Journal of Experimental Psychology:
Learning, Memory, and Cognition
2010, Vol. 36, No. 6, 1441–1451

0278-7393/10/$12.00 DOI: 10.1037/a0020636

Test-Enhanced Learning of Natural Concepts: Effects on Recognition Memory, Classification, and Metacognition

Larry L. Jacoby, Christopher N. Wahlheim, and Jennifer H. Coane
Washington University in St. Louis

Three experiments examined testing effects on learning of natural concepts and metacognitive assessments of such learning. Results revealed that testing enhanced recognition memory and classification accuracy for studied and novel exemplars of bird families on immediate and delayed tests. These effects depended on the balance of study and test trials during training. Metacognitive measures provided results suggesting that participants were aware of the beneficial effects of testing. A new measure of metacognition at the level of categories is introduced and shown to be potentially useful for theory and applied purposes. It is argued that focusing on optimizing the learning of natural concepts encourages the convergence of theorizing about memory, concept learning, and metacognition and holds promise for the development of applications to education.

Keywords: testing effects, concept learning, metacognition

Testing memory enhances learning and reduces forgetting more than does repeated study in a variety of contexts (for a review, see Roediger & Karpicke, 2006a). Testing has been shown to improve memory for paired associates (e.g., Carrier & Pashler, 1992), recall of information from prose passages (e.g., Chan, McDermott, & Roediger, 2006; Roediger & Karpicke, 2006b), and test performance in the classroom (e.g., McDaniel, Anderson, Derbish, & Morrisette, 2007). For example, Karpicke and Roediger (2008) examined effects of retrieval practice on learning of a foreign language using Swahili–English paired associates as materials (e.g., *mashua*–boat). Results revealed that delayed recall of the English response (boat) when given the Swahili cue (*mashua*–?) was dramatically enhanced for pairs that were repeatedly tested rather than repeatedly presented for study. However, students were unaware of the benefits of retrieval practice, as evidenced by their not predicting that repeated recall of vocabulary words would enhance their later recall performance more than did repeated study. University students seem to be generally unaware, despite its powerful effects, that testing enhances memory performance (for a review, see Karpicke, Butler, & Roediger, 2009).

This article was published Online First August 30, 2010.

Larry L. Jacoby, Christopher N. Wahlheim, and Jennifer H. Coane, Department of Psychology, Washington University in St. Louis.

Jennifer H. Coane is now at the Psychology Department, Colby College.

We are grateful to Andrea Hughes and Tammy Duguid for the selection and preparation of materials for this project and to Mitchell Waite and WhatBird.com for permission to use the materials. We thank Sarah Arnspiger, Rissa Ivens, Carole Jacoby, Kate Koch, and Rachel Teune for their assistance with data collection and Julie Bugg, Brigid Finn, and Ruthann Thomas for their helpful comments on an earlier version of this article. This research was supported by Grant 22020166, Applying Cognitive Psychology to Enhance Educational Practice, from the James S. McDonnell Foundation, awarded to Larry L. Jacoby.

Correspondence concerning this article should be addressed to Larry L. Jacoby, Department of Psychology, Washington University, St. Louis, MO 63130. E-mail: lljacoby@artsci.wustl.edu

The experiments reported in this article extended the investigation of testing effects to the learning of natural concepts by examining testing effects on learning to classify studied and novel exemplars of bird families. Doing so is of interest from a practical or pedagogical standpoint and promises to be useful for the development of theories of testing effects and theories of metacognition. For theorizing about both testing effects and metacognition, it is potentially important to examine effects using materials that have an underlying category or similarity structure. Prior experiments investigating testing effects have largely been limited to the verbal learning tradition, as illustrated by the example of vocabulary learning described above (Karpicke & Roediger, 2008), or have employed multiple-choice or recall tests of information presented in prose passages. Cues presented at test have been generally unrelated, resulting in only pairwise structure, as in the case of Swahili–English paired associates, and testing has been said to have its effects by means of retrieval practice. Similarly, investigations of metamemory have been largely restricted to the verbal learning tradition and memory for prose passages (see Dunlosky & Metcalfe, 2009). Findings of testing effects on classification and metacognition in the context of concept learning would extend knowledge of testing effects and provide a basis for broadening theorizing about metacognition.

