Memory & Cognition
2006, 34 (3), 538-549

Determinants of retrieval solutions during cognitive skill training: Source confusions

SERGE V. ONYPER, WILLIAM J. HOYER, and JOHN CERELLA
Syracuse University, Syracuse, New York

Diverse outcomes, both facilitative and disruptive, have been reported for the effect of interpolated item recognition tests on the acquisition of a cognitive skill. We collected data from a repeated set of 12 artificial arithmetic problems, soliciting compute/retrieve strategy reports after every trial. In one condition, a recognition test was administered after every three blocks of training. Recognition testing was found to depress retrieve frequencies in both younger and older adults, particularly for newly acquired items. Pairing training items with similar recognition foils mitigated these effects. This pattern of results could be explained by assuming that the participants based compute/retrieve decisions on item familiarity or frequency, tracked across both skill trials and recognition trials, and on a threshold influenced by source confusion. Variations in the threshold parameter could lead to depressed reports of item retrieval (our findings) or to elevated retrieval decisions, as has been shown in some other studies.

The development of skilled performance in natural settings relies on an individual's accumulated knowledge of reoccurring instances (see, e.g., Loft, Humphreys, & Neal, 2004; McCarley, Kramer, Wickens, Vidoni, & Boot, 2004). Participants trained in the laboratory on artificial computational skills such as alphabet arithmetic show a similar development. As the same problems are encountered repeatedly, solution times for both easy and hard problems converge to a consistent, short duration, suggesting that memory retrieval has supplanted computation as the means for obtaining solutions (e.g., Barrouillet & Fayol, 1998; Logan, 1988). If participants are questioned about their responses, they report just such a shift, from "I computed the answer" to "I remembered the answer" (Compton & Logan, 1991; Rickard, 1997).

Logan (1988) presented a quantitative model of skill acquisition that accounted for many aspects of the shift from computation to retrieval, such as changes in the probability and the duration of retrievals with repetitions. The basic mechanism driving skill formation was the creation of problem–answer associations in episodic or long-term memory, which was postulated to occur automatically (see Rickard, 1997, for a related account; differences between the two models will be considered below).

In 1994 (Experiment 1), Ackerman and Woltz reported results that challenged *bottom-up*, *strategy-free* models. Specifically, in a noun–noun table lookup task, they found that a third of their participants never switched to consistent retrieval. This suggested that the memory strategy might have been optional, not obligatory, contrary to Logan's (1988) theory. In a second condition (Experiment 3), they found that recognition tests interspersed between skill-training blocks reduced the number of nonretrievers from a third to a fifth. Ackerman and Woltz conjectured that the imposition of the recognition tests served to heighten awareness of the memory strategy, leading more participants to retrieve, instead of relying on table lookup. These and other findings have implicated the role of *top-down* factors in skill acquisition. The picture emerging from these studies is that exposure to a skill domain poses a *problem-solving* challenge and that participants actively try to discover and deploy strategies that minimize the overall demands of the task (Ackerman & Woltz, 1994; Haider & Frensch, 2002; Siegler & Shipley, 1995).

In the present study, we looked at one of the pieces of this emergent picture, the effect of recognition testing on concurrent skill acquisition. Ackerman and Woltz (1994) found that such testing increased the frequency of retrieval solutions in skill trials. The finding suggested that a memory strategy, rather than being an invariable outcome of repetition training, was an option elected by some participants (perhaps by all those who discovered it) and declined by others (see also Rogers & Gilbert, 1997; Rogers, Hertzog, & Fisk, 2000; Touron & Hertzog, 2004). But this is not the only explanation of the available findings. Recognition testing necessarily involves additional exposures to the training problems—in effect, boosting their repetition count. If there is an automatic, repetition-driven process responsible for the emergence of retrieval solutions in untested participants, this process may simply be accelerated in tested participants. Results from an experiment by Palmeri (1997) were interpreted in just this way.

This research was supported by Research Grant AG11451 from the National Institute on Aging. We thank Greg Mangan and Korena Onyper for assistance with data collection and Patrick Lemaire and Tim Rickard for constructive comments on an earlier draft. Correspondence should be addressed to W. J. Hoyer, Department of Psychology, Syracuse University, Syracuse, NY 13244-2340 (e-mail: wjhoyer@syr.edu).

Note—This article was accepted by the previous editorial team, when Colin M. MacLeod was Editor.

Participants reported the numerosity of repeated random-dot patterns. Rather than introducing recognition trials in comparison with a baseline condition, Palmeri introduced additional training patterns that were visually similar to a target pattern and mapped to the same response category. Acquisition of the target was accelerated, as compared with baseline, apparently because of the features it shared with the companion patterns.

The argument here is an intuitive one: When cued with a target item, memory associations are activated not only to that item, but also to similar items; thus, a target benefits from its own occurrences and also from related occurrences. Reder and colleagues (Reder & Ritter, 1992; Schunn, Reder, Nhouyvanisvong, Richards, & Stroffolino, 1997) have reported results of the same sort. In Schunn et al., participants were trained on repeated multiplication problems (of the form A * B = ?, where A and B are two-digit numbers) and *sharp arithmetic* problems (of the form C # D = ?, where # is an artificial operator of about the same degree of computational difficulty as two-digit multiplication). The measure of interest to these investigators was repetition-based increases in the frequency of participants' estimates that they would be able to retrieve, and the findings were striking: Retrieval estimates were predicted better by the combined frequencies of the A and B operands than by their conjoined frequencies. That is, the participants were inclined to report *retrieve* to the problem A * B (or C # D) if both operands were high frequency (achieved through other pairings, A * E, F * B, C # G, H # D, etc.), even though the composite problem (A * B or C # D) was rare. Schunn et al. drew two conclusions from this work: Presolution decisions to compute or retrieve were based on the familiarity of the problem, and the familiarity of the whole was based on the familiarity of the parts.

Here, then, is an alternate account of the facilitative effect of recognition testing on skill learning: Additional exposures to targets on recognition trials elevate their familiarity, which in turn induces more retrieval attempts on training trials. This account is strategy free, based on frequency tracking, a process characterized as being largely automatic and effortless (see, e.g., Ofen-Noy, Dudai, & Karni, 2003; Zacks & Hasher, 1982).

Reder and colleagues observed that the tendency to track frequency was strong enough to cross over different exposure contexts (*source confusions*): A retrieve report would be elicited to A * B even if it had acquired its familiarity from strategy probes, rather than from training trials, or from training on A # B, rather than training on A * B. As a consequence of source confusion, participants were consistently lured into attempting retrievals for problems for which they did not know the answers. In such cases, incorrect responses were generated (responses that Woltz, Gardner, & Bell, 2000, termed "strong-but-wrong"), or a corrective computation was undertaken after a failed retrieval (signaled by an inflated response time).

The source activation confusion model of strategy selection discussed by Schunn et al. (1997) is built on frequency (and recency) tracking, coupled with a threshold parameter such that only items whose familiarity exceeds the threshold trigger a retrieve. In fitting their model to individuals, Schunn et al. observed that threshold differences accounted for a large amount of between-subjects variance. Note also that in Siegler and Shipley's (1995) model of arithmetic skills, a homologous parameter played a similar role in explaining individual differences (see also Siegler & Lemaire, 1997). In particular, some participants apparently set a high retrieval threshold, in order to minimize the possibility of responding on the basis of a false sense of knowing produced by item familiarity.

The retrieval threshold can be viewed as a top-down influence on skill acquisition—an influence that is sharply defined and highly circumscribed but, nonetheless, of great import. In particular, and returning to the consequences of concurrent recognition testing, elevated retrievals may not be the only possible outcome. If participants in the tested group recognize the potential for confusion and raise their retrieval criterion to guard against false alarms, the result may be depression (i.e., fewer retrieves). Indeed, in another of Palmeri's (1997) numerosity conditions, similar patterns were introduced whose numerosity differed from, rather than matched, a target pattern. The related patterns were "enemies," as opposed to "friends," in Palmeri's terminology. In that condition, acquisition of the target was delayed, rather than facilitated, due to the increased difficulty of meeting a retrieval criterion. In short, from this joint frequency-tracking-plus-threshold perspective, recognition testing may be associated with a spectrum of outcomes, both facilitative and disruptive. We will show that the published reports in this area do exhibit a spectrum of outcomes.

EXPERIMENT 1A

In an experiment by Hoyer, Cerella, and Onyper (2003), participants were trained on a fixed set of alphabet–arithmetic problems, and compute/retrieve strategy reports were elicited after every trial. (The focus of that experiment was the effect of item difficulty on retrieval, which is not pertinent to the present question.) Recognition tests were inserted between every three blocks of skill training. Contrary to Ackerman and Woltz's (1994) finding, there was no sign of a facilitation effect due to recognition testing in the data from college-aged adults. And the data from older adults showed the opposite result, a transient disruption of skill learning due to the recognition testing. That is, there was a sharp drop in retrievals after each recognition test, followed by a rapid recovery over the next couple of blocks. Given the opposite outcome reported by Ackerman and Woltz, Experiment 1A was undertaken to test the generalizability of our previous findings. In Experiment 1A, we assessed the effects of recognition testing on a skill task different from the noun–noun table lookup task used by Ackerman and Woltz and from the alphabet arithmetic task used in Hoyer et al. In the present experiment, we used an artificial arithmetic task akin to that used by Reder and Ritter (1992) and Rickard (1997).

Let us suppose that disruptive effects are obtained in the present experiment. We have argued that these effects may

reflect a raised retrieval threshold, elevated in reaction to confusions between training trials and testing trials. This interpretation is consistent with previous findings showing more conservative response criteria for older adults in a variety of tasks (e.g., Ratcliff, Spieler, & McKoon, 2000; Strayer & Kramer, 1994). One testable prediction of this interpretation is as follows. Item dropout (operationalized as the number of items retrieved in the training block preceding a recognition test and not retrieved in the block following the test) should be selective: lower familiarity items should be more susceptible to dropout. To evaluate this prediction, item familiarity was defined behaviorally: For any given recognition test, unlearned items were coded as low familiarity; recently learned items were coded as medium familiarity, and early learned items were coded as high familiarity. Following every test, we compared the dropout rate for the three categories of items.

The notion of item selectivity has broader implications for the familiarity-based, source-independent, strategy selection model adopted here. The skill items affected by recognition testing should be correlated with the composition of the test list. Training items were divided into two sublists: one that was paired with similar foils in the recognition tests, and another that was paired with distinct foils. We compared retrieval levels for the two sublists, *similar items* versus *distinct items*, so as to test for item-specific transfer from recognition trials to skill trials. Given the flexibility of the source activation confusion model, it seemed to us that the direction that such transfer might take could be either positive or negative: Retrievals for similar items would be elevated if their familiarity scores benefited from exposure to related foils; retrievals for similar items would be depressed if their familiarity scores were flagged as untrustworthy via associations to confusable foils (i.e., the participants maintained two thresholds, high for confusable items and low for distinct items). In either event, item-specific transfer would reflect the operation of low-level frequency tracking, over and above any item-general motivational, instructional, or disruptive influence occasioned by the recognition tests.

In regard to disruptive effects, an entirely different explanation needs to be considered. The interpolated tests may have the force of an unrelated distractor task, disrupting memory traces through interference or delay. In that case, the amount of disruption engendered by a given test ought to be proportional to its (subject-determined) duration. Again, this explanation seems especially pertinent to the age-related disruption reported by Hoyer et al. (2003), given that older adults often show exaggerated distractor effects (e.g., Hedden & Park, 2003).

The distractor explanation shares a prediction with the frequency-tracking explanation. In both cases, new acquisitions would be more prone to drop out following a disruptive recognition test. The reasoning differs in the two cases: Selective drop out could be due to either the fragility of new stimulus–response associations or the near-threshold familiarity of newly identified stimuli. The composite results may settle this ambiguity: A correlation between test duration and the magnitude of disruption would point to the distractor account, whereas a separation between similar and distinct skill items would point to the familiarity account.

Method

Participants. Forty-two younger adults (18–24 years) and 43 older adults (60–80 years) were tested. Younger adults were recruited from the human participants pool of the Department of Psychology at Syracuse University. Older adults were community-residing volunteers recruited from the registry of the Adult Cognition Laboratory at Syracuse University. Prior to testing, the participants reported their education level and their overall physical health, using a 5-point scale. Individuals who reported that they were not taking any medications known to affect memory or learning, who rated their health as average, good, or excellent (ratings of 3, 2, or 1, respectively), and who had corrected or noncorrected near visual acuity of 20/30 or better were eligible. The Digit Span and the Digit–Symbol Substitution subtests of the Wechsler Adult Intelligence Scale–Revised (WAIS–R; Wechsler, 1981), the Number subtest from the Primary Mental Abilities Test (Thurstone & Thurstone, 1949), and a test of working memory that assessed spans for digit reordering were administered for the purpose of describing the age samples. Means and standard deviations for these measures are reported for the younger and older adults in Table 1. The measures reported in this table do not include the participants whose error rates exceeded a specified cutoff described in the Results section.

Stimuli and Procedure. The participants were trained on a four-digit summation task. Problems had the form

$$ab \wedge cd = ?,$$

where a, b, c, and d were single digits and the answer was also a single digit, given by the formula MOD 10 ($a + b + c + d$). To illustrate, to solve the problem $24 \wedge 57 = ?$, the four digits are first added together ($2 + 4 + 5 + 7 = 18$), and then the units position of the sum is reported [MOD 10 (18) = 8]. The grouping of the four digits into two pairs is irrelevant.

Five different sets of 12 problems were generated, with several constraints: At least three of the four digits had to be unique; their sum had to fall in the range 11–29, with 20 excluded; and the units position of the sum had to span the range 1–9 across the 12 problems. For skill training, one problem set was assigned to each participant,

Table 1
Means (With Standard Deviations) for Measures of the Characteristics of the Research Participants

	Experiment 1A		Experiment 1B
	Younger Adults	Older Adults	(Younger Adults)
N	35	42	36
Age	18.8 (0.8)	71.0 (5.4)**	19.1 (1.0)
Male	40%	41%	31%
Education	12.8 (0.9)	15.5 (3.1)**	12.9 (1.1)
Health	1.9 (0.6)	2.2 (0.7)*	1.9 (0.7)
Arithmetic speed	20.8 (7.7)	31.0 (12.5)**	21.0 (8.5)
Digit span	16.7 (3.1)	17.3 (3.9)	14.7 (4.0)
Digit symbol	72.6 (10.5)	53.0 (10.3)**	72.3 (10.6)
Number ordering	16.7 (3.8)	17.4 (4.1)	16.5 (3.9)

Note—For *N*, the participants dropped for low accuracy are not included (see the text). Education, self-reported number of years of formal education; Health, self-reported using a scale from 1 (*excellent*) to 5 (*poor*); Arithmetic speed, Number subtest score from the Primary Mental Abilities test (Thurstone & Thurstone, 1949); Digit span, measure combines the forward span and backward span scores for the WAIS–R Digit Span test (Wechsler, 1981); Digit symbol, WAIS–R Digit–Symbol Substitution test score; Number ordering, measure of working memory span in which strings of digits are reordered numerically. $^{*}p < .05$. $^{**}p < .001$ (t tests for age differences).

selected at random from the five sets. (The selected items also appeared as targets in the recognition tests; see below.) Instructions emphasized both speed and accuracy, and the participants were given practice and demonstrated proficiency in keyboard use and in solving practice problems prior to testing. The problem set was presented repeatedly, in 12-trial blocks, for a total of 18 blocks. During testing, the participants were allowed rest breaks after every few blocks.

Each trial consisted of the presentation of a fixation cross at eye level in the center of a computer monitor for 500 msec, followed by the presentation of one of the problems, which subtended about 6º of visual angle at a typical viewing distance. The stimulus remained on the screen until a keypress was made. An error message (with a blip sound) followed an incorrect response. After 500 msec, a correct response was followed by a strategy probe. The probe requested an introspective report from the participant as to whether the response just made was determined by a computation, by memory retrieval of the solution, or otherwise. Three keys on the computer keyboard were labeled C, M, and O for the participant to respond "compute," "memory," or "other" to the probe. A 1,000-msec blank screen followed each problem.

An online recognition test was administered after the 3rd, 6th, 9th, 12th, and 15th blocks of skill training. Each test consisted of 24 items: the 12 problems from the original list (target problems) and 12 newly generated problems (foils). For the purposes of recognition testing, the original list was divided in two sublists of 6 items each. Recognition foils paired with the target items of one sublist were generated with prefixes that matched the target items (e.g., the foil might be 34 ^ 69 = ?, where the corresponding target was 34 ^ 27 = ?). The other six targets were paired with foils generated with unique prefixes (as well as unique suffixes). Thus, in one sublist, targets were associated with similar foils, and in the other, with dissimilar foils. A different set of foils was created for each recognition test.

The 24 items of a recognition test were presented simultaneously on a single computer screen. The participants were instructed to use the mouse to mark a checkbox next to the items they had seen and were required to check exactly 12 items. The participants were allowed to revise their checks and clicked on a "Finished!" button when they were satisfied with their selections. The session was terminated 3 blocks after the last recognition test—that is, after the 18th block of skill training.

Results and Discussion

Eight participants (7 young, 1 old) failed to achieve 90% accuracy averaged over the 18 blocks of training and were removed from the data set. The error rates of the remaining sample did not differ by age (mean for 35 young adults, 5.8%; mean for 42 old adults, 4.1%; $p > .05$). The data from these 77 participants were analyzed in order to assess the effect of interpolated recognition tests on "*item learning*" during skill training (i.e., on the likelihood of retrieval solutions to a training item) and the effect of similarity between the recognition foils and the training items. The primary dependent measure was the number of retrieval solutions reported by a participant on a given repetition of the training set, a value with the range 0–12 (or a range of 0–6 when retrieves are separated by sublist). The *other* strategy option was rarely reported (comprising 1.10% of all strategy reports for the young and 0.03% for the old). Therefore, the number of computational solutions on a given repetition of the training set is given almost exactly by the 12s (or 6s) complement of the number of retrieval solutions.

Item learning. The number of items retrieved is presented in Figure 1 as a function of blocks and age group (broken down further by target–foil similarity, discussed below). It can be seen that retrieves increased with training and that the retrieval counts of older adults fell below those of the younger adults. These trends replicate the well-documented shift from computational solutions to retrieval solutions as a function of item repetition, as well as the existence of a substantial age deficit in the level of item learning. The trends were confirmed by a repeated measures ANOVA, conducted on Blocks 2–18 and collapsing over target–foil similarity. There were significant main effects due to age [$F(1,75) = 10.04$, $MS_e = 113.80$, $p < .01$] and to block [$F(16,1200) = 21.90$, $MS_e = 2.74$, $p < .01$]; the interaction between age and block was not significant ($p = .14$). Thus, the age effect was expressed as a uniform reduction in retrieval rate throughout training, rather than as a difference in the rate of acquisition, as was found by Touron, Hoyer, and Cerella (2004). The purely additive age effect in these data is probably due to truncation of the acquisition curve at Block 18, before asymptotic performance had been reached.

Conspicuous in Figure 1 are dips in the item retrieval curves for both age groups that occur immediately after the recognition tests and that are more prominent early in the session. This pattern was evaluated in a repeated measures ANOVA to determine the effects of the recognition test number (1–5), age (young vs. old), and training phase (the block immediately preceding a recognition test vs. the block immediately following the test) on the number of reported retrieves. The number of items retrieved increased with each subsequent test [$F(4,300) = 23.30$, $MS_e = 4.256$, $p < .001$], from about 1.7 items (Test 1) to 3.9 items (Test 5), at a rate comparable in the young and the old ($p = .25$ for the age $\times$ test interaction). Training phase also had an effect [$F(1,75) = 35.2$, $MS_e = 1.387$, $p < .001$], indicating that the number of items retrieved in the block before a recognition test was significantly higher (3.2 items) than the number of items retrieved in the block after the test (2.7 items), regardless of age ($p = .82$ for the age $\times$ phase interaction). The effect of training phase diminished with each subsequent test [$F(4,300) = 2.35$, $MS_e = 1.353$, $p = .05$]: More items were lost after Recognition Test 1 (0.95 items lost) than after subsequent tests (0.27 items lost following Recognition Test 5), and this effect did not differ with age (the three-way interaction was not significant, $p = .12$). The effects of test and phase were superimposed on a main effect of age [$F(1,75) = 9.95$, $MS_e = 65.147$, $p = .002$]: Older adults retrieved far fewer items overall (2.0 items) than did younger adults (3.9 items). These effects are illustrated in Figures 2A and 2B. (Note that Figure 2 sums over the sublists; hence, retrieval frequencies are about double those in Figure 1.)

The interpolated tests disrupted retrievals in both age groups by an amount that diminished with each successive test. What caused the disruption? We explored two hypotheses. The duration of the tests was controlled by the participants, who clicked an exit button when they were

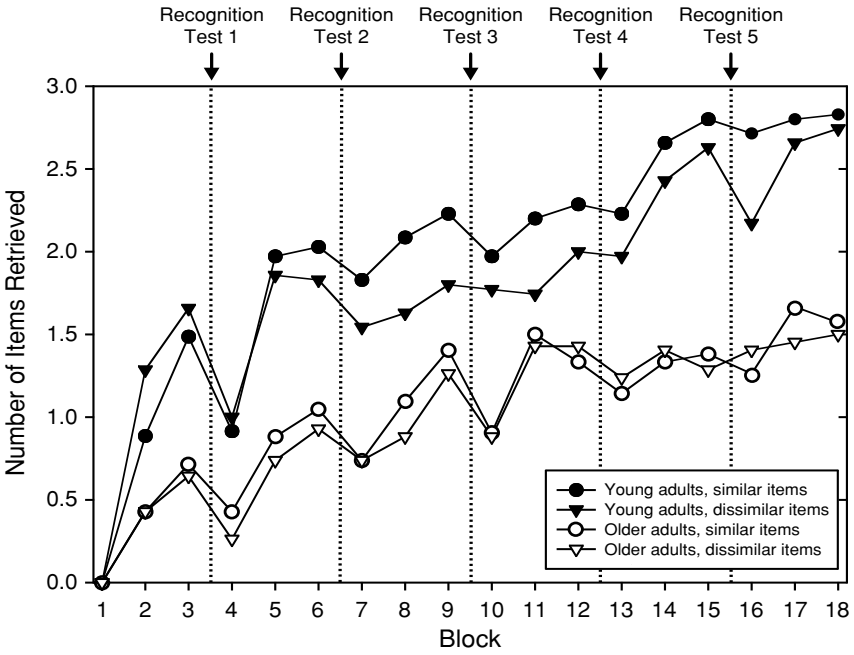

Figure 1. Frequency of item retrieval (from two sublists of six items each) as a function of repetition (block). Retrieval frequencies increased with training and were higher for younger adults than for older adults. Interpolated recognition tests are indicated by dotted vertical lines. Retrievals show a transient disruption following a recognition test, an effect that diminished as training proceeded.

finished. Each test took up to several minutes to complete, as is shown in Figure 2C. The figure also shows that test durations decreased with practice. Given that both the item loss scores (retrieved before test minus retrieved after test, shown in Figure 2B) and the test durations diminished with practice, we asked whether the second variable explained the first. This is the idea that an interpolated test may have functioned as an unrelated distractor task; in such a case, the amount of memory loss would be proportional to the amount of interpolated activity. Accordingly, we calculated the correlation between a participant's loss score for a given test and the duration of the test (excluding the participants who reported no retrieves before or after). Note that if the correlation were computed from scores pooled from the five tests combined, it would be spuriously high, due to the common trend in the two measures. Instead, we computed separate correlations for each test (and each age group). The results are given in Table 2. The correlations were uniformly low (range, $-.09$ to $+.22$), and none approached significance. Evidently, the item loss occasioned by a test was not due to its distracting effect.

The second hypothesis was suggested by the negatively accelerated shape of the item acquisition curves, as seen in Figure 1 (see also Figure 5). The number of items gained per block, or per three-block epoch, diminishes with training, as does the number of items lost on successive recognition tests. We tested the idea that newly acquired items were especially vulnerable to disruption. If that were the case, losses would decrease with training, because the number of new acquisitions decreased.

To evaluate this hypothesis, we counted the number of newly acquired items that were retrieved in the block before a recognition test and compared this with the number of these same items that were retrieved in the block after the test. In other words, how many of the items lost were newly acquired? An item was scored as newly acquired if it was retrieved in the block preceding a recognition test, but not in the block immediately following the previous test. The data were summed across Recognition Tests 2–5 and across participants. The counts are given in Table 3 and show that the percentage of retrieved items that were newly acquired fell from 45% (52% for the old adults) in the block before a test to 34% (35% for the old adults) in the block after the test. That is, a disproportionate number of newly acquired items were lost. The before and after difference in these frequencies was confirmed by chi-square tests (see Table 3): Item status (newly acquired or previously acquired) interacted with phase (before or after) for both the young and the old. Thus, newly acquired items were more prone to disruption, perhaps because the memory traces for those items were weaker or because they had lower familiarity scores.

Item recognition. For each recognition test, hits and false alarm rates were transformed to d' scores separately for the similar target–foil subset of items and the dissimilar target–foil subset. The resulting scores were subjected to a repeated measures ANOVA with age, recognition test number, and similarity as factors. The analysis demonstrated significant main effects due to test [$F(4,300) = 18.1$, $MS_e = 0.480$, $p < .001$] and similarity [$F(1,75) =$

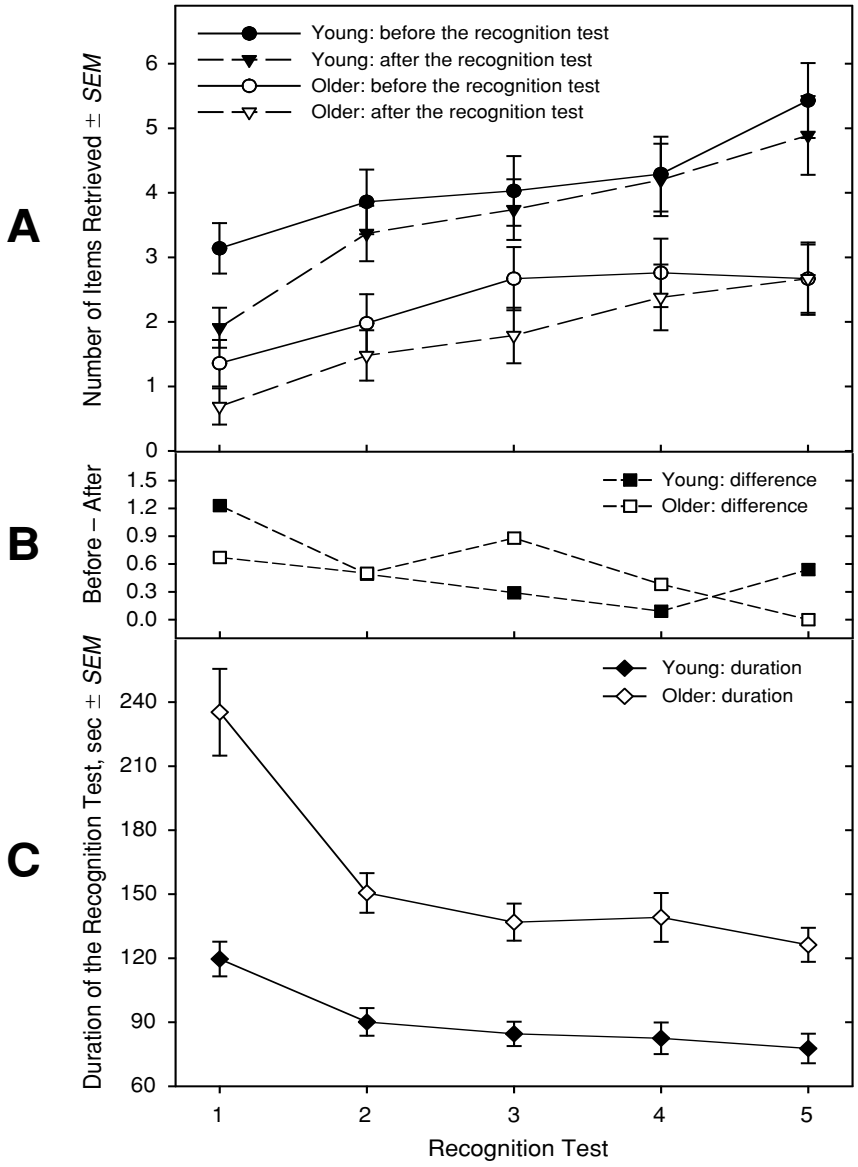

Figure 2. (A) Frequency of item retrieval as a function of recognition test number, shown separately for the block preceding a test and the block following the test. The two sublists in Figure 1 are combined. Plot shows that overall frequencies increased and that the retrieval penalty diminished with successive tests. (B) The retrieval penalty from panel A as a function of recognition test number for younger and older adults. (C) Average duration of self-timed recognition tests for younger and older adults for Recognition Tests 1–5.

48.2, $MS_e = 0.689$, $p < .001$] and a significant test × similarity interaction [$F(4,300) = 4.6$, $MS_e = 0.478$, $p < .01$]. Overall, d' values improved with each successive recognition test, and dissimilar targets were recognized better than similar targets. The similarity effect increased over tests. The age effect was marginally significant [$F(1,75) = 3.5$, $MS_e = 1.911$, $p = .06$], suggesting that the younger adults had somewhat higher d' values (.76) than did the older adults (.57). These effects are illustrated in Figure 3.

Thus, the foil manipulation was successful in modulating new/old judgments, increasingly so as training proceeded. The interaction can be understood simply in terms of the growing familiarity of targets with training. Paired with dissimilar foils, higher target familiarity led to enhanced d' scores; paired with similar foils, it led to proportionately reduced d' scores. The demonstrable confusability of targets and similar foils sets the stage for the assessment of similarity effects on skill trials performed below.

Table 2
Correlations Between the Duration of the Recognition Test and the Number of Items Lost Following the Recognition Test, for Tests 1–5

Recognition Test	Young Adults ($N = 33$)	Older Adults ($N = 30$)
1	.15	.22
2	.03	−.07
3	.12	−.21
4	−.09	.002
5	−.03	.21

Response times. Figure 4 gives response times as a function of block, averaged across each age group. Several effects are conspicuous in the figure. Old participants were slower than young; increased training was associated with a downward trend, more pronounced in the young than in the old, and the interpolated recognition tests generated a cyclic perturbation superimposed on the downward trend. A mixed model ANOVA confirmed the effects of age [$F(1,75) = 7.47$, $MS_e = 21.6 * 10^6$, $p = .008$] and of block [$F(17,1275) = 34.65$, $MS_e = 2.9 * 10^5$, $p < .001$; linear trend, $F(1,75) = 136.35$, $MS_e = 1.05 * 10^6$, $p < .001$] and of the age × block interaction [$F(17,1275) = 5.72$, $MS_e = 2.9 * 10^5$, $p < .001$; linear trend, $F(1,75) = 11.29$, $MS_e = 1.05 * 10^6$, $p < .001$]. Thus to an, admittedly crude, first approximation, response times fell on a linear trend at a rate greater in the young than in the old. These latency effects are well established in the literature on skill learning and age (e.g., Touron et al., 2004).

Rickard (1997) advanced an important interpretation of response times measured during skill training (see also Delaney, Reder, Staszewski, & Ritter, 1998). Given that a typical block of trials comprises a mixture of computational solutions and retrieval solutions, it follows that the mean response time for the block will be determined by the formula for a statistical mixture of the two component times. Because computation times are typically far longer than retrieval times, it follows that the primary predictor of change in the mean is change in the mixture proportion. The upshot is that the response time curve and the retrieval proportion curve will be mirror images, reflected through their common abscissa (blocks). Qualitatively, this interpretation fits our data comfortably. The mirror image correspondence can be seen quite clearly in our Figures 1 and 4: Upward trends in the former are transformed into downward trends in the latter.

Table 3
Total Number of Items Retrieved Before and After Each Recognition Test, Summed Across Tests and Participants

Phase	Item Status: Newly Acquired	Item Status: Previously Acquired
	Young Adults*	
Before	277	339
After	138	272
	Old Adults†	
Before	220	203
After	91	166

*$\chi^2(1) = 31.187$, $p < .001$. †$\chi^2(1) = 17.939$, $p < .001$.

The correspondence implies that computations had longer latencies than did retrievals. We verified this implication by examining the component times, separating computes and retrieves by means of participants' trial-by-trial strategy reports, averaged over block but broken down by similarity (similar or dissimilar). A mixed model ANOVA established that computes were indeed longer than retrieves [$F(1,58) = 116.5$, $MS_e = 7.05 * 10^5$, $p < .001$; computes, 4,041 msec; retrieves, 2,865 msec] and that the old were slower than the young [$F(1,58) = 10.9$, $MS_e = 4.29 * 10^6$, $p = .002$; young, 3,009 msec; old, 3,897 msec]. Age and component did not interact [$F(1,58) = 0.965$, $p = .330$]. The compute–retrieve latency differences replicate the findings of Rickard (1997) and Delaney et al. (1998), as well as the age-related slowing observed by Touron et al. (2004). Considered jointly, these *component* frequencies and latencies support the *mixture* interpretation of skill measures elaborated by Rickard (1997).

Returning to the ANOVA on latencies, the main effect of similarity was not significant [$F(1,58) = 0.450$, $p = .505$], but there was a significant similarity × age interaction [$F(1,58) = 6.25$, $MS_e = 2.63 * 10^5$, $p = .015$]. Follow-up ANOVAs established the basis of this interaction. For the young, responding to similar items (2,904 msec) was significantly faster than responding to distinct items (3,115 msec) [$F(1,32) = 4.85$, $MS_e = 3.03 * 10^5$, $p = .035$]; for the old, there was no difference [$F(1,26) = 1.87$, $p = .183$]. The contrast between the young and the old is apparent in the figure.

The recognition test results showed that targets were confused with similar foils. This had one repercussion on skill trials: Response times for similar targets were somewhat speeded. The effect was seen equally in compute trials and retrieve trials, but only for the young. One interpretation of the facilitation is that the strategy decision, to compute or to retrieve, was streamlined for similar targets—exactly the result observed by Palmeri (1997) and attributed to spillover to the target, of the strengthening effects of exposure to similar nontargets.

Similarity effects. Here, we assess the effects of target similarity on the frequency of computes and retrieves, as opposed to their latency. To this end, we returned to the item retrieval counts analyzed earlier and added the similarity factor to the ANOVA. Again, younger and older samples were analyzed separately. The blocks factor ranged from 4 through 18. (Similar and dissimilar training items were not defined prior to the first recognition test. A separate analysis on Blocks 2 and 3 alone showed that there were no initial differences between items that were to become similar and dissimilar, for either the younger or the older adults.)

For the young, the ANOVA revealed a significant main effect of similarity [$F(1,34) = 4.81$, $MS_e = 3.45$, $p < .05$] and of block [$F(14,476) = 9.39$, $MS_e = 1.72$, $p < .01$]. Similar training items were retrieved with higher frequency (2.23 per block) than were dissimilar items (1.98 per block). The interaction between similarity and block

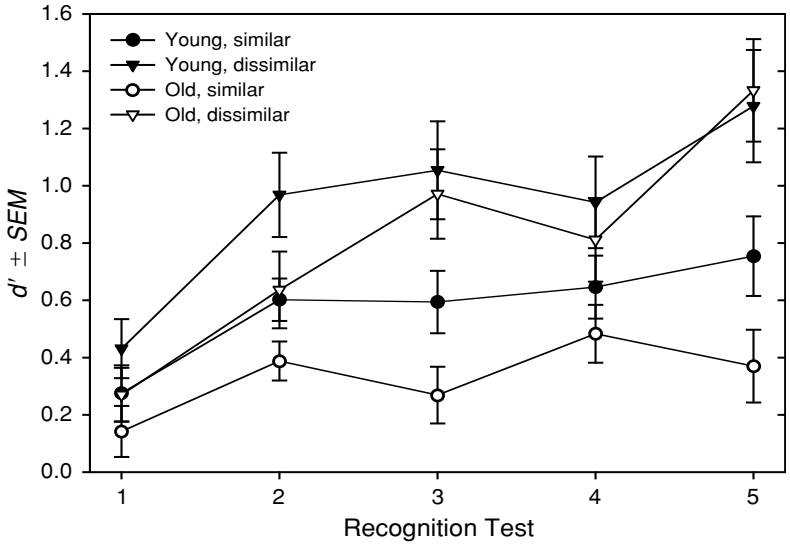

Figure 3. Item recognition in units of *d′* as a function of recognition test number. *d′* increased over successive tests and was higher in young adults than in old adults. Target–foil similarity depressed *d′* by an amount that grew with test number.

was not significant, indicating that the similarity benefit was uniform across training. The corresponding analysis for the old demonstrated a significant effect of block [$F(14,574) = 11.10, MS_e = 0.90, p < .01$] but no effect of item similarity [$F(1,41) < 1$] and no interaction. The existence of a similarity effect for the young, and its absence in the old, can be clearly seen in Figure 1.

In the young, the effect of item similarity on retrieval frequency complements its effect on solution times; in both cases, there was carryover from the recognition tests to the skill trials. For retrieval frequency the nature of this exchange was of particular interest, because both positive and negative transfer seemed to us to be possibilities beforehand. It seemed possible that retrieval of similar items would be depressed because of uncertainty over their target/foil status; it also seemed possible that similarity-induced familiarity would lead to more retrievals. The data point to the latter outcome. Palmeri's (1997)

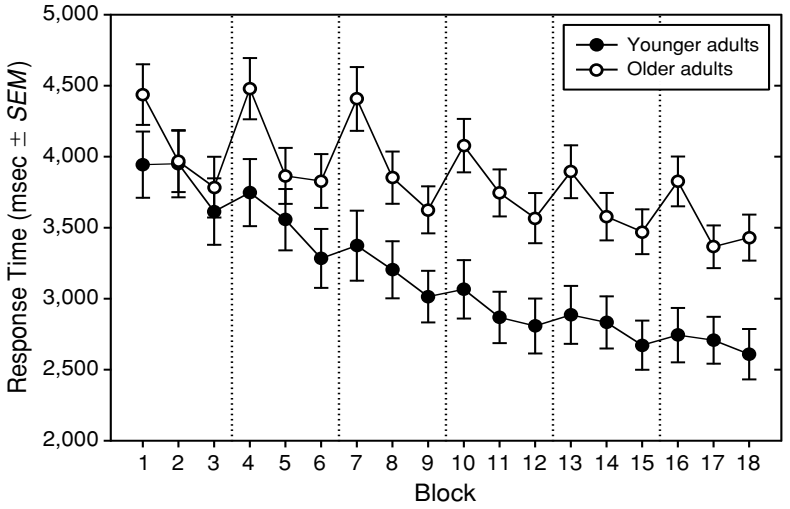

Figure 4. Response time as a function of repetition (block). Overall, response time dropped with training and was higher for older adults than for younger. Interpolated recognition tests are indicated by vertical lines. Latencies show a transient disruption following recognition tests, a mirror image of the effect seen in Figure 1 (see the *mixture* interpretation in the text).

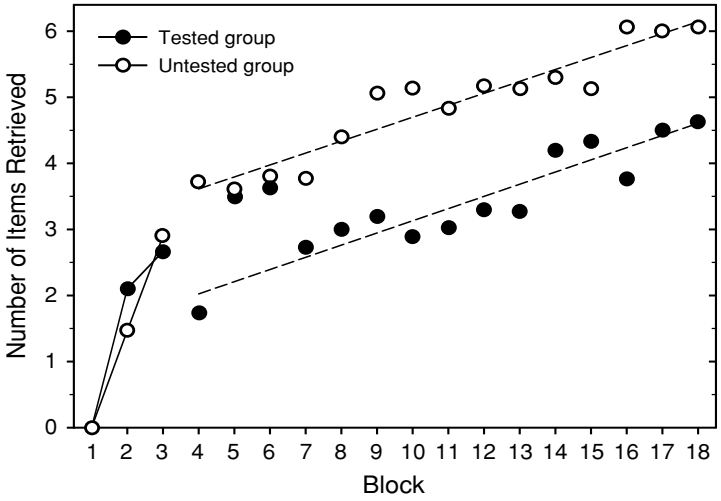

Figure 5. Number of items retrieved by younger adults in the condition with interpolated tests (filled circles) and without interpolated tests (open circles). Retrieval frequencies were consistently higher in the test-free condition, as evidenced by the dotted regression lines.

distinction between *friends* and *enemies* in a training list seems pertinent here. Friends and enemies are subsets of items, both of which are confusable with a target item. Friends have the same response as the target; enemies have a different response. To the extent that items are, in fact, confused, friends enhance item retrieval (a case of getting the right answer for the wrong reason), and enemies retard item retrieval—due to increased cautiousness or, in the worse case, "strong-but-wrong" errors (Woltz et al., 2000). In our experiment, confusable foils evidently played the role of friends. Because they did not activate misleading responses (recognition items were never associated with answers), they merely contributed to the overall familiarity of related targets and, hence, made retrieval attempts more likely for those items—attempts that were not accompanied by any increase in the error rate [$t(34) = 1.466, p = .152$ (young), and $t(41) = 0.305, p = .75$ (old), for the similar-item/distinct-item comparison in error frequency].

We argued in the introduction that any transfer between recognition trials and skill trials, regardless of its sign, would be evidence for frequency-based compute/retrieve strategy selection—in particular, for a frequency-tracking mechanism subject to source confusion. At least in the young participants, all exposures to an item (or to its parts) apparently incremented the same familiarity counter, regardless of context.

The results from the older participants stand in striking contrast. Although recognition tests disrupted retrievals in young and old alike, no sublist effects were seen in the old. In their case, neither response speed nor retrieval frequency was influenced by the target–foil similarity manipulation. Perhaps it is not so surprising that exposure to one additional half-matching foil had no detectable impact on these participants, given the much lower level of item learning of any sort seen in their data.

EXPERIMENT 1B

In Experiment 1A, intermittent recognition tests disrupted skill learning, in apparent contradiction to the facilitative effect found by Ackerman and Woltz (1994). However, we too found a facilitative effect in young participants, one confined to training items that were paired with similar foils in the recognition tests. We conjecture that those items benefited from increased familiarity induced by the recognition foils. If this were so, an analogous benefit should attach to every training item, induced by the recognition targets, as opposed to the foils. That is to say, after Recognition Test 1, every training item had been seen four times, not three (and similar training items had been seen somewhere between four and five times, thanks to the confusable foils). This could be the source of a *universal* facilitation that would have gone undetected in Experiment 1A, because there was no untested control group. Indeed, increased familiarity due to exposure to recognition test targets may have been responsible for Ackerman and Woltz's result, rather than the motivational influence to which they attributed it.

In Experiment 1B, a sample of young adults was recruited and given 18 blocks of skill training, exactly as in Experiment 1A, without the interim recognition tests. It was an open question whether the item-learning curve for young adults in this condition would rise above or fall below that of the young adults in Experiment 1A.

Method

Forty-two younger adults (18–24 years) were trained with the same procedure and stimuli as those in the previous experiment,

without the recognition tests. The same accuracy criterion was applied (at least 90% accuracy averaged over 18 blocks), resulting in the exclusion of 6 of the participants from the data pool.

Results and Discussion

The experimental design requires that the *tested* and the *untested* groups do not differ in their retrieval levels over the first three training blocks. As it turned out, the new sample reported considerably fewer retrievals over this period. Almost surely, this difference resulted from the fact that the second sample had been recruited during the summer term at our university, whereas the first sample was recruited from the regular school year term. We adjusted for the sample differences by matching pairs of treated and untreated participants on the basis of their retrieval counts for Blocks 2 and 3 (Block 1 had no retrievals). Matched samples of 30 participants each were formed in this way; the retrieval frequencies from these subsamples are shown in Figure 5. A repeated measures ANOVA with condition (tested or untested) and block (2 or 3) as factors showed that the two groups did not differ in early session retrievals [$F(1,58) = 0.15$, $MS_e = 2.14$, $p = .70$].

This paved the way for the principal analysis, a comparison of item learning over Blocks 4–18. An ANOVA on these data demonstrated that the control (untested) group reported more retrieves per block (4.9 items) than did the experimental (tested) group (3.4 items) [$F(1,58) = 5.74$, $MS_e = 80.012$, $p < .05$]. There was also a significant effect of block [$F(14,812) = 11.6$, $MS_e = 2.814$, $p < .001$], due to a linear increase in retrievals from Block 4 to Block 18 (the linear trend accounted for 91% of the total variance associated with the blocks effect). These main effects were qualified by a significant condition × block interaction [$F(14,812) = 2.5$, $MS_e = 2.814$, $p < .01$], which lay in the higher order components of the blocks trend, due to the cyclic perturbations specific to the experimental condition. For our purposes, item acquisition for the two groups over Blocks 4–18 can be represented by two parallel lines, one above the other, as depicted in the figure.

Interim testing had a twofold effect on item learning. There was the pattern demonstrated in Experiment 1A of transient disruption followed by rapid recovery, which diminished in intensity as training progressed. Superimposed on this transient effect was the sustained effect demonstrated in Experiment 1B, a consistent depression in retrieval, seen over the entire course of training. It is, of course, quite possible that the latter effect is entirely due to the former—a matter of incomplete recovery between tests.

GENERAL DISCUSSION

The formation of a cognitive skill is a complex process in which direct retrieval comes to replace computation as the means by which solutions to previously encountered problems are obtained (Delaney et al., 1998; Logan, 1988; Rickard, 1997). Critical to the shift in solutions is the accumulation of item information in memory. One of the unsettled issues is the extent to which item encoding (or *item learning*) occurs automatically as a concomitant of computation (assumed by the aforementioned researchers) or is the consequence of a deliberate strategy adopted in the training situation (Haider & Frensch, 2002). In a noun–noun table lookup task, for example, Ackerman and Woltz (1994) found that a third of their participants never switched to consistent retrieval, suggesting that the memory strategy may have been optional, not obligatory. In a second condition, it was found that recognition tests interspersed between skill-training blocks reduced the number of nonretrievers to one fifth. Ackerman and Woltz conjectured that the recognition tests heightened awareness of the memory strategy, leading more participants to adopt it.

In the introduction, we argued that this finding could also be understood as the outcome of a low-level frequency-tracking mechanism, which summed across skill trials and recognition trials, without recourse to executive-level strategy changes. The idea is that of Reder and Ritter (1992), that a training item triggers a retrieval attempt if its *familiarity* exceeds a threshold value. Because items would benefit from their appearance in recognition trials, as well as from their appearance in skill trials, retrieval attempts would be elevated as a result of concurrent recognition testing.

What is interesting is that this same mechanism allows for the opposite outcome as well. Participants who realize the potential for confusion may raise their familiarity threshold, thereby depressing the frequency of retrievals. This was just the outcome obtained by Hoyer, Cerella, and Onyper (2003): Interpolated recognition tests depressed, rather than enhanced, item learning. Their outcome appears to contradict that of Ackerman and Woltz (1994), but they measured only the transient effects of recognition testing; there was no untested control group against which to assess sustained effects.

The present experiments were designed to reassess both transient and sustained effects of recognition testing, using an artificial arithmetic task. In Experiment 1A, training items were divided into one sublist that was paired with similar foils in the recognition tests and another sublist that was paired with distinct foils. The intention here was to test for the transfer of specific information from the recognition tests to the skill trials, beyond any general motivational, instructional, or disruptive influence.

Experiment 1A replicated the findings of Hoyer et al. (2003) in demonstrating a transient depression of retrieval solutions to training problems following recognition testing (see Figures 1 and 2). An item-level analysis showed that the depression was due primarily to the loss of recently acquired items. New memory traces were more *fragile*. Further observations gave some insight into the cause of the fragility. We found that the amount of deflation was unrelated to the (subject-controlled) duration of the interpolated tests. Thus, the memory loss was not due to the interruption or the distraction occasioned by the recognition tests.

Schunn et al.'s (1997) source confusion strategy selection model presents an alternate interpretive framework.

Retrieval deflation is viewed as a consequence of an elevated decision criterion—elevated to protect participants against false alarms, consequent to their exposure to foils. An elevated criterion would affect newly acquired items (low familiarity) more than firmly established items (high familiarity), in keeping with our observation.

The source confusion view is supported by other aspects of the data from Experiment 1A. We found that retrievals were elevated for skill items paired with similar foils (an effect superimposed on the depression effect). This proves that exposure to recognition foils fed back to skill trials (and further weakens the idea that the disruption occasioned by a recognition test may have been due to its impact as an unrelated distractor). The positive sign of the feedback loop accords with the theory: Similar targets enjoyed higher similarity scores and, hence, were selected for retrieval more often. We also found that similar targets were processed more quickly. Again the sign of the feedback accords with theory: Strategy decisions were faster for more familiar items, a latency effect demonstrated by Palmeri (1997).

Evidently, recognition testing in Experiment 1A had a two-sided effect on skill retrievals. The similarity of target items was elevated, leading to higher retrieval levels; and the retrieval threshold was increased, leading to lower retrieval levels. Experiment 1B allowed the combined effect of the two tendencies to be assessed by providing an untested baseline. The comparison of untested and tested participants was clear-cut: In our experiment, the net effect of recognition testing was negative; the retrieval levels of tested participants were depressed by about 30% throughout training (Figure 5).

Changes in familiarity can be combined with changes in threshold to explain any possible finding, depending on the relative strength of the two opposing influences. Is this two-factor account of interpolated memory testing falsifiable? It makes one prediction that seems especially telling. The theory allows (but does not demand) a spectrum of outcomes, from positive to negative. If diverse outcomes were, in fact, observed (across experiments), this would point strongly to a dual-factor account and would salvage an otherwise confusing group of studies.

In fact, diverse outcomes appear to be the rule, rather than the exception, in this area. In our own studies in which two age groups and two tasks were used, three outcomes were negative (i.e., skill learning was disrupted by recognition testing), and one outcome was neutral. Ackerman and Woltz's (1994) outcome was positive: At the end of training, 80% of the tested group retrieved consistently, in comparison with 67% of the untested group. In two other studies, the noun–noun table lookup task was used (Rogers & Gilbert, 1997; Touron & Hertzog, 2004). In both of these studies, the outcome for younger adults was negative, and the outcome for older adults was positive. The two negative outcomes have been slighted and deserve to be highlighted here. In Rogers and Gilbert, by the end of training, there was no difference between the tested and the untested young participants; 75% of the former and 70% of the latter consistently retrieved. But at the midpoint of training, only 50% of the tested group retrieved, in comparison with 70% of the untested group. This suggests that recognition testing retarded the rate of item learning, although both groups eventually reached the same asymptote. In Touron and Hertzog, compute/retrieve strategies were probed over the course of skill training, rather than inferred post facto from response times. Touron and Hertzog plotted retrieval frequencies as a function of blocks, and their outcome for young participants matches our findings (in Figure 5): The curve from the tested group falls consistently below that from the untested group.

The inconsistencies are striking; they cannot be separated by task or by age group. To us, the situation points to multiple, opposing factors, the balance of which differs from study to study, leading to one outcome or another in each particular case. Accepting this interpretation, what has been learned about the principles governing skill acquisition? In our view, these studies heighten the importance of Schunn et al.'s (1997) notion of strategy selection. Strategy choice appears to be determined in part by an item familiarity variable, which derives its strength not only from problem–answer sequences encountered during training, but also from problem-only exposures outside of training.

This, of course, is nothing other than frequency-of-occurrence tracking, a process thought to operate autonomously, apart from strategy. Its very autonomy, however, opens a trainee to false alarms, decisions to retrieve unsupported by recorded solutions, owed to extratraining exposures or interitem confusions. This hazard is countered by the imposition of a subject-controlled retrieval criterion, which can be set high or low as the opportunities for misinformation wax or wane.

We began by contrasting *bottom-up* with *top-down* views of skill learning. With respect to the studies at hand, bottom-up processes appear to tell most, but not all, of the story. To bottom-up processes are owed a participant's tracking of item frequency, as well as the accumulation of solution information. But these assets do not determine a solution method on their own; the final decision depends on a strength threshold under control of the participant. Here, then, is scope for top-down influences—"strategy" differences due, in the present case, to training concomitants and, in other cases, to instruction (Ackerman & Woltz, 1994, Experiments 4 and 5), to pretraining (Touron & Hertzog, 2004), or to individual preferences (Rogers et al., 2000).

It seems to us that the bottom-up models are well positioned to accommodate such influences. Rickard's (1997; see also Rickard, 2004) neural network model includes compute and retrieve subgoals, whose base-level activation determines the decision to compute or retrieve, together with the strength of the current memory trace. Palmeri's (1997; see also Nosofsky & Palmeri, 1997) adaptation of Logan's (1988) race model involves a series of micro-retrievals that accumulate evidence or else are supplanted by a completed computation. Both subgoal activation and the accumulator cap have the force of threshold variables.

We need only imagine that they are placed at the disposal of an *executive* to see how task-level, top-down influences may be wedded to item-level, bottom-up processes.

REFERENCES

Ackerman, P. L., & Woltz, D. J. (1994). Determinants of learning and performance in an associative memory/substitution task: Task constraints, individual differences, volition, and motivation. *Journal of Educational Psychology*, **86**, 487-515.

Barrouillet, P., & Fayol, M. (1998). From algorithmic computing to direct retrieval: Evidence from number and alphabetic arithmetic in children and adults. *Memory & Cognition*, **26**, 355-368.

Compton, B. J., & Logan, G. D. (1991). The transition from algorithm to retrieval in memory-based theories of automaticity. *Memory & Cognition*, **19**, 151-158.

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological Science*, **9**, 1-7.

Haider, H., & Frensch, P. A. (2002). Why aggregated learning follows the power law of practice when individual learning does not: Comment on Rickard (1997, 1999), Delaney et al. (1998), and Palmeri (1999). *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 392-406.

Hedden, T., & Park, D. C. (2003). Contributions of source and inhibitory mechanisms to age-related retroactive interference in verbal working memory. *Journal of Experimental Psychology: General*, **132**, 93-112.

Hoyer, W. J., Cerella, J., & Onyper, S. V. (2003). Item learning in cognitive skill training: Effects of item difficulty. *Memory & Cognition*, **31**, 1260-1270.

Loft, S., Humphreys, M., & Neal, A. (2004). The influence of memory for prior instances on performance in a conflict detection task. *Journal of Experimental Psychology: Applied*, **10**, 173-187.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, **95**, 492-527.

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. *Psychological Science*, **15**, 302-306.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, **104**, 266-300.

Ofen-Noy, N., Dudai, Y., & Karni, A. (2003). Skill learning in mirror reading: How repetition determines acquisition. *Cognitive Brain Research*, **17**, 507-521.

Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 324-354.

Ratcliff, R., Spieler, D., & McKoon, G. (2000). Explicitly modeling the effects of aging on response time. *Psychonomic Bulletin & Review*, **7**, 1-25.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 435-451.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General*, **126**, 288-311.

Rickard, T. C. (2004). Strategy execution in cognitive skill learning: An item-level test of candidate models. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 65-82.

Rogers, W. A., & Gilbert, D. K. (1997). Do performance strategies mediate age-related differences in associative learning? *Psychology & Aging*, **12**, 620-633.

Rogers, W. A., Hertzog, C., & Fisk, A. D. (2000). An individual differences analysis of ability and strategy influences: Age-related differences in associative learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 359-394.

Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stroffolino, P. J. (1997). To calculate or not to calculate: A source activation confusion model of problem familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 3-29.

Siegler, R. S., & Lemaire, P. (1997). Older and younger adults' strategy choices in multiplication: Testing predictions of ASCM using the choice/no-choice method. *Journal of Experimental Psychology: General*, **126**, 71-92.

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 31-76). Hillsdale, NJ: Erlbaum.

Strayer, D. L., & Kramer, A. F. (1994). Aging and skill acquisition: Learning-performance distinctions. *Psychology & Aging*, **9**, 589-605.

Thurstone, L. L., & Thurstone, T. G. (1949). *Examiner manual for the SRA Primary Mental Abilities Test* (Form 10-14). Chicago: Science Research Associates.

Touron, D. R., & Hertzog, C. (2004). Distinguishing age differences in knowledge, strategy use, and confidence during strategic skill acquisition. *Psychology & Aging*, **19**, 452-466.

Touron, D. R., Hoyer, W. J., & Cerella, J. (2004). Cognitive skill learning: Age-related strategy shifts and speed of component operations. *Psychology & Aging*, **19**, 565-580.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale–Revised*. New York: Psychological Corporation.

Woltz, D. J., Gardner, M. K., & Bell, B. G. (2000). Negative transfer errors in sequential cognitive skills: Strong but wrong sequence application. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 601-625.

Zacks, R. T., & Hasher, L. (1982). Automatic encoding of event frequency: Further findings. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **8**, 106-116.

(Manuscript received August 27, 2004;
revision accepted for publication April 28, 2005.)