson and Tulving's (1970) major conclusions were correct might be to suggest that there is nothing wrong with using free report as a measure of cued recall performance. But this suggestion assumes, at the very least, that participants can perfectly monitor their own recall processes. Such a notion, in the extreme form necessary to justify the use of free report in measuring performance, must be wrong. Participants do not always know when they are correctly retrieving targets and should be reporting them, nor do they always save incorrect responses for forced report. To the extent that monitoring and control processes are imperfect in cued recall, just as they are in recognition and other memory tasks, free report as a measure of recall performance will be distorted.[5]

Part of the reason that generation failure may not have been discovered in the 1970s when the encoding specificity principle was being heavily researched, along with related topics such as recognition failure of recallable words (see Nilsson & Gardiner, 1993, for a review), may have been the choice of control group. In many cued recall studies, the control group is a no-cue or free-recall group whose members attempt to recall targets without the benefit of any cues (see, e.g., Higham, 2002; Roediger & Payne, 1983; Thomson & Tulving, 1970). By comparing performance in the no-cue group to that of other groups provided with various test cues, researchers have attempted to determine the efficacy of those cues. However, the problem with the no-cue control group is that participants are likely to use some kind of cue to access memory, but these participant-defined cues are invisible to experimenters. Consequently, by comparing cued-recall performance to the performance of a no-cue control group, one is actually comparing performance between two cued recall groups, without the benefit of knowing what cues were actually used by one of the groups.

For these reasons, we prefer to compare two groups that were both given the same test cues, but with the control participants unable to make use of recent prior encounters because they had been exposed to either a sham study list or no study list at all (see, e.g., Higham & Tam, 2005). Another alternative is to include some cues in the test phase that have no corresponding target presented in the study phase (see, e.g., Vokey & Higham, 2005). We believe such controls are particularly necessary when participants have several routes to producing correct answers, such as when the cues used have some preexperimental association with the target items. The most important function of a control condition in this scenario is to determine the extent to which participants can produce targets without the benefit of having recently encountered those targets in the study phase. And indeed, it was the inclusion of control conditions of this sort in our own research that led to our discovery of generation failure.

We conclude by saying that we suspect that effects analogous to generation failure are fairly widespread. The fact that comparable limitations have been discussed in problem solving, creativity, and the philosophy of science adds credence to this suspicion. Our modest contribution has been to demonstrate how generation failure occurs in a domain that is well-known to memory theorists, and one where it is quite unexpected. Who would have predicted that an encoding and retrieval context could be fashioned in such a way that participants are unlikely to think of primary associates to common English words? Our goal—and, we believe, an important goal for memory theorists more generally—will be to determine the extent to which generation failure affects performance not just in cued recall, but in many other tasks requiring free-form responses from participants.

REFERENCES

Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition*, **30**, 67-80.

Higham, P. A., & Brooks, L. R. (1997). Learning the experimenter's design: Tacit sensitivity to the structure of memory lists. *Quarterly Journal of Experimental Psychology*, **50A**, 199-215.

Higham, P. A., & Gerrard, C. (2005). Not all errors are created equal: Metacognition and changing answers on multiple-choice tests. *Canadian Journal of Experimental Psychology*, **59**, 28-34.

Higham, P. A., & Tam, H. (2005). Generation failure: Estimating metacognition in cued recall. *Journal of Memory & Language*, **52**, 595-617.

Higham, P. A., & Vokey, J. R. (2004). Illusory recollection and dual-process models of recognition memory. *Quarterly Journal of Experimental Psychology*, **57A**, 714-744.

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory & Language*, **48**, 704-721.

Koriat, A. (2000). Control processes in remembering. In E. Tulving & F. I. M. Craik (Eds.), *Oxford handbook of memory* (pp. 333-346). Oxford: Oxford University Press.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, **103**, 490-517.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Light, L. L., & Carter-Sobell, L. (1970). Effects of changed semantic context on recognition memory. *Journal of Verbal Learning & Verbal Behavior*, **9**, 1-11.

Murphy, M. D., & Wallace, W. P. (1974). Encoding specificity: Semantic change between storage and retrieval cues. *Journal of Experimental Psychology*, **103**, 768-774.

Nilsson, L.-G., & Gardiner, J. M. (1993). Identifying exceptions in a database of recognition failure studies from 1973 to 1992. *Memory & Cognition*, **21**, 397-410.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition*, **21**, 89-102.

Roediger, H. L., III, & Adelson, B. (1980). Semantic specificity in cued recall. *Memory & Cognition*, **8**, 65-74.

Roediger, H. L., III, & Payne, D. G. (1983). Superiority of free recall to cued recall with "strong" cues. *Psychological Research*, **45**, 275-286.

Santa, J. L., & Lamwers, L. L. (1974). Encoding specificity: Fact or artifact? *Journal of Verbal Learning & Verbal Behavior*, **13**, 412-423.

Schooler, J. W., Dougal, S., & Johnson, M. K. (1998, November). *The self-discovery effect: When solving is confused with remembering.* Paper presented at the 39th Annual Meeting of the Psychonomic Society, Dallas.

Thomson, D. M., & Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology*, **86**, 255-262.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, **26**, 1-12.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, **80**, 352-373.

Vokey, J. R., & Higham, P. A. (2005). *Components of recall: The semantic specificity effect and the monitoring of cued recall.* Unpublished manuscript.

Wertheimer, M. (1959). *Productive thinking.* New York: Harper & Row.

Zeelenberg, R. (2005). Encoding specificity manipulations do affect retrieval from memory. *Acta Psychologica*, **119**, 107-201.

NOTES

1. Although Tulving and Thomson (1973) argued against the viability of a particular class of the generate–recognize model—that which assumes a single representation per item—they and others found it useful to distinguish between separate generation and recognition stages of recall:

> it is helpful to remember that the procedure we used rendered many recallable words not recognizable, but it did not seem to affect the ability of subjects to successfully generate copies of target words in response to extralist cues. The failure of retrieval as envisaged by the two-process theory had its source in the recognition phase and not the generation phase. (p. 365)

Like Tulving and Thomson, we adopt the distinction in the present article between a generation (or memory access) stage of recall and a separate recognition (or monitoring) stage. However, our use of these terms should in no way be interpreted as either our promoting classic generate–recognize theory or our attributing acceptance of specific versions of such models to Tulving and colleagues.

2. One degree of freedom was lost in this analysis because one word trio (desk–*classroom–information*) was not available on EAT.

3. The terms *reinstated* and *extralist* are used here to refer to conditions of the same name for Experiment 3, but for Experiment 1 they refer to the "weak" and "strong" conditions, respectively. Because cue strength and context reinstatement were confounded in Experiment 1, this change merely constitutes a relabeling of the conditions, which was done so that the analysis of nontargets could be more easily described.

4. Because the reported ANOVAs were based on items, group was a within-subjects factor and cue type was a between-subjects factor. Whether the nontarget response was given in free or in forced report was not considered. For some cues, no participant gave any incorrect responses, so these cues were eliminated from the analysis. To be specific, in Experiment 1 no weak cues and six strong cues were eliminated; in Experiment 3, eight reinstated cues and three extralist cues were eliminated. The main effects of cue and of the cue × group interaction on the data from Experiment 1 likely resulted from the fact that primary associates of many of the strong cues were systematically eliminated from the analysis because they were targets. In addition to decreasing the overall associative strength between strong cues and nontarget responses, this systematic elimination of primary associates likely compressed any group differences for these cues because of floor effects, leading to the interaction. However, regardless of the reason for effects involving cue

type, it is important not to be distracted from the principal result derived from these item analyses: A group main effect was observed in Experiment 1 but not in Experiment 3.

5. The extent to which free-report performance is a distorted index of recall performance is a function of a complex interplay of several factors, including monitoring, report bias, retrieval probability, and the probability of producing targets using processes other than recall (e.g., preexperimental associations), which affects monitoring. The point system that we used (+1 for a correct response and −4 for an incorrect response) may well have produced a more conservative report bias than standard cued recall instructions in which participants are encouraged, but not forced, to guess. Because conservative bias exacerbates the distortion associated with free-report measures, it is conceivable that we encountered more free-report distortion than is standard in the literature. Nonetheless, in our view, because free-report measures are influenced by so many factors and because it is impossible to determine the role of these factors by examining free-report performance alone, we believe most circumstances call for a forced-report measure as well. Such data are obtained very straightforwardly, and by comparing free- and forced-report performance it is possible to obtain measures of monitoring and report bias (for further discussion, see Higham, 2002; Higham & Gerrard, 2005; Higham & Tam, 2005; Kelley & Sahakyan, 2003; Koriat & Goldsmith, 1996).

(Manuscript received July 16, 2004;
revision accepted for publication February 11, 2005.)