Testing effects on classification learning might differ from those on learning to retrieve responses to unrelated cues. When informed about the advantages of testing for memory, educators sometimes respond that those advantages might be restricted to memory for facts and not extend to higher levels of learning, such as the learning of concepts (H. L. Roediger, personal communication, April 13, 2010). In this vein, Bloom's (1956) taxonomy of learning describes recall and recognition as requiring "knowledge," which is held to be a lower level of learning than is "comprehension," which is said to be required for classification. It could be argued that the focus on memory for particulars (facts or exemplars) that is produced by retrieval practice is counter to the discovery of general characteristics that are required for higher levels of intel-

lectual behavior, such as classification learning. A claim of this sort can be seen as resting on an abstractionist theory of concept learning and ignores the possibility that memory for exemplars or prototypes might serve as a basis for concept learning (for a review of theories of concept learning, see Murphy, 2002). Regardless, concerns such as potential tradeoffs between focus on memory for particulars and the learning of general characteristics points to the possibility that testing effects in classification learning might differ from those found in tasks in which testing effects typically have been examined.

Although a theory of testing effects on classification learning ultimately will require theorizing about concept learning, our experiments were not designed to choose among theories of concept learning, nor do the results of our experiments allow us to do so. Our interests differ from those that have motivated most investigations of concept learning in that we are interested in specifying conditions that optimize the learning of natural concepts. In contrast, laboratory investigations of concept learning typically have focused on choosing among theories, using artificial materials (for an extensive review, see Murphy, 2002). Relatively little research or theorizing in the concept-learning literature has been aimed at specifying manipulations that optimize classification learning. Investigations of perceptual expertise (e.g., Gauthier & Tarr, 1997; Gauthier, Williams, Tarr, & Tanaka, 1998; Tanaka, Curran, & Sheinberg, 2005; Tanaka & Taylor, 1991) are more directly relevant, but the primary focus of that research has been on the effects of expertise rather than on optimizing its development.

To our knowledge, no studies have investigated testing effects on classification of nonstudied exemplars of natural concepts. However, concept-learning experiments using artificial materials have examined such effects by comparing results from feedback and observational training conditions (e.g., Ashby, Maddox, & Bohil, 2002; Estes, 1976, 1994; Izawa, 1967; Reber & Millward, 1968). *Observational* training is similar to repeated study in that an exemplar is presented before or simultaneously with its category label on all trials during training. In contrast, *feedback* training includes retrieval practice in that exemplars are presented and participants are asked to retrieve the category label. Participants' responses are followed by presentation of the category label as feedback and as an opportunity for study. Feedback training is typically found to enhance categorization performance for studied and novel exemplars beyond that of observational training.

A difference between testing effects as investigated in the memory literature and feedback training is that for investigations of testing effects in the memory literature, the number of study presentations and retrieval attempts are varied. As done in memory experiments, we varied the number of study presentations, along with the number of tests, to compare the relative value of the two. We examined testing effects on recognition memory and classification accuracy for studied and novel exemplars of bird families. Compared with repeated study, we expected repeated testing to result in more attention and extensive processing of exemplars and expected this further processing to enhance later recognition memory of studied exemplars. Although not directly relevant to our experiments, the finding of an advantage of feedback training over observational learning in investigations of concept learning gave reason for us to predict that we would also find beneficial effects of testing on the learning of natural concepts. We predicted that repeated testing, compared with repeated study, would enhance later classification of both studied and novel exemplars.

An advantage of repeated testing for classification learning might come about because testing results in greater attention to and memory for studied exemplars, as could be argued by advocates of an exemplar model of concept learning (e.g., Brooks, 1978; Medin & Schaffer, 1978; Nosofsky, 1988), or because testing encourages attempts to discover features that are shared by members of a bird family, or the abstraction of a prototype (e.g., Rosch, 1975). As illustrated by exemplars of the Thrasher family (see Figure 1), there were features that were typical of members of a family but no features that occurred for all exemplars of a family and could serve to identify uniquely family members (e.g., many but not all Thrashers have a long, slightly hooked beak). This is generally true for bird families. At best, then, the abstraction of rules to be used for classification can rely only on characteristic features, rather than on defining features.

Figure 1 illustrates the materials used in our experiments by showing exemplars from one of the several bird families that were presented for learning. Prior to training, participants were instructed that they would be tested later for their recognition memory of studied exemplars and would then be asked to classify the studied and novel exemplars of the bird families. For a *repeated study* condition, exemplars from a portion of the families were repeatedly presented for study accompanied by their family name, whereas for a *repeated testing* condition, exemplars from the remaining families were presented accompanied by their family name, and then the ability to correctly classify exemplars by producing their family name was repeatedly tested. Feedback following these attempts was varied across experiments, as was the retention interval between training and the later tests of recognition memory and classification. Various measures of metacognition, in the form of predictions (e.g., predicted probability of later accurate classification) and postdictions (e.g., confidence judgments for classifications made at test), were included in later experiments.

In Experiments 2 and 3, participants predicted their recognition memory and classification performance for each studied exemplar and rated their confidence in their classification of studied and novel exemplars at test. Results from these metacognitive measures relate to the question of whether participants were unaware of the beneficial effects of testing for classification learning, as has been found for memory tasks (e.g., Karpicke & Roediger, 2008). More important, perhaps, we also introduce a new metacognitive measure designed to evaluate judgments of learning at the category level. Category learning judgments (CLJs) were made by participants after the learning phase, and prior to test. For CLJs, participants were asked to predict the likelihood of correctly classifying novel exemplars from each of the studied families. This measure

Figure 1. Exemplars from the Thrasher family.

allowed us to assess testing effects on metacognition at the level of categories in terms of participants' sensitivity to differences in classification difficulty among the families.

Prior investigations of classification learning have not included measures of metacognition, nor has theorizing about metacognition often relied on experiments using materials that have an underlying category or similarity structure. Doing so has the potential to inform theories of concept learning and broaden theorizing about metacognition and, as described in the General Discussion, is also of interest from a pedagogical standpoint.

Experiment 1

Method

Participants. Forty Washington University (St. Louis, Missouri) undergraduates participated in exchange for course credit or $10. All participants were tested individually.

Design and materials. A 2 (Training: repeated study vs. repeated testing) × 2 (Exemplar: studied vs. novel) within-participants design was used. In the repeated study condition, exemplars were presented along with their family names on four study trials (SSSS). In contrast, the repeated testing condition included an initial study trial and three subsequent test trials followed by feedback ($ST^ST^ST^S$). Eight bird families were randomly divided into two groups of four families. Each group was assigned to either the repeated study or repeated testing condition. In each group, 10 exemplars from each family were divided into two subgroups of five exemplars, with each subgroup being assigned to either the studied or novel exemplar condition. Groups were rotated through study conditions, and the subgroups were rotated through the exemplar conditions such that, across formats, materials were the same for the different conditions.

Color images of perching birds from the taxonomic order Passeriformes were obtained from http://www.whatbird.com, a website for bird identification. The images were equally scaled and presented against a light blue background (see Figure 1 for examples). Families were chosen from the same taxonomic order such that there was sufficient between-family similarity to avoid ceiling effects. Further, exemplars were chosen such that the genus of each was specific to each family, thus providing additional within-family similarity. Experiment 1 included exemplars from each of the following families: Buntings, Finches, Jays, Orioles, Swallows, Thrashers, Thrushes, and Wrens.

Procedure. Participants first completed a training phase. On each trial, birds were presented individually for 5 s in the center of a computer monitor with a family name or test prompt presented below, followed by a 500-ms interstimulus interval. Forty exemplars (five exemplars from each of eight families) were presented four times each for a total of 160 trials. For repeated study, the family name was displayed below the exemplar and read aloud on each trial. For the repeated testing condition, the family name was displayed below the exemplar and read aloud on the first trial, and the name was replaced by a row of question marks (???) that prompted classification attempts on the remaining trials. The experimenter keyed the classification attempts into the computer. Feedback displaying the message "correct" or "incorrect" was then presented for 1.5 s along with the correct family name. Participants were encouraged to remember the exemplars for an upcoming memory test. Repetitions were spaced by five, seven, or nine intervening items. Blocks with varying repetition lags were arranged in a fixed random order to decrease the likelihood that participants would detect a pattern.

At the time of test, participants made recognition memory and classification decisions for studied and novel exemplars. Ten exemplars (five studied, five novel) from each of the eight families were each presented once, for a total of 80 exemplars. Exemplars were presented in a fixed random order, with the restriction that exemplars from the same condition did not appear on more than three consecutive trials. The order of items on the test list remained fixed across participants, and groups of exemplars were rotated through list positions. Exemplars remained on the screen until a response was made. For recognition memory decisions, participants were told to respond "old" to studied exemplars and "new" to novel exemplars. After each old/new decision, all eight family names appeared below the exemplar. Participants then selected the family to which the exemplar belonged, regardless of whether it had been judged "old" or "new." All responses were made aloud and recorded by the experimenter.

Results and Discussion

In each of the experiments, effects reported as being significant were significant beyond $\alpha = .05$, unless otherwise stated. Accuracy in the repeated testing condition increased across the three attempts at naming the family of tested exemplars (.82, vs. .94 vs. .97.), $F(2, 78) = 68.79$, $\eta_p^2 = .64$.

Recognition memory. Testing effects on recognition memory (see Table 1) were largely limited to an influence on the probability of a hit, calling studied exemplars "old," with there being little difference between conditions in false alarms, calling novel exemplars "old." The probability of a hit was higher in the repeated testing condition than in the repeated study condition, $t(39) = 4.75$, $d = 0.73$. The manipulation of training condition did not significantly influence the probability of a false alarm, $t(39) = 1.42$.

Table 1
Proportion of "Old" Responses for Recognition Memory and Proportion of Correctly Classified Exemplars as a Function of Study and Exemplar Conditions: Experiment 1

	Study condition			
	SSSS		$ST^ST^ST^S$	
Test	*M*	*SEM*	*M*	*SEM*
Recognition				
Studied	.72	.02	.81	.02
Novel	.20	.02	.17	.02
Classification				
Studied	.71	.03	.79	.02
Novel	.45	.02	.53	.02

Note. The study condition labels represent the order of study trials (S) and test trials with feedback (T^S).

A signal-detection analysis examined differences in discriminability (d′) and criterion (c; see Macmillan & Creelman, 1991).[1] Results of that analysis revealed that repeated testing increased d′ (1.97 vs. 1.57), $t(39) = 3.90$, $d = 0.87$, and resulted in less conservative responding (smaller c) than did repeated study (.003 vs. .015), $t(39) = 2.86$, $d = 0.32$.

Perhaps the influence on criterion was produced by repeated testing enhancing memory for characteristic features that were shared by studied and novel family members (e.g., the long, slightly hooked beak shared by some Thrashers), as well as memory for more global features and other features that were unique to individual, studied exemplars. Enhanced memory for shared features would result in increased false alarms, whereas enhanced memory for features that were unique to studied exemplars would produce an increase in hits and a decrease in false alarms. The advantage in hits for the repeated testing condition, along with offsetting effects on false alarms, would produce the observed results of an increase in hits with little difference in false alarms, producing effects on both d′ and c.

The possibility that false alarms reflected category-relevant information was supported by the finding that novel exemplars were more likely to be falsely recognized as old if they were correctly classified (.23) rather than incorrectly classified (.14), $t(39) = 4.00$, $d = 0.76$. The finding of a higher false-alarm rate for exemplars that were correctly classified can be explained in terms of either a prototype model (Posner & Keele, 1968) or an exemplar model (Hintzman, 1986) of concept learning, as well as by a model that postulates the abstraction of characteristic features.

Classification. Studied exemplars were classified more accurately than were novel exemplars, $F(1, 39) = 197.66$, $\eta_p^2 = .84$. The bottom two rows of Table 1 show that classification performance for both studied and novel exemplars was more accurate in the repeated testing condition than in the repeated study condition, $F(1, 39) = 16.30$, $\eta_p^2 = .30$. The interaction of study and exemplar conditions did not approach significance, $F < 1$.

In sum, Experiment 1 demonstrated that testing effects extend to the learning of natural concepts. Both recognition memory and classification performance were facilitated by repeated testing.

Experiment 2

In Experiment 2, effects of repeated study were compared with those of two levels of testing so as to investigate trade-offs between repeated study and testing. Training conditions included a repeated study condition with six study trials, an intermediate testing condition with three study trials followed by three test trials, and an extended testing condition with one study trial followed by five test trials. As in Experiment 1, tests were followed by feedback. We expected that recognition and classification performance would be best in the extended testing condition.

In addition to assessing recognition and classification performance, the accuracy of metacognitive judgments was examined. After training, participants predicted their performance on later tests by making recognition memory and classification judgments of learning (JOLs), judging the probability of their later recognizing and that of their later correctly classifying the exemplars presented during training. Participants also made category learning judgments (CLJs), being given the family names and predicting future classification performance for novel exemplars from each of the families. Finally, at the time of test, participants made confidence judgments regarding the likelihood of their correctly classifying studied and novel exemplars.

We examined the extent to which testing effects on metacognitive judgments were the same as testing effects on accuracy of recognition memory and classification performance. As an example, one question is whether testing increases the prediction of future accurate classification as well as the actual accuracy of classification performance. For these comparisons, metacognitive judgments and accuracy are averaged across items representing each of the testing conditions. We also examined the resolution of metacognitive measures, which refers to the correlation at the item level between a metacognitive measure and the accuracy of performance. As an example, the gamma correlation between confidence judgments and accuracy at the level of items was computed for each participant and then analyzed by means of an analysis of variance (Nelson, 1984). Doing so allows one to determine whether judgments that were held in highest confidence were ones that were most likely to be accurate.

Experiments in the metacognition literature (for a review, see Dunlosky & Metcalfe, 2009) have not generally examined classification learning, and investigations of classification learning have not included measures of metacognition (for a review, see Murphy, 2002). Prior experiments examining testing effects have not examined effects on confidence judgments, but we expected effects on confidence to be similar to those on accuracy. Predictions of future classification performance (JOLs or CLJs) also have not been examined.

To examine the accuracy of predictions at the category level, we correlated CLJs for each family with classification accuracy at the level of families and averaged them across exemplars for each family. This was done separately for studied and novel exemplars. The magnitude of the correlation between CLJs and classification accuracy at the level of families measures sensitivity to differences among families in the difficulty of classification. A high correlation provides evidence that participants were aware of which families produce the greatest difficulty for accurate classification. If CLJs and classification of novel exemplars have a common basis, one would expect CLJs and classification accuracy for novel exemplars at the level of categories to be highly correlated. As is later discussed, the accuracy of metacognition at the category level is potentially important from a pedagogical standpoint as well as for the development of theory.

Method

Participants. Thirty-six Washington University undergraduates participated in exchange for course credit or $10 per hour. All participants were tested individually.

Design, materials, and procedure. A 3 (Training: repeated study vs. intermediate testing vs. extended testing) × 2 (Exemplar: studied vs. novel) within-participants design was used. The repeated study condition consisted of six study trials (SSSSSS), the intermediate testing condition consisted of three initial study trials followed by three test trials with feedback ($SSST^ST^ST^S$), and the

[1] Hit rates of 1.0 and false alarm rates of 0 were dealt with by means of the standard correction of subtracting or adding $1 - 2N$, the number of observations in each cell.

extended testing condition consisted of one study trial followed by five test trials with feedback ($ST^ST^ST^ST^ST^S$). Ten exemplars were selected from 12 families that included the original eight families from Experiment 1 and four additional families (i.e., Flycatchers, Sparrows, Vireos, and Warblers), for a total of 120 exemplars. For counterbalancing purposes, the families were divided into three groups matched on overall recognition and categorization performance from Experiment 1 (and a pilot experiment that included families not included in Experiment 1) and rotated through the study conditions. The groups were further divided into two subgroups and rotated through exemplar conditions, for a total of 60 studied and 60 novel exemplars at test.

During training, exemplars were presented six times for 3 s each (360 presentations), accompanied by their family name or followed by a test. For tests, the family name was presented as feedback for responding. The lags between repetitions were five, six, seven, or eight intervening items. JOLs, CLJs, and confidence judgments were made on a scale from 0 (*wild guess*) to 100 (*certain correct*). Participants were encouraged to use the full range of the scale. For JOLs, they were instructed to use the scale to predict their ability to later recognize and to later classify studied exemplars. JOLs were made for recognition memory and for classification performance following the final presentation of each exemplar during training (postfeedback on retrieval attempts). After training, and prior to test, all 12 family names were presented individually in random order for category-level judgments. Participants were instructed to use the 0–100 scale to predict the accuracy of their later classification performance for novel exemplars from each of the families. Finally, at the time of test, confidence judgments were made for each exemplar following classification decisions. Other details of the materials and procedures were the same as in Experiment 1.

Results and Discussion

In the intermediate testing condition, accuracy increased from the first to second test, at which point it was near ceiling (.92 vs. .98 vs. .98), $F(2, 70) = 14.62$, $\eta_p^2 = .30$. In the extended testing condition, accuracy increased from the first to third test and remained near ceiling thereafter (.75 vs. .95 vs. .98 vs. .99 vs. .99), $F(4, 140) = 85.01$, $\eta_p^2 = .71$. Accuracy on the final test in the intermediate testing condition did not significantly differ from that of the extended testing condition (.98 vs. .99), $t(35) = .72$, $d = 0.28$.

Recognition memory. Recognition memory performance in Experiment 2 agreed with results found in Experiment 1 by showing that repeated testing enhanced recognition memory performance beyond that produced by repeated study (see Table 2). The effect of training condition was significant for the probability of a hit, $F(2, 70) = 5.91$, $\eta_p^2 = .14$. Pairwise comparisons revealed that only the difference between the repeated study and the extended testing conditions was significant, $t(35) = 3.45$, $d = 0.52$. Training condition did not produce a significant effect on false alarms ($F < 1$).

Although the main effect of study condition on d′ was not significant, $F(2, 70) = 2.00$, $p = .14$, $\eta_p^2 = .05$, d′ for the extended testing condition (1.57) was greater than that for the repeated study condition (1.41) when assessed with a one-tailed t test, $t(35) = 1.96$, $d = 0.39$. The main effect of study condition on the measure of criterion was also not significant, $F(2, 70) = 1.92$, $p = .16$, $\eta_p^2 = .05$, but the pattern of results agreed with results in Experiment 1. As revealed by a smaller value of c, responding was less conservative in the extended testing condition compared with the repeated study condition (.001 vs. 013), when assessed with a one-tailed test, $t(35) = 1.88$, $d = 0.35$.

Table 2
Proportion of "Old" Responses for Recognition Memory, and Proportion of Correctly Classified Exemplars as a Function of Study and Exemplar Conditions: Experiment 2

	Study condition					
	SSSSSS		$SSST^ST^ST^S$		$ST^ST^ST^ST^ST^S$	
Test	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
Recognition						
Studied	.68	.03	.72	.02	.76	.02
Novel	.21	.02	.21	.02	.21	.02
Classification						
Studied	.60	.03	.69	.03	.73	.03
Novel	.38	.03	.44	.02	.46	.03

Note. The study condition labels represent the order of study trials (S) and test trials with feedback (T^S). Hits and false alarms correspond to the proportion of old responses for studied and novel exemplars, respectively.

Similar to results in Experiment 1, testing effects were restricted to hits, with no effect on false alarms. Again, the lack of an effect on false alarms might reflect offsetting effects of testing on memory for category relevant information (e.g., characteristic features by some members of a family), and effects on memory for features that were unique to studied exemplars. Enhanced memory for category-relevant information would be expected to increase false alarms to novel exemplars, whereas enhanced memory for unique features would provide a basis for rejecting novel exemplars that did not have those features. In line with the possibility that false alarms reflected category-relevant information, the probability of a false alarm was higher for novel exemplars that were correctly classified (.28) than for novel exemplars that were not correctly classified (.17), $t(35) = 5.49$, $d = 0.99$.

Classification. Studied exemplars were classified more accurately than were novel exemplars, $F(1, 35) = 210.02$, $\eta_p^2 = .86$. As shown in the bottom two rows of Table 2, repeated testing enhanced classification performance for both studied and novel exemplars. The main effect of training condition was significant, $F(2, 70) = 10.88$, $\eta_p^2 = .24$, and the interaction between training and exemplar conditions did not approach significance ($F < 1$). Paired-samples t tests revealed that classification accuracy, collapsed across studied and novel exemplars, for the intermediate (.57) and extended testing (.60) conditions were both greater than that for the repeated study condition (.49), $ts(35) \geq 3.18$, $ds \geq .54$, but did not significantly differ from one another, $t(35) = 1.39$.

Again, the classification results provide evidence that testing effects extend to the learning of natural concepts. However, extended testing provided little advantage over a mixture of repeated study and repeated testing in the intermediate testing condition.

Metacognitive judgments. Results for metacognitive judgments are presented in Table 3. Testing effects on the magnitude of predictions of performance, measured by JOLs, for recognition memory and classification, were in general agreement with effects

Table 3
Metacognitive Judgments for Recognition Memory and Classification Performance as a Function of Study and Exemplar Conditions: Experiment 2

	Study condition					
	SSSSSS		$SSST^ST^ST^S$		$ST^ST^ST^ST^ST^S$	
Measure	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
JOLs						
Recognition (studied)	.67	.03	.76	.03	.75	.03
Classification (studied)	.63	.03	.73	.03	.72	.03
Confidence						
Classification (studied)	.60	.03	.66	.03	.68	.03
Classification (novel)	.45	.03	.48	.03	.48	.03
CLJs						
Classification (novel)	.52	.03	.62	.03	.63	.03

Note. JOL = judgment of learning; CLJ = category learning judgment. The study condition labels represent the order of study trials (S) and test trials with feedback (T^S).

on the accuracy of performance. Testing effects on JOLs were significant for both recognition memory, $F(2, 70) = 17.01$, $\eta_p^2 = .33$, and classification, $F(2, 70) = 18.49$, $\eta_p^2 = .35$. JOLs were lower for the repeated study condition than for the intermediate testing condition for both recognition memory and classification judgments, $ts(35) \geq 5.33$, $ds \geq .58$. JOLs in the extended testing condition did not differ significantly from those in the intermediate testing condition for recognition memory or classification judgments ($ts < 1$).

To measure the resolution of predictions, gamma correlations (G) between JOLs and accuracy were computed for all participants (see Nelson, 1984) and then analyzed by means of an analysis of variance. Gammas for the correlation between recognition JOLs and accuracy could not be computed for two participants, and gammas for classification JOLs could not be computed for one participant because of lack of variability (constant values) on at least one of the variables. In contrast to effects found for the magnitude of JOLs, training condition did not significantly influence the resolution of JOLs for either recognition memory ($G = .12$), or classification ($G = .31$; $Fs < 1$). Thus, although testing increased JOLs averaged across items, testing did not increase the ability to predict the particular items for which recognition memory judgments or classification would be accurate.

Confidence in classification was higher for studied than for novel exemplars, $F(1, 35) = 247.79$, $\eta_p^2 = .88$. Testing effects on the magnitude of confidence judgments were revealed by a significant main effect of study condition, $F(2, 70) = 6.76$, $\eta_p^2 = .16$. The interaction of training condition with studied versus novel exemplars, $F(2, 70) = 3.30$, $\eta_p^2 = .09$, revealed that the effects of training condition were largely restricted to studied exemplars. Confidence judgments for studied exemplars were greater in the intermediate and extended testing conditions than in the repeated study condition, $ts(35) \geq 2.86$, $ds \geq .32$. In contrast, confidence judgments in the intermediate and extended testing conditions did not differ from those in the repeated study condition for novel exemplars, $ts(35) \leq 1.58$, $ds \leq .15$.

An analysis of the resolution of confidence judgments by means of gamma correlations between confidence and accuracy excluded one participant for whom a correlation could not be computed because of constant values on one variable. Results from that analysis revealed that resolution for confidence judgments was greater for studied (.68) than for novel (.50) exemplars, $F(1, 34) = 19.61$, $\eta_p^2 = .37$. In addition, resolution increased across repeated study, intermediate testing, and extended testing conditions (.50 vs. .60 vs. .68), $F(2, 68) = 5.59$, $\eta_p^2 = .14$. The interaction between study and exemplar conditions did not approach significance ($F < 1$). That is, testing had the desirable effect of increasing the correlation between confidence and accuracy at the level of items for both studied and novel exemplars.

Metacognitive judgments at the category level. The magnitude of CLJs varied across study conditions, $F(2, 70) = 9.83$, $\eta_p^2 = .22$, producing a pattern of results that agreed with that produced by training condition on classification accuracy for novel exemplars. Repeated study produced lower CLJs than did the intermediate testing condition, $t(35) = 3.59$, $d = 0.55$, which did not significantly differ from the extended testing condition, $t(35) = .52$, $d = 0.08$.

Resolution at the category level was measured by computing a Pearson product-moment correlation between CLJs, averaged across participants for each family, and averaged classification accuracy for each of the families. There was a strong positive correlation between mean CLJs and mean classification performance for novel exemplars, averaged at the levels of families and participants ($r = .82$). Averaged CLJs were also strongly correlated with classification accuracy for studied exemplars averaged across exemplars within a family and participants ($r = .95$). These results suggest that CLJs were based on accurate predictions of differences among families in the difficulty of classifying studied and novel exemplars.

To show that individual participants were aware of differences among the families in the difficulty of classifying exemplars, we computed Pearson product-moment correlations between CLJs and mean classification accuracy for novel exemplars from each family for each participant. That correlation was significant ($r = .28$, $p < .001$) but much smaller than was the correlation computed by collapsing across individuals. It is not surprising that the correlation found at the level of participants was smaller than that found by averaging across participants. Collapsing across participants

increases the reliability of measures because of the larger number of observations contributing to each score. Also, analyzing at the level of participants confounds effects of training conditions with differences among families because of the counterbalancing of materials across training conditions. We do not report analyses of the effect of training conditions on the correlation of CLJs with accuracy. Analyzing effects of training conditions for individual participants not only confounds effects of training conditions with materials but also results in only four observations for each of the training conditions. This is because each training condition was represented by only four families for each participant.

The high correlation of CLJs with classification accuracy of studied exemplars suggests that differences in classification difficulty among studied exemplars served as a basis for CLJs. An alternative basis for CLJs is reliance on more general information, such as features that are characteristic of a particular family. As an example, the long, slightly hooked beak (see Figure 1) is a salient characteristic of some members of the Thrasher family, and this characteristic might be used as a basis for predicting that novel Thrashers are likely to be relatively easy to identify.

Experiment 3

Results from Experiment 2 revealed that additional testing in the extended testing condition tended to further enhance recognition and classification performance beyond that of the intermediate testing condition. In Experiment 3, we examined whether this trend depended on the additional study opportunities provided by corrective feedback following tests. Experiment 3 included the same conditions as Experiment 2, with the exception that no feedback was provided following tests. We expected the removal of feedback to be particularly disadvantageous for the extended testing condition, the condition that engaged in five tests after a single study presentation. The greater opportunity for study in the intermediate testing condition might make feedback less important.

Positive effects of repeated testing have been shown to be larger on a delayed test than on a test that immediately follows study (for a review, see Roediger & Karpicke, 2006a). In Experiment 3, we compared testing effects in recognition memory and classification performance on immediate and delayed tests. We expected the positive effects of testing to be larger after the longer retention interval. Experiment 3 included the same metacognition measures as did Experiment 2 so as to allow us to examine the effects of feedback and those of test delay on metacognition.

Method

Participants. Seventy-two Washington University undergraduates participated in exchange for course credit or $10 per hour. Thirty-six participants were randomly assigned to each between participant condition (test: immediate vs. delayed). All participants were tested individually.

Design, materials, and procedure. The design, materials, and procedure were identical to those used in Experiment 2, except that there was no feedback following retrieval attempts. Also, retention interval (immediate vs. delayed) was manipulated between participants. Participants in the delayed testing condition returned to the lab 1 day following study at the same time as their initial arrival.

Results and Discussion

The accuracy of performance on tests during training did not differ between immediate and delayed test conditions ($F < 1$). There were also no significant interactions between accuracy during study and delay of testing for intermediate testing or extended testing conditions (Fs < 1). Collapsed across test delays, accuracy during study in the intermediate testing condition increased slightly from the first to second test (.92 vs. .93 vs. .93), $F(2, 142) = 5.50$, $\eta_p^2 = .07$. Accuracy in the extended testing conditions also increased across tests (.78 vs. .83 vs. .83 vs. .84 vs. .85), $F(4, 184) = 25.51$, $\eta_p^2 = .36$. Accuracy on the final test during training was greater for the intermediate than for the extended testing condition (.93 vs. .85), $t(47) = 4.68$, $d = 0.77$.

It is interesting that classification accuracy increased across tests, although feedback was not given. The high level of classification accuracy on the first test trial might be important for this finding. The presence of a large number of correctly classified items may have resulted in the discovery of shared features across trials that allowed correct classification of exemplars that were not initially correctly classified. Alternatively, from the vantage point of an exemplar view of concept learning, repeated testing of correctly classified exemplars might have resulted in changes in generalization of a sort that improved the classification accuracy of exemplars that were not initially correctly classified.

Recognition memory. An analysis of the recognition memory results (see Table 4) revealed a significant effect of training condition on the probability of a hit, $F(2, 140) = 17.53$, $\eta_p^2 = .20$, as well as a significant effect of test delay, $F(1, 70) = 9.38$, $\eta_p^2 = .12$. The interaction of the two variables was not significant, $F(2, 140) = 1.08$. Delaying the recognition memory test reduced the probability of a hit. Pairwise comparisons of training conditions, collapsed across immediate and delayed tests, revealed that the probability of a hit in the repeated study condition (.66) was lower than that in both the intermediate testing condition (.72), $t(71) = 4.06$, $d = 0.40$, and the extended testing condition (.75), $t(71) = 5.80$, $d = 0.57$. The difference in the probability of a hit in the latter two conditions was not significant, $t(71) = 1.70$. The analysis of false alarms revealed only a significant interaction between training condition and test delay, $F(2, 140) = 3.61$, $\eta_p^2 = .05$. The interaction arose because the probability of a false alarm was lower for the intermediate testing condition than for the other two conditions on the immediate test but higher than that of the other two conditions on the delayed test. We do not have an explanation for this interaction.

A signal-detection analysