Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited

PHILIP A. HIGHAM

University of Southampton, Southampton, England

Performance on tests in which there is control over reporting (e.g., cued recall with the option to withhold responses) can be characterized by four parameters: free- and forced-report retrieval (correct responses retrieved from memory when the option to withhold responses is exercised and when it is not, respectively), monitoring (discrimination between correct and incorrect potential responses), and report bias (willingness to report responses). Typically, researchers do not examine all these components in cued-test performance; blanks are sometimes counted the same as errors, meaning that the (free-report) performance index is contaminated with report bias and monitoring ability. In this research, a two-stage testing procedure is described that allows measures of free- and forced-report retrieval, monitoring, and bias to be derived from the original encoding specificity experiments (Thomson & Tulving, 1970). The results show that their cue-reinstatement manipulation affected free-report retrieval, but once report bias and monitoring effects were removed by forcing output, retrieval was unaffected.

The accuracy of memory reports can be improved by the exertion of some control over how much and what is reported. For example, people who choose to report only memories about which they are very confident are likely to have more accurate memory reports than people who report everything that comes to mind (see, e.g., Koriat & Goldsmith, 1996c). Although the effect of report option on memory accuracy is well known, very few analytical tools are currently available to memory researchers to estimate or otherwise control report bias. Within the context of free recall, Roediger and Payne (1985) have stated, "It is rather puzzling that the problem of recall criteria [report option] has received so little attention, since some of our most popular theories make strong assumptions about their role, as do commonsense folk theories" (p. 7).

In laboratory research, when participants are required to give responses to a finite set of cues (e.g., questions,

words, pictures, word stems, etc.), report option has generally been considered a nuisance factor. Researchers typically try to control or eliminate any influence of report option by, for example, forcing participants to respond to every test cue, having them guess if necessary. However, not all researchers force responses to all cues, and in some cases, this allows metacognitive and decision-making parameters to distort the memory index. Indeed, the experiment described in this paper, which revisits the effect of context reinstatement on memory, demonstrates that previous failures to control report option in a cued-testing paradigm might have led to some misleading interpretations of experimental data.

The problem of dealing with the effect of report option on memory performance is just as serious in many everyday memory situations, in which memory correspondence and accuracy is emphasized. For example, consider the literature on the cognitive interview and its effect on memory (e.g., Fisher & Geiselman, 1992; see Memon & Higham, 1999, for a recent review). Developers of the cognitive interview have incorporated techniques into the interview that have been shown to enhance memory performance in laboratory studies (e.g., mental reinstatement of context). The hope is that these techniques will increase the likelihood that interviewees will be able to remember details and/or events that otherwise they would not have been able to recall during the interview process. In other words, there is an assumption that these techniques actually enhance the accessibility of memory traces. However, some concern has been raised lately over the possibility that the cognitive interview merely affects the report criterion (i.e., it renders it more

Preparation of this article was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) research grant to the author. Portions of this research were presented at the Ninth Annual Meeting of the Canadian Society of Brain, Behaviour and Cognitive Science, June 1999, in Edmonton, Alberta, Canada, and at the 40th Annual Meeting of the Psychonomic Society, November, 1999, in Los Angeles. I thank Casey Hrabchuk for helpful discussion and research assistance and Peter Winters for additional research assistance. I also thank John Dunlosky, Asher Koriat, and Morris Goldsmith for helpful comments on earlier drafts of this article. Correspondence concerning this article should be addressed to P. A. Higham, Department of Psychology, University of Southampton, Highfield, Southampton SO17 1BJ, England (e-mail: higham@soton.ac.uk).

— Accepted by previous editorial team

liberal relative to a standard interview) and has no effect on the accessibility of memory traces (see, e.g., Fisher, 1996; Higham & Roberts, 1996; Memon & Higham, 1999; Memon & Stevenage, 1996; Roberts & Higham, in press, for discussion). Similar claims have been made regarding the effect of hypnosis on memory performance (e.g., Dywan & Bowers, 1983; Klatzky & Erdelyi, 1985); that is, hypnosis serves merely to increase the amount of information provided in the memory report, not to make memory traces more accessible. Despite all this concern, however, there is no well-accepted analytical tool or method that can be used to gain an estimate of report bias in situations in which reporting is optional. One purpose of this paper is to introduce a methodology that is simple to apply in cued testing procedures and that allows various memory parameters—including report bias—to be determined. Each parameter reflects a distinct aspect of memory performance that is qualitatively and quantitatively separable from the others.

Koriat and Goldsmith (1996c; see also Koriat & Goldsmith, 1994, 1996a, 1996b, and Koriat, Goldsmith, & Pansky, 2000) also gained an estimate of participants' report criterion when they were given the opportunity to withhold responses. To do so, they gave participants a set of general knowledge questions and had them provide answers to the questions under conditions of either forced report or free report. For forced report, participants had to provide an answer to every test question, guessing if necessary. For free report, participants answered questions with an accuracy incentive; they were penalized (by having money subtracted from their total) for incorrect answers and rewarded (by having money added to their total) for correct answers. However, they could avoid being penalized (and being rewarded) by leaving questions blank, in which case no money was added or subtracted from the total. It is assumed in Koriat and Goldsmith's (1996c) model that an input question causes a "best-candidate" answer to be retrieved from long-term memory and be considered for report. A monitoring mechanism estimates the probability that the best candidate is the correct answer (the *assessed probability*— P_a) and assigns it to the candidate. In the free-report condition, the candidate is reported if its associated P_a is greater than or equal to some predetermined *report criterion probability*— P_{rc} —otherwise, it is withheld. P_{rc} reflects the report criterion and can be estimated with their model by comparing performance under free- versus forced-report conditions. However, their model also allows for other memory parameters to be determined. For example, *monitoring effectiveness* is "the extent to which the assessed probabilities successfully differentiate correct from incorrect candidate answers" (Koriat & Goldsmith, 1996c, p. 494). Also, *retrieval* is the percentage of input questions that are answered correctly under forced-report conditions. The model and methodology advocated in this paper also allows for the estimation of monitoring and retrieval in addition to report bias, but this is achieved by using a model analogous to that underlying signal detection theory (SDT).

Although Koriat and Goldsmith's (1996c) model is similar in spirit to an SDT model in that there is an attempt to separate response effects from memory effects, they argue that an SDT model is unsuitable for gaining separate estimates of bias, monitoring, and retrieval.

Thus in the forced-report "old/new" paradigm to which signal-detection methods are typically applied, control is isolated in terms of the parameter β , yet retention (overall memory strength) and monitoring effectiveness (the extent to which a person's confidence distinguishes old–studied from new–foil items) cannot be operationally or conceptually separated: Both are equally valid interpretations of d' (Koriat & Goldsmith, 1996c, p. 506).

Koriat and Goldsmith's arguments are true for the typical application of SDT to standard yes/no testing paradigms (e.g., old/new recognition); it is impossible to obtain separate estimates of bias, monitoring, and retrieval with standard SDT because there are only two parameters (β and d'). However, because report option is generally not an issue with yes/no recognition research (participants typically respond "yes" or "no" to all test items), there is no need to gain these separate estimates. The more relevant question is whether or not these separate estimates can be obtained by using an SDT analogue model that can be applied to paradigms in which a report option is exercised. The present study addresses exactly this question, and it suggests that such an application is not only feasible, but has the potential to be very informative about memory and the decision processes involved with tests containing a finite set of cues.

The SDT Model and Indices of Performance

The psychological model that is adopted in the present study is essentially the same as the one presented in Koriat and Goldsmith (1996c), but the methodology, performance indices, and methods of data analysis are quite different. In response to an input cue (word stem/fragment, word associate, general knowledge question, etc.), it is assumed that a "best-candidate" answer is retrieved from long-term memory and considered for report. The best-candidate answer is the one to which the monitoring mechanism assigns the highest confidence of accuracy. Like Koriat and Goldsmith (1996c), no claims are made in regard to the nature of retrieval (e.g., whether the candidate is retrieved from episodic or semantic memory); only the accuracy of the retrieval and the decision processes that follow retrieval of the best candidate are of interest. Participants are assumed to set a report criterion along the confidence dimension so that any items with accuracy confidence above the criterion are reported, whereas items with accuracy confidence below the criterion are withheld. The setting of the report criterion is assumed to depend on situational demands and accuracy incentives, just as with Koriat and Goldsmith's (1996c) model.

Both Koriat and Goldsmith's (1996c) model and the SDT approach require information regarding the accuracy of both reported candidates and initially withheld candidates. In other words, the frequencies are needed

for all four cells of the 2 (accurate/inaccurate candidate) $\times 2$ (report/withhold response) contingency table shown in Table 1. The a cell in Table 1 represents the number of responses that are correct and reported in free report; b is the number of incorrect/reported responses; c is the number of initially withheld responses that are correct at forced report; and d is the number of initially withheld/incorrect responses at forced report.

To obtain these frequencies, the participants' performance was evaluated under free-report and forced-report conditions, just as with Koriatic and Goldsmith's (1996c) methodology. For free report, participants were tested with a finite number of cues that required responses, but an incentive for leaving responses blank was provided with a pay-off schedule; a reward was attached to correct-reported answers, a penalty was attached to incorrect-reported answers, but no penalty was attached to blanks. The number of correct and incorrect responses made during the free-report phase of the experiment provided data for cells a and b , respectively, in Table 1. In forced report, the participants were persuaded to provide responses to all test cues, guessing if necessary. The number of new correct and incorrect initially withheld answers revealed under these forced-report circumstances provided data for cells c and d , respectively, in Table 1.

According to SDT, in the present research, correct and incorrect candidates are assumed to form two overlapping distributions over a dimension of subjective confidence of accuracy, with correct candidates having higher confidence on average than incorrect candidates. The assignment of a reasonably placed report criterion on the confidence dimension splits the correct and incorrect candidate distributions into two, rendering four areas corresponding to hits, false alarms, misses, and correct rejections as in standard SDT. Once the hit and false-alarm rates are derived from the a , b , c , and d frequencies (see Table 1 for the formulae), various memory parameters can be derived. The index chosen for *monitoring* was A' , a nonparametric measure of discrimination (e.g., Grier, 1971). Essentially this measure indicates the degree to which participants have the tendency to report correct candidates and withhold incorrect ones. The index chosen for *report bias* was B_D'' (Donaldson, 1992), a nonparametric measure of bias. This index indicates the participants' tendency to report candidates, regardless of their accuracy. Other measures of discrimination (monitoring) and (report) bias might have

been chosen, but nonparametric measures avoid the strong distributional assumptions that would be necessary to use, say, d' and β .

The performance indices chosen here for monitoring and report bias differ from those of Koriatic and Goldsmith (1996c) in that they are based on the actual reporting behavior of participants, not on their confidence ratings. It will be argued that such measures are more direct indices of participants' performance, and they are not subject to problems associated with the analysis and interpretation of confidence ratings. These problems, and the implications of the difference between the two types of performance index, are discussed in more detail below.

In addition to the SDT measures of monitoring and bias, two types of *retrieval* were also calculated. The first, *free-report retrieval*, refers to the number of correct responses on the cued-memory test, divided by the number of cues on the test, when participants were permitted to leave cues blank. The second, *forced-report retrieval*, refers to the number of correct responses on the cued-memory test, divided by the number of cues on the test, when participants were forced to give a response to every cue. The formulae based on the a , b , c , and d frequencies that were used to derive the retrieval measures are also shown in Table 1.

Overview of the Experiment

Whenever participants exercise a report option, they potentially fail to report correct candidates that have been retrieved, but this cannot occur under forced-report conditions. Thus, free-report retrieval is generally an underestimate of forced-report retrieval. The SDT model assumes that the degree of this underestimation depends on at least two factors:

1. *The placement of the report criterion*—the more conservative the placement of the report criterion, the more free-report retrieval will be an underestimation of forced-report retrieval (i.e., c will be high; see Table 1).
2. *The level of monitoring*—the less effective the monitoring, the more free-report retrieval will be an underestimation of forced-report retrieval. If participants have perfect monitoring and report only correct answers at their report criterion (free report), no new correct responses will be revealed when reporting is forced, so retrieval will be unaffected. However, if monitoring is less than perfect, some correct answers are likely to be withheld at the report criterion. When these correct answers are revealed in forced report, retrieval will be increased relative to free-report retrieval.

These two assumptions were used to clarify the interpretation of Thomson and Tulving's (1970) results demonstrating the encoding specificity principle. In a study phase, participants were given targets to memorize that were paired with weak associates (e.g., *bats*—BLOOD, where BLOOD was the target and *bats* was the weak associate). At test, participants were asked to recall target words under three different conditions. In the weak-cue condition, the same weak cue that was presented with the

Table 1
A 2×2 Contingency Table and Formulae Used to Derive the Various Measures Discussed in the Text

Response	Candidate Answer	
	Correct	Incorrect
Reported	a	b
Withheld	c	d

Note—Free-report retrieval = $a/(a+b+c+d)$; forced-report retrieval = $(a+c)/(a+b+c+d)$; hit rate (h) = $a/(a+c)$; false-alarm rate (fa) = $b/(b+d)$; monitoring = $A' = .5 + [(h-fa)(1+h-fa)]/[4h(1-fa)]$; report bias = $B_D'' = [(1-h)(1-fa)-h*fa]/[(1-h)(1-fa)+h*fa]$.

target at study was presented again, and participants were asked to recall the associated target (e.g., *bats*-?). In the strong-cue condition, a new strong associate of the target was presented, and participants were asked to recall the related target word (e.g., *donor*-?). Finally, as a baseline, participants in the no-cue condition were simply asked to recall as many of the targets as possible without the assistance of any cues.

In Thomson and Tulving's (1970) original version of this experiment, they found that retrieval in the weak-cue condition was much better than retrieval in the strong-cue condition. In fact, the strong cues did not seem to augment retrieval at all; for the most part, performance under strong cues was no better than performance under no cues. In their abstract, they stated that after participants studied weak associate–target pairs,

recall of the TBR [to-be-remembered] words in the presence of these [weak] cues was greatly facilitated in comparison to noncued recall; recall of the TBR words in the presence of their strongest normative associates, which had not been seen at input, did not differ from noncued recall. (p. 255)

However, Thomson and Tulving (1970) did not force output to all test cues, rendering their measure of recall analogous to free-report retrieval, which is potentially contaminated with monitoring and report bias. To investigate whether Thomson and Tulving's results and conclusions were affected by these variables, I applied the SDT analogue model outlined above and determined the effect of cue type (weak vs. strong) on retrieval, report bias, and monitoring. Also, by comparing free- and forced-report retrieval, the contaminating effect of report bias on free-report retrieval could be assessed.

METHOD

Participants

Forty-eight undergraduates participated in the experiment either for course credit or for financial compensation. Thirty-two participants were assigned to two cued-recall groups: 16 to the phase group and 16 to the trial group. Sixteen participants were assigned to the no-cue group. The participants were tested in groups of 1–4 at individual work stations. All participants in a given session were assigned to the same experimental group.

Design & Materials

One hundred target words, each with two associated cue words, were chosen from the Edinburgh Associative Thesaurus.¹ One cue word for each target word was a strong associate of the target (mean probability of target production for all 100 strong cues = 35%), whereas the second was a weak associate (mean probability of target production for all 100 weak cues = 1%). No word was repeated across the lists of weak cues, strong cues, or targets. The 100 targets, 100 strong cues, and 100 weak cues are presented in the Appendix.

The experiment was divided into two or three phases, depending on the particular experimental group. In Phase 1, all 100 targets, along with their weak associates, were presented in pairs in a random order to all participants for study. For counterbalancing purposes in the cued-recall (trial and phase) groups, half the participants were presented in Phase 2 with strong cues to elicit targets

1–50 and were presented with weak cues to elicit targets 51–100, whereas this was reversed for the other half of the participants. This counterbalancing procedure was not necessary for the no-cue group because no associates were presented to the participants in Phase 2.

In Phase 2, the participants in the trial and phase groups were presented with retrieval cues. The participants were initially given the choice of providing a response to a given cue or leaving it blank (free report; see the Procedure section below). However, responses to cues that were left blank were obtained later in the experiment by using two forced-report methods. For the phase group, cues that were left blank were presented to the participants again in a third phase of the experiment, which commenced immediately after all 100 cues were presented in Phase 2. In the trial group, immediately after a cue was left blank, the cue was presented again and a response was required before moving on to the next trial. Thus, there were two experimental phases in the trial group, but three phases in the phase group. The order of presentation of the cues in Phase 2 was randomized for each participant. After randomization (which was different for each participant), data from the first six trials of Phase 2 in both the trial and phase groups were counted as practice trials and dropped from the analysis. Thus, analyses in these groups were based on 94 items, with an average of 47 (range, 44–50) data-points for each of the strong and weak conditions.

For the no-cue group, immediately following the study phase, the participants were asked to recall as many uppercase words from Phase 1 as possible, without the assistance of any cues. Thus, there were two experimental phases in the no-cue condition. Because there were no cues presented to the participants in the no-cue group, the effect of strong/weak manipulation could not be assessed, nor could the effect of free versus forced report. The purpose of the no-cue condition was merely to establish the participants' unassisted free-report retrieval ability, given the particular study conditions, which were held constant across all three experimental groups.

Procedure

In Phase 1, participants in all three experimental groups studied 100 weak cue–target word pairs presented individually for 3 sec each, centered on a computer monitor. The cue words were displayed in lowercase letters to the left of the target words that were presented in uppercase letters. The participants were instructed to study the uppercase words for a later memory test, but to attend to the lowercase words as possible cues to assist in recalling the uppercase words at test. These instructions were given verbally by the experimenter and also appeared on the computer monitor. The order of presentation of the word pairs was randomized.

In Phase 2, the participants in the phase group were instructed that they would be presented with words, one at a time on the computer monitor. Each cue word was centered on the monitor and was displayed with a question mark to its immediate right indicating that a response was requested. The participants were told that each word presented during this phase had an uppercase word from the training list that was related to it and were instructed to use the word as a cue to assist in recalling the uppercase word. The participants were also informed that each correct answer would earn 1 point, and each incorrect answer would cost 4 points, but that they could avoid the point system by not entering a response to the cue; blank (no response) trials would neither earn nor cost any points. If the participants chose to enter a response, they typed the response on the computer keyboard. If they chose not to respond, they entered "B" (for "blank"). If a response was entered, the participants were then asked to give a confidence rating on a separate screen, using a scale from 1 to 6, where 1 = *extremely low confident correct*, 2 = *very low confident correct*, 3 = *low confident correct*, 4 = *high confident correct*, 5 = *very high confident correct*, and 6 = *extremely high confident correct*. This scale appeared on the screen whenever a confidence rating was required in any of the experimental groups. After entering a confidence rating, the cumulative points, the points

earned/lost for that particular trial (either +1 or -4), and the number of trials remaining were displayed. After receiving feedback, the participants pressed the space bar to initiate the next trial. If "B" was entered, no confidence rating was gathered, and the program skipped to the feedback display. The feedback display for blank trials again consisted of the cumulative points, points for that trial (0), and the number of trials remaining.

The participants in the phase group completed all 100 trials before moving on to Phase 3. During Phase 3, all cues that the participants had chosen to leave blank in Phase 2 were presented again, one at a time. The participants were informed that blanks were no longer a viable response option to the cues; a response was required even if it was necessary to guess, and the experimental program would not advance to the next stage until a response was entered ("B" was no longer accepted as a response). The participants were also told, however, that responding in this phase did not affect point totals because the point system was no longer operative.² Immediately after entering a response to each cue in Phase 3, the participants were asked for a confidence rating using the same 6-point confidence scale as in Phase 2. No feedback screen was displayed during Phase 3, because point totals were unaffected by the accuracy of the responses.

The procedure for the participants in the trial group was similar to that for the phase group, except in the method used to collect responses for cues initially left blank. Instead of gathering responses to cues initially left blank, in a separate third phase of the experiment, as in the phase group, responses in the trial group were gathered on a trial-by-trial basis. As in the phase group, immediately after the participants entered a "B" to leave a trial blank, the feedback screen was displayed. However, instead of advancing to the next cue when the space bar was pressed, the same cue was presented again in the trial group, and the participants were required to enter a response, guessing if necessary. No points were at stake for the second presentation of the cue. As in the phase group, confidence ratings (1-6) were gathered for both initially reported and initially withheld responses.

Following the study phase, the participants in the no-cue group were asked to try to recall as many of the uppercase words from Phase 1 as possible. The participants entered their responses into the computer, typing each word next to a prompt (Response →), which appeared in the center of the computer monitor. After each response, a confidence rating was gathered by using the same 6-point scale as in the trial and phase groups. When the participants could remember no more words, they were instructed to type DONE, which terminated the program.

RESULTS AND DISCUSSION

Two 2×2 contingency tables, similar to that shown in Table 1, were created for each participant in the trial and phase groups. The first table corresponded to performance for strong cues and the second to performance for weak cues. Given that there were 94 valid test trials (100 in total minus 6 practice trials), the sum of the frequencies across both contingency tables for each participant was 94, with as many as 50 and as few as 44 observations for a given table. Once both contingency tables were created for each participant, it was possible to determine four performance measures from each table using equations presented in Table 1: free- and forced-report retrieval, monitoring, and report bias.³ Table 2 shows the frequencies for each table (weak-phase, strong-phase, weak-trial, and strong-trial) averaged across participants.

Before conducting the SDT analyses, the hit and false-alarm rates were adjusted according to Snodgrass and Corwin's (1988) recommendation to avoid having to eliminate participants because of undefined values (i.e., .5 was added to the numerator and 1 was added to the denominator). However, for some non-SDT analyses that did not involve hit and false-alarm rates, the participants with undefined cells were eliminated from the analysis. The exact number of participants eliminated can be determined from the degrees of freedom presented with the analyses.

Retrieval

Free- and forced-report retrieval is shown in Figure 1. The method used to collect initially withheld responses (trial vs. phase) had no effect on either the free- or forced-report retrieval measures, nor did it interact with any other variables [largest $F(1,30) = 1.27$], so the means presented are collapsed across cued-recall group. Retrieval in the no-cue condition is shown as the solid line in the figure. The pattern of results indicates that when the participants exercised the option to withhold answers (free report), their retrieval for strong cues was much worse (.07) than if reporting was forced (.26). However, retrieval for weak cues was unaffected by report option (free, .26; forced, .28). The pattern was confirmed by a 2 (group: trial/phase) $\times 2$ (cue type: strong/weak) $\times 2$ (report type: free/forced) mixed analysis of variance (ANOVA) that revealed a significant cue type (strong/weak) $\times$ report type (free/forced) interaction [$F(1,30) = 97.22$, $MS_e = 0.002$, $p < .001$]. A post hoc Tukey honest significant difference (HSD) test on the interaction revealed that performance in the free-strong cell was significantly less than performance in any of the other three conditions (free-weak, forced-strong, forced-weak; all $ps < .001$). Performance between the free-weak, forced-strong and forced-weak cells did not differ (smallest $p = .14$).

Table 2
The *a*, *b*, *c*, and *d* Cells of the 2×2 Contingency Tables for Strong- and Weak-Retrieval Cues as a Function of Experimental Group

Cue Type and Response	Candidate Answer	
	Correct	Incorrect
Phase Group		
Weak cues		
Reported	13.31	7.44
Withheld	0.94	25.19
Strong cues		
Reported	3.38	6.94
Withheld	9.13	27.69
Trial Group		
Weak cues		
Reported	10.94	9.88
Withheld	1.06	24.56
Strong cues		
Reported	3.44	8.63
Withheld	8.31	27.19

To compare the obtained results to those of Thomson and Tulving (1970), it is important to reemphasize that they treated blank responses the same as incorrect responses; that is, they used free-report retrieval as their index of memory performance. In examining Figure 1, it is apparent that the results for free-report retrieval replicated their pattern exactly; weak-cue performance was superior to no-cue performance, but strong-cue performance was not. In support of this pattern, a one-way ANOVA (group: trial/phase/no cue) comparing free-report retrieval in the weak-cue conditions (trial and phase) to retrieval in the no-cue condition revealed a significant main effect [$F(1,45) = 14.89$, $MS_e = 0.013$, $p < .001$]. A post hoc Tukey HSD test revealed that free-report retrieval was significantly greater in the weak-cue conditions (trial = .24; phase = .29) than in the no-cue condition (.08; both $ps < .001$) and that the weak-cue conditions did not differ from each other ($p = .43$). An analogous one-way ANOVA comparing free-report retrieval in the strong-cue conditions with the no-cue condition did not reveal a significant effect ($F < 1$). Thus, in free report, only weak cues elicited performance that was superior to the no-cue group.

However, a very different pattern of results is revealed if the forced-report measure is examined in Figure 1, in which report bias and monitoring could no longer contaminate performance. Now, both strong- and weak-cue retrieval were much better than no-cue retrieval. This conclusion was supported by 2 one-way ANOVAs (group: trial/phase/no cue) on the forced-report data, one for strong cues and one for weak cues. Both ANOVAs revealed significant main effects [$F(1,45) = 31.97$, $MS_e = 0.005$, $p < .001$, and $F(1,45) = 16.75$, $MS_e = 0.014$, $p < .001$, for strong and weak cues, respectively]. Post hoc Tukey HSD tests indicated that forced-report retrieval in both the strong-cue conditions (trial = .25; phase = .27) and the weak-cue conditions (trial = .26; phase = .31) was superior to no-cue retrieval (.08, all $ps < .001$). Forced-report retrieval did not differ between the strong-cue conditions in the first analysis ($p = .76$), nor did it differ between the weak-cue conditions in the second analysis ($p = .51$). Thus, when output was forced, both weak and strong cues facilitated retrieval relative to the no-cue group.

Monitoring

The model assumes that free-report retrieval will underestimate forced-report retrieval more if monitoring is low than if it is high. Thus, the weak-cue condition, in which the underestimation of forced-report retrieval was small, should have higher monitoring than the strong-cue condition, in which the underestimation was large. Consistent with this assumption, monitoring was much better for weak cues (.88) than for strong cues (.61). The reliability of this difference was confirmed with a 2 (group: trial/phase) $\times$ 2 (cue type: strong/weak) ANOVA on the monitoring measure (A'), which revealed only a significant main effect of cue type [$F(1,30) = 100.50$,

$MS_e = 0.012$, $p < .001$]. Because monitoring was good in the weak-cue condition, the participants were in a position to set their report criterion so that the vast majority of correct answers were reported. Consequently, making the report criterion more liberal by forcing report did not elicit much new correct information, so retrieval was unaffected by the free/forced manipulation. Conversely, poor monitoring in the strong-cue condition meant that many correct responses were withheld at the report criterion. Forcing responses, however, elicited these correct responses and caused retrieval to increase in the strong-cue condition. Thus, strong-cue retrieval was greatly affected by the free/forced manipulation.

Report Bias

The model also assumes that free-report retrieval will underestimate forced-report retrieval to a greater extent if the report criterion is conservative than if it is liberal. The monitoring difference between the strong- and weak-cue conditions described in the previous analysis might have been enough to account for the differential level of forced-report retrieval underestimation that was found. Nonetheless, the differential underestimation might also have been partially due to the participants' having a more conservative report criterion in the strong-cue condition than in the weak-cue condition. The results indicated that report bias, as measured by the SDT index (B''_D), was indeed more conservative in the strong-cue condition (.62) than in the weak-cue condition (−.31). A 2 (group: trial/phase) $\times$ 2 (cue type: strong/weak) mixed ANOVA on B''_D calculated for each participant revealed only a significant main effect of cue type [$F(1,30) = 85.93$, $MS_e = 0.061$, $p < .001$].

Some caution should be exerted when interpreting the cue-type effect on B''_D . Although the effect could be interpreted as a shift in the placement of the criterion between the two cue types, it is also possible that a single criterion was used for both classes of cues, but that mean confidence differed between them. If a single report criterion were placed between the means of strong- and weak-cue confidence distributions, B''_D would be conservative relative to the strong-cue distribution, but liberal relative to the weak-cue distribution. This interpretation of the difference in report bias seems more likely than one in which participants are assumed to be adopting multiple criteria. Importantly, however, debates regarding the source of the bias effect are not relevant to the testing of the report bias predictions made by the SDT model; conservative placement means high free–forced retrieval discrepancy regardless of whether there has been a “true” criterion shift or not.

Support for a difference between mean confidence ratings in the strong- and weak-cue conditions was found in a 2 (group: phase/trial) $\times$ 2 (cue type: strong/weak) $\times$ 2 (accuracy: correct/incorrect) ANOVA on the mean confidence ratings. It revealed a main effect of cue type [$F(1,30) = 312.04$, $MS_e = 0.253$, $p < .001$]; responses to weak cues were assigned higher confidence (3.41) than

Figure 1. Free- and forced-report retrieval in the strong- (nonreinstated) and weak- (reinstated) cue conditions. Retrieval in the no-cue condition is shown as the horizontal line.

answers to strong cues (1.84). The main effect of accuracy was also significant [$F(1,30) = 525.01, MS_e = 0.207, p < .001$], as was the accuracy $\times$ cue type interaction [$F(1,30) = 374.81, MS_e = 0.172, p < .001$]. Correct answers were assigned higher confidence (3.55) than incorrect answers (1.71), but this difference was greater for weak cues (+3.26) than for strong cues (+.42). This interaction is consistent with the above monitoring analysis that revealed higher monitoring in the weak-cue condition than in the strong-cue condition.

Other Measures of Monitoring and Report Bias

Nelson (1984) has advocated the Kruskal–Goodman gamma (γ) correlation as the most suitable accuracy measure of metamemory predictions (but see Schraw, 1995, and commentary by Nelson, 1996, and Wright, 1996). Similarly, Koriatic and Goldsmith (1996c) used γ correlation between confidence and accuracy as a measure of monitoring. So that the SDT measure (A') could be compared directly to γ , two γ coefficients were calculated for each participant in the trial and phase groups. The first, $\gamma - rc$, is the correlation calculated from the 2×2 contingency table presented in Table 1. The second, $\gamma - conf$, is the correlation between confidence ratings and accuracy as derived from the 2 (accuracy: correct/incorrect) $\times 6$ (confidence scale level: 1–6) contingency tables. Both can be considered as measures of monitoring, only $\gamma - rc$, like A' , is based on the reporting behavior of the participants, whereas $\gamma - conf$ is based on the confidence ratings.

Just as with the analysis on the SDT measure (A'), $\gamma - rc$ and $\gamma - conf$ were analyzed with separate 2 (group: trial/phase) $\times 2$ (cue type: strong/weak) ANOVAs. Both

ANOVAs revealed only a significant main effect of cue type [$F(1,28) = 39.93, MS_e = 0.188, p < .001$, and $F(1,30) = 63.87, MS_e = 0.128, p < .001$, for $\gamma - rc$ and $\gamma - conf$, respectively]. $\gamma - rc$ and $\gamma - conf$ were both higher in the weak-cue condition (both $M_s = 0.95$) than in the strong-cue condition (both $M_s = 0.24$). These results are consistent with the analysis of A' .

The SDT index of report bias (B_D'') was compared with Koriatic and Goldsmith's (1996c) index of bias (Prc). As before, the comparison was of interest because B_D'' is based on reporting behavior, whereas Prc is based on confidence ratings. To calculate Prc , first the number of reported and withheld candidates was determined for each level of confidence. Then five *fit ratios* (Koriatic & Goldsmith, 1996c) were calculated for each participant, representing placement of the criterion at 2 through 6. Fit ratios were determined by averaging (weighted according to the number of observations available) (1) the proportion of withheld candidates assigned confidence below a given criterion and (2) the proportion of reported candidates assigned confidence equal to or higher than the criterion. The scale value that yielded the highest fit ratio represented Prc for that participant. If more than one scale value was associated with the highest fit ratio, the mean of the values was taken as Prc .⁴ A 2 (group: trial/phase) $\times 2$ (cue type: strong/weak) ANOVA revealed a significant main effect of cue type [$F(1,28) = 5.57, MS_e = 0.596, p < .05$]; the mean Prc was more conservative in the strong-cue condition (4.03) than in the weak-cue condition (3.57). This result is consistent with the analysis on the SDT measure of report bias (B_D'').

Although both the analysis of B_D'' and Prc revealed significant effects of cue type, the effect sizes were consid-

erably different ($\eta^2 = .74$ and $.17$ for B_D'' and Prc , respectively). This difference might actually reflect dissimilarity in the nature of the information that each index is measuring. Although B_D'' is sensitive both to "true" criterion shifts between conditions and to differences in the mean subjective confidence of the underlying distributions (as explained above), Prc is only sensitive to the former. That is, Prc is designed to "hone in on" the particular subjective confidence level that determines whether or not a candidate response is reported. If one assumes that participants assign scale values consistently between free and forced report and between experimental conditions, Prc will not be influenced by differences in reporting behavior owing to confidence distribution shifts, only by changes to the subjective level of confidence associated with the criterion. Consequently, the cue-type effect size may have been smaller for Prc than for B_D'' simply because Prc is sensitive to one influence, whereas B_D'' is sensitive to two influences.

Some readers may consider it a problem that differences in B_D'' can be attributed to either criterion or distribution shifts. However, this will obviously depend on what information one is hoping to glean from the bias measure. Prc may be preferable to B_D'' if one is hoping to estimate the subjective confidence level at which the report criterion is set. But B_D'' may be preferable if one is simply wanting a summary statistic of the prevalence of a particular type of behavior to a given set of items (e.g., the tendency to respond "yes" in traditional SDT or the tendency to report candidates in the present treatment). Furthermore, it is not clear what the relationship is between Prc and free-/forced-report retrieval discrepancy. Suppose, for example, that Prc increases from experimental Condition A to B. If this were true of B_D'' , and monitoring did not vary, the prediction is clear: The free-/forced-retrieval discrepancy should be greater for Condition B than for Condition A. However, this is not necessarily true for Prc . For example, if the means of the underlying incorrect and correct candidate distributions also increase from Condition A to B by approximately the same amount as Prc (again, keeping monitoring constant between the conditions), the free-/forced-retrieval discrepancy will be invariant between the conditions (i.e., the number of misses—cell c —will not vary). In short, it is the prevalence of reporting (B_D'') that is important when generating report-bias-related predictions, not the confidence scale value that marks whether or not candidates will be reported (Prc).

There are other reasons to use the SDT bias index instead of Prc : B_D'' is not subject to some of the problems that are inherent with measures such as Prc , where there is an attempt to map subjective confidence to external scale values. Consider, for example, the phase group. The participants in this group assigned confidence ratings to initially withheld answers in a forced-report block at the end of the experiment, consisting wholly of the cues that were left blank in an earlier phase of the ex-

periment. Most of these cues were likely to elicit mostly incorrect candidate answers in a limited range of (low) subjective confidence. This would likely lead the participants to redistribute the six objective confidence scale values over this limited range of low subjective confidence, rather than strictly adhering to using only scale values below the report criterion. Should this occur, the meaning of Prc as a measure of the report criterion becomes questionable.

In summary, the SDT measure of monitoring (A'), the gamma correlation based on the report criterion ($\gamma - rc$), and the gamma correlation based on confidence ratings used by Koriatic and Goldsmith (1996c; $\gamma - conf$) all indicated better monitoring in the weak-cue condition than in strong-cue condition. Also, the SDT measure of report bias based on reporting behavior (B_D'') and Prc based on fit ratios weighed according the number of reported and withheld responses, both indicated that the bias was more liberal for weak cues than for strong cues. Consistent with both these findings, free-report retrieval underestimated forced-report retrieval more in the strong-cue condition than in the weak-cue condition. This underestimation of retrieval was sufficiently great for strong cues that, when output was forced, they facilitated performance relative to no cues, a result that is at odds with the conclusions of Thomson and Tulving (1970).

GENERAL DISCUSSION

In the present research, a model analogous to those seen in SDT was used to estimate retrieval, report bias, and monitoring in Thomson and Tulving's (1970) classic cued-recall task. The effect of retrieval cues, when participants were given the option of omitting responses to cues, replicated the effect found in Thomson and Tulving, who also allowed omissions: Reinstating the same cues at retrieval that were encoded specifically with to-be-remembered material facilitated memory performance more than did nonreinstated cues. Indeed, under these reporting conditions, nonreinstated cues did not facilitate memory performance *at all* relative to no-cue recall. However, because Thomson and Tulving did not force participants to provide answers to all the cues, their measure of memory performance was free-report retrieval as it is defined here. The SDT model assumes that free-report retrieval will underestimate forced-report retrieval to the extent that report bias is conservative and monitoring is poor. The results showed both of these conditions were met for strong cues, and, as expected, the free-forced retrieval difference was greater in the strong-cue condition than in the weak-cue condition. Only 28% of the correct answers potentially produced to strong cues were actually offered in free report, as compared with 92% for weak cues.

Consistent with the results obtained here, other researchers have found ways of improving performance to strong, nonreinstated cues using Thomson and Tulving's

(1970) paradigm. However, some of these manipulations might also have been confounded with report-bias differences, just as Thomson and Tulving's manipulation of cue type was. For example, Santa and Lamwers (1974, Experiment 1) replicated a portion of Thomson and Tulving's methodology by having the critical lists in the no-cue and strong-cue conditions preceded by practice study-test lists. Each practice study list consisted of weak paired associates, and each practice test list consisted of reinstated weak cues and new, unrelated distractor cues. The final, critical study list of weak associates was followed by a recall phase with no cues or one with a mixture of nonreinstated strong cues and unrelated distractor cues. They reasoned that participants performed poorly in the strong-cue condition of Thomson and Tulving's experiments because the instructions failed to warn participants that correct responses on the final list were strong associates of the cues, not weak associates as in the previous lists. Without the warning, Santa and Lamwers argued, participants might continue to generate weak associates to the cues on the final critical list and, consequently, respond incorrectly in the majority of cases.

To alleviate the problem of the surprise switch from weak to strong associates in the final list for participants in the strong-cue condition, Santa and Lamwers (1974) told participants in the warning group that the cues were not presented in the study list, but that some of them were strong associates of the target words. This small inclusion had a dramatic effect on performance. Replicating Thomson and Tulving's (1970) results, the mean number of words recalled for participants provided with strong cues and no warning was equivalent to performance in the no-cue condition ($M_s = 5.33$ and 5.40 , respectively). However, for participants given strong cues and told that the correct responses were strong associates of the cues, performance jumped to 10.07 words recalled. Thus, Santa and Lamwers' results suggest that it is critical to warn participants of the nature of the association between the cues on the test list and the target words so that they can adjust their mental set for the final, critical list.

However, before accepting the mental set interpretation out of hand, it is worth noting that their warning manipulation was confounded with a difference in report bias. The mean responses to distractor cues was $.33$ in the strong-cue-no-warning group, which more than doubled to $.80$ in the strong-cue-warning group. A shift in report bias was shown, in the present experiment, to be associated with a boost in performance to strong cues, without the need for any warning. Thus, it is conceivable that Santa and Lamwers's instruction was effective simply because it varied participants' report criterion, not because it changed their mental set.

There are two important points to be made about the reinterpretation of Thomson and Tulving's (1970) and Santa and Lamwers's (1974) results in terms of the SDT

framework. First, it is clear that not all researchers are considering report bias and its effect on indices of recall performance in their experiments. For example, had Santa and Lamwers been studying recognition, the difference in bias would likely have been an important consideration because of potential distortions to indices like percent correct recognition. However, in the context of recall, report bias and its potential to undermine even major conclusions of the research is not even discussed. Second, the reinterpretation shows that the present framework, in which report bias, retrieval, and memory parameters were considered separately, is both feasible and potentially enlightening when applied to experimental data. The framework offers a synthesis of memory, metamemory, and decision-making mechanisms, which is much needed in the recall literature (see Roediger & Payne, 1985).

Appropriateness of Measures

Confidence ratings were used to derive $\gamma - conf$ and Prc , measures for monitoring and report bias, respectively, advocated by Koriart and Goldsmith (1996c). In contrast, two measures of monitoring (A' and $\gamma - rc$) and a measure of report bias (B_D'') were dependent on the participants' reporting behavior—the distribution of responses across the four cells of the 2×2 contingency table shown in Table 1. With the present paradigm, measures based directly on reporting behavior are probably preferable to those based on confidence ratings because of the potential for inconsistency in the mapping between subjective confidence and the objective scale values between free and forced report.

With the methodology advocated here, confidence data were gathered when responses were offered, meaning that some confidence data were collected in forced report and some in free report. In contrast, Koriart and Goldsmith (1996c) gathered all confidence data from participants during a forced-report stage of the experiment and assumed that both the generated candidates and their associated confidence would be the same as at free report. Their methodology is more likely to lead to consistent application of the scale values than is the present methodology because the confidence data were all collected during the same phase of the experiment. However, even with their methodology, they found that a few responses to the same test questions did not match between free and forced report, and these responses had to be eliminated from the analysis. Furthermore, even for the candidates that were similar between free and forced report, there is no guarantee that the subjective confidence level in those candidates was the same. These cases are more tricky to pinpoint because Koriart and Goldsmith (1996c) did not have participants provide confidence at free report. With the general knowledge questions that they used, the candidates generated and their associated confidence might be fairly stable. However, with other materials intended to test episodic mem-

ory, such as the paired associates used here, it is not clear how stable either candidate generation or confidence might be across time.

For these reasons, it is recommended that the present procedure be used for gathering candidate responses and confidence data, as long as confidence rating-based indices of performance are avoided. But apart from the problems associated with using confidence data to derive performance indices, there are other reasons to prefer indices based on reporting behavior. What could be a more direct measure of participants' monitoring of the reportability of retrieved candidates than their decisions regarding whether or not to report them? Confidence ratings are peripheral to this issue; in fact, the SDT measures can, and perhaps should, be derived without considering confidence data at all.

Type 1 Versus Type 2 Signal Detection Models

Some readers may recognize the similarity between the SDT model used in this study and early, so-called Type 2 SDT models (e.g., Banks, 1970; Clarke, Birdsall, & Tanner, 1959; Healy & Jones, 1973; Lockhart & Murdock, 1970; Murdock, 1966, 1974). The main differences between Type 1 and Type 2 SDTs are the methods by which the underlying item distributions are generated and the nature of the dimension over which items are distributed. In Type 1 (stimulus contingent) SDT, the experimenter generates the items by choosing them for the recognition experiment, and the items are assumed to be distributed over *activation*, *strength*, or *familiarity*. On the other hand, in Type 2 (response contingent) SDT, the items are generated by participants when cues initiate a search of long-term memory. Furthermore, generated items are assumed to be distributed over *confidence*, not *strength*. Although application of Type 1 SDT has been widespread in research on recognition memory, the use of Type 2 SDT with recall data has been limited. This may be partly attributable to the fact that there has been some confusion in the literature in regard to the distinction between Types 1 and 2 SDT, which have led to some inappropriate criticism of Type 2 SDT (see discussion and examples of this confusion in Healy & Jones, 1973). For example, if it is erroneously assumed that the underlying dimension over which items are distributed is *strength* for both Type 1 and Type 2 SDTs, then discrimination reflects differences in the *strength* between old and new items. Thus, under this assumption, manipulations thought to enhance the *strength* of old items should also enhance Type 2 discrimination. However, because discrimination more accurately represents a metacognitive index with Type 2 SDT, manipulations that enhance memory for items (retrieval) should not (and do not) necessarily enhance discrimination (monitoring).

To illustrate, consider an experiment by Murdock (1966). He presented participants with short lists of unrelated paired associates. Following list presentation, one of the cues was presented, and participants were re-

quired to provide the associated target, guessing if necessary, and to give a confidence rating. He found a large effect of serial position on retrieval; the proportion of targets correctly provided in forced response to cues was .29, .29, .44, .59, and .85 for Serial Positions 1–5, respectively. However, the effect of serial position on discrimination (as derived from confidence ratings) was negligible. That is, to use the present terminology, serial position affected forced-report retrieval but not monitoring.

Murdock's (1966) data are at odds with a strength interpretation of discrimination. In his experiment, more recent items were more retrievable, so they should also have had more *strength* (higher d'). If, instead, the Type 2 discrimination index is more appropriately interpreted as a metacognitive index of monitoring rather than of *strength*, Murdock's results are understandable; serial position affected participants' ability to retrieve correct candidate responses, but did not affect the relationship between assigned confidence and accuracy. Although there may be situations in which items may be both more retrievable at forced report and monitored better, rendering a strength interpretation of Type 2 discrimination plausible, the present study, and that of Koriatic and Goldsmith (1996c), Murdock, and others, suggests that the two memory parameters are dissociable. As Lockhart and Murdock (1970) have stated, "It is gross oversimplification to say that the Type 2 d' is a measure of . . . strength" (p. 107).

A second reason that Type 2 SDT may not have been adopted in the past may relate to the different ways that the item distributions are generated. Because participants generate their own candidates, with the best candidates' being distributed across confidence, it is not clear what the nature of the underlying distributions might be like. For example, monitoring may be very high for some materials, meaning that participants' responses in free report are all correct and are assigned very high confidence, whereas those in forced report are all incorrect and are assigned very low confidence. Such materials would produce highly skewed distributions, peaking at the extremes of the confidence dimension. The distributions associated with these materials would render parametric indices such as d' and β unsuitable because they assume normal underlying distributions of equal variance. However, this is more of a problem with using d' and β than it is a criticism of Type 2 SDT more generally. To circumvent assumptions regarding the nature of the underlying distributions in the present study, I used a nonparametric index of discrimination (A' , which represents the area under the receiver operating characteristic curve) and its associated measure of bias (B''_D ; Donaldson, 1992). Until more is known about the nature of the underlying distributions, nonparametric statistics are probably the safer bet.

With the growing interest in metacognitive processes over the last few decades and the need for analytical tools to estimate and control report bias in cued-testing paradigms, the value of Type 2 SDT models might be more readily accepted today. Certainly there is no need

to dismiss them out of hand simply because their interpretation is inconsistent with more traditional Type 1 SDT models. As Healy and Jones (1973) have stated, "... Type II analysis can illuminate the important evaluation processes which go on in recall and which may affect the measure of memory loss in such experiments" (p. 340).

The "Lucky Guessing" Criticism and the Status of Encoding Specificity

If memory researchers were to be surveyed and asked to list the most important principles of memory, it is likely that the encoding specificity principle would come out on top. The importance of reinstating the same retrieval cues at test, which were encoded specifically with the to-be-remembered material during study, has been demonstrated in many different domains with a wide variety of materials. Certainly, it is not being argued that the present experiment undermines this general principle of memory. However, I *am* arguing that there are flaws in both the design and interpretation of Thomson and Tulving's (1970) original experiments with strong and weak cues that generated interest in the encoding specificity principle in the first place.

One obvious criticism of using forced-report retrieval is that participants in the forced-report condition may have responded correctly to several cues because of memory for *pre-experimentally derived associations* (PEDAs); that is, a given candidate may have been simply the first word related to the cue that "came to mind" and would have been produced even by participants not given any study list. This "contamination" of the retrieval measure with PEDAs in the forced-report condition is likely to have been particularly great with strong cues because the normative probability of producing a target word was, by definition, higher than with weak cues. A critic might argue, therefore, that Thomson and Tulving's (1970) choice of free-report retrieval is preferable to forced-report retrieval because participants are not affected by PEDAs in free report, which only serve to confound the strong-/weak-cue comparison.

Before one accepts this argument out of hand, however, I think it is important to consider its assumptions. First, it assumes that participants are able to more or less perfectly monitor retrieved information that derives from a pre-experimental source; that is, participants withhold all candidates that are retrieved because of PEDAs in the strong-cue/free-report condition. Tulving and Thomson (1973) explicitly supported this assumption by suggesting that "guessing from semantic memory" occurred rarely if free-report retrieval was used as the memory index (see, e.g., their discussion on p. 355). However, the argument also assumes that participants more or less perfectly monitor the retrieval of episodic content derived from the study list; that is, participants withhold none of the experimentally derived candidates that they retrieve in the strong-cue/free-report condition. Thomson and

Tulving implicitly supported this assumption by treating free-report omissions as incorrect responses. By counting omissions as errors, the assumption must be that if responses were gathered to cues initially left blank, they would all have been wrong—either the wrong word would have been produced, or the right word would have been produced, which was not really right because it was a "lucky guess" (i.e., not "true" episodic retrieval).

This latter assumption is particularly questionable. First, the extra search time and effort induced by forcing report from participants to strong cues is likely to have produced at least some new experimentally derived candidate memories. This is particularly likely given the large number of correct answers that were withheld to strong cues in free report in the present study (i.e., poor monitoring and conservative report bias). Additionally, Bahrck (1969, 1970) found that nonreinstated cues consistently elicited target words in forced report, across a wide range of associative strengths, with greater probability if the target words were previously studied in a paired associate study list than if no paired-associate list was studied. That is, the probability of the target words' being elicited by nonreinstated cues, after participants studied a paired-associate list containing the target words, was higher than the normative probability of the targets' being generated by the same cues by using PEDAs. The results of Bahrck's research suggest that nonreinstated cues that are strong associates of the target words result in memory of targets derived from encoding the experimental study list, not just from PEDAs.

It is necessary to point out, however, that this "lucky guessing" criticism speaks only to whether or not the present results undermine the encoding specificity principle, which was never the intended purpose of the study. The present study was conducted in order to demonstrate the problems that can arise by failing to consider such issues as monitoring and report bias when using free-report indices of retrieval and to use Thomson and Tulving's (1970) original research as an example. At the very least, their use of free report carries with it some very strong, unwarranted assumptions regarding the nature of withheld information, which could only have obtained support if output was forced. Thus, Thomson and Tulving's experiments, by themselves, provide, at best, weak evidence for the encoding specificity principle, despite the fact that they are considered classic and are cited in textbooks throughout the world as providing its very foundation. This remains true regardless of whether or not the correct withheld candidates in free report were "lucky guesses" or "true" episodic retrieval.

Today, there are various analytical tools available to cognitive psychologists (e.g., the remember/know paradigm, Higham, 1998; Tulving, 1985; the independent scales technique, Higham & Vokey, 2002; the process dissociation paradigm, Jacoby, 1991; and the opposition paradigm, Higham, Vokey, & Pritchard, 2000; Jacoby, Woloshyn, & Kelley, 1989) that allow for the separation

of different sources of retrieval. Currently, we are investigating the nature of the reported and withheld information in the strong-/weak-cue paradigm using some of these tools. Hopefully, this line of research will allow us to determine the nature of the correct withheld information and more directly assess the role of encoding specificity in this paradigm.

REFERENCES

- BAHRICK, H. P. (1969). Measurement of memory by prompted recall. *Journal of Experimental Psychology*, **79**, 213-219.
- BAHRICK, H. P. (1970). Two-phase model for prompted recall. *Psychological Review*, **77**, 215-222.
- BANKS, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, **74**, 81-99.
- CLARKE, F. R., BIRDSALL, T. G., & TANNER, W. P. (1959). Two types of ROC curves and definitions of parameters. *Journal of the Acoustical Society of America*, **31**, 629-630.
- DONALDSON, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General*, **121**, 275-277.
- DYWAN, J., & BOWERS, K. (1983). The use of hypnosis to enhance recall. *Science*, **222**, 184-185.
- FISHER, R. P. (1996). Misconceptions in design and analysis of research with the cognitive interview. *Psychology*, **7**(35), witness-memory.12.fisher.
- FISHER, R. P., & GEISELMAN, R. E. (1992). *Memory enhancing techniques for investigative interviewing: The cognitive interview*. Springfield, IL: Thomas.
- GRIER, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, **75**, 424-429.
- HEALY, A. F., & JONES, C. (1973). Criterion shifts in recall. *Psychological Bulletin*, **79**, 335-340.
- HIGHAM, P. A. (1998). Believing details known to have been suggested. *British Journal of Psychology*, **89**, 265-283.
- HIGHAM, P. A., & ROBERTS, W. T. (1996). Measuring recall performance. *Psychology*, **7**(38), witness-memory.13.higham.
- HIGHAM, P. A., & VOKEY, J. R. (2002). *Illusory recollection*. Manuscript submitted for publication.
- HIGHAM, P. A., VOKEY, J. R., & PRITCHARD, J. L. (2000). Beyond task dissociations: Evidence for controlled and automatic influences in artificial grammar learning. *Journal of Experimental Psychology: General*, **129**, 457-470.
- JACOBY, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language*, **30**, 513-541.
- JACOBY, L. L., WOLOSZYN, V., & KELLEY, C. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General*, **118**, 115-125.
- KLATZKY, R. L., & ERDELYI, M. H. (1985). The response criterion problem in tests of hypnosis and memory. *The International Journal of Clinical & Experimental Hypnosis*, **33**, 246-257.
- KORIAT, A., & GOLDSMITH, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General*, **123**, 297-315.
- KORIAT, A., & GOLDSMITH, M. (1996a). Memory as something that can be counted versus memory as something that can be counted on. In D. Hermann, C. McEvoy, C. Hertzog, P. Hertel, & M. Johnson (Eds.), *Basic and applied memory research: Practical applications* (Vol. 2, pp. 3-18). Hillsdale, NJ: Erlbaum.
- KORIAT, A., & GOLDSMITH, M. (1996b). Memory metaphors and the real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioural & Brain Sciences*, **19**, 167-228.
- KORIAT, A., & GOLDSMITH, M. (1996c). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, **103**, 490-517.
- KORIAT, A., GOLDSMITH, M., & PANSKY, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology*, **51**, 481-537.
- LOCKHART, R. S., & MURDOCK, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, **74**, 100-109.
- MEMON, A., & HIGHAM, P. A. (1999). A review of the cognitive interview. *Psychology, Crime & Law*, **5**, 177-196.
- MEMON, A., & STEVENAGE, S. V. (1996). Interviewing witnesses: What works and what doesn't? *Psychology*, **7**(6), witness-memory.1.memon.
- MURDOCK, B. B. (1966). The criterion problem in short term memory. *Journal of Experimental Psychology*, **72**, 317-324.
- MURDOCK, B. B. (1974). *Human memory: Theory and data*. New York: Wiley.
- NELSON, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, **95**, 109-133.
- NELSON, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item: Comments on Schraw (1995). *Applied Cognitive Psychology*, **10**, 257-260.
- ROBERTS, W. T., & HIGHAM, P. A. (in press). Selecting accurate statements from the cognitive interview using confidence ratings. *Journal of Experimental Psychology: Applied*.
- ROEDIGER, H. L., III, & PAYNE, D. G. (1985). Recall criterion does not affect recall level or hypernesia: A puzzle for generate/recognize theories. *Memory & Cognition*, **13**, 1-7.
- SANTA, J. L., & LAMWERS, L. L. (1974). Encoding specificity: Fact or artifact? *Journal of Verbal Learning & Verbal Behavior*, **13**, 412-423.
- SCHRAW, G. (1995). Measures of feeling-of-knowing accuracy: A new look at an old problem. *Applied Cognitive Psychology*, **9**, 321-332.
- SNODGRASS, J. C., & CORWIN, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, **117**, 34-50.
- 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.
- WRIGHT, D. B. (1996). Measuring feeling of knowing: Comment on Schraw (1995). *Applied Cognitive Psychology*, **10**, 261-268.

NOTES

1. At the time the materials were being prepared, the Edinburgh Associative Thesaurus was available on line at <http://www.cis.rl.ac.uk/proj/psych/eat.html>. However, a more recent check indicated that the thesaurus is no longer available. For the reader's convenience, the target words, weak associates, and strong associates are listed in the Appendix.
2. The point system was dropped at forced report in order to keep the participants motivated during the task. I suspected that if the participants chose to withhold a response to avoid penalty, but were then forced to respond to the cue and be penalized anyway, that they would find the task pointless and stop trying to perform well. Incentive was also dropped in forced report in Koriat and Goldsmith's (1996c) design.
3. A response was considered correct in the cued-recall (trial and phase) groups only if the associated target was given in response to either the strong or weak cue. Targets associated with other cues were not counted as correct. However, because there were no cues presented in the no-cue group, this constraint could no longer be applied; instead, any target produced anywhere in the free-recall list was counted as a correct response, given that it was not repeated. Thus, retrieval in the no-cue group was defined as the number of correctly recalled targets divided by 100 (the number of possible retrievable targets).
4. The scale value "1" was not included in the analysis because all (1.0) reported items are assigned confidence "1" or higher, and none (0.0) of the withheld items are assigned a confidence value less than "1," meaning that it always had a fit ratio of .50.

APPENDIX
The 100 Target Words, 100 Strong Associates, 100 Weak Associates, and the
Probabilities of Target Production for Each Strong and Weak Associate

Target Words	Strong Associates	Weak Associates	$P(\text{Target/Strong})$	$P(\text{Target/Weak})$
loft	attic	pigeon	.15	.01
secret	agent	societies	.19	.01
plane	air	elevation	.10	.02
clock	alarm	antique	.34	.01
ship	anchor	nelson	.28	.01
money	bill	dealer	.22	.01
knife	blade	arrow	.28	.01
James	bond	Walter	.25	.01
read	book	carrots	.18	.01
beer	bottle	tanks	.17	.01
brigade	charge	postal	.06	.01
cheddar	cheese	scotch	.12	.01
hammer	chisel	tilt	.41	.01
tea	coffee	cabby	.39	.01
hot	cold	footed	.34	.02
Russia	communism	tractor	.20	.01
state	condition	union	.19	.01
goods	consumer	textile	.28	.01
food	cook	warmer	.31	.01
law	criminal	family	.16	.01
hill	dale	chalk	.26	.02
today	date	topical	.14	.01
owe	debt	vow	.18	.01
sea	deep	nauseous	.18	.01
life	death	equation	.19	.01
fragile	delicate	spindly	.11	.01
same	equal	parrot	.15	.01
continent	Europe	metric	.15	.01
precise	exact	tidiness	.33	.01
prison	cell	quilt	.42	.01
necessary	essential	unable	.33	.01
everlasting	eternal	procession	.15	.01
blue	eyes	bells	.20	.01
fiction	fact	bible	.42	.01
true	false	hollow	.32	.01
down	fallen	abdicate	.11	.01
last	first	fifth	.51	.01
blood	donor	bats	.88	.01
leader	follow	wisdom	.16	.01
kind	generous	various	.31	.01
gap	generation	void	.20	.01
lady	gentleman	seventy	.38	.01
hand	glove	torch	.66	.01
white	black	pastry	.58	.01
transplant	heart	surgery	.18	.01
neck	giraffe	snob	.34	.01
enemy	foe	watchtower	.42	.01
light	neon	wick	.60	.01
soft	hard	option	.44	.01
low	high	wages	.60	.01
glory	splendor	wings	.12	.01
home	address	farms	.38	.01
thought	idea	impulse	.23	.01
guilty	verdict	wary	.33	.01
wrong	incorrect	cane	.62	.01
god	almighty	Venus	.80	.01
knowledge	knowing	unsure	.16	.01
steel	girders	submarine	.33	.01
bird	nest	gibbon	.50	.01

APPENDIX (Continued)

Target Words	Strong Associates	Weak Associates	<i>P</i> (Target/Strong)	<i>P</i> (Target/Weak)
diamond	jewel	rare	.17	.01
paper	journal	darts	.34	.01
murder	homicide	villain	.35	.01
queen	royalty	visit	.51	.01
door	hinge	van	.67	.01
tie	knot	weld	.34	.01
sheep	flock	valley	.76	.01
found	lost	jungle	.61	.01
rung	ladder	outer	.19	.01
woman	man	ailing	.67	.01
cow	moo	dumb	.71	.01
early	late	unready	.35	.01
small	large	imply	.45	.01
least	minimum	total	.14	.01
mistress	master	sir	.26	.01
old	new	earl	.64	.01
day	night	feast	.53	.01
one	unity	track	.33	.01
tree	oak	napalm	.67	.01
subject	predicate	undergo	.27	.01
duty	obligation	unfit	.13	.01
even	odd	were	.27	.01
present	past	brain	.41	.01
fear	dread	flutter	.63	.01
fail	pass	list	.21	.01
sentence	phrase	verb	.24	.01
explosion	population	stress	.32	.01
bed	mattress	crimson	.61	.01
meat	butcher	keep	.48	.01
back	front	weeding	.58	.01
army	salvation	serge	.73	.01
plan	scheme	systematic	.45	.01
electric	shock	scope	.21	.01
wild	tame	eerie	.24	.01
nervous	tense	unrest	.19	.01
you	thank	lad	.74	.01
over	under	leper	.27	.01
bottom	top	fidget	.51	.01
handy	useful	umbrella	.15	.01
cleaner	vacuum	soot	.43	.01
rubbish	dump	poetry	.45	.01

(Manuscript received July 20, 1999;
revision accepted for publication October 27, 2001.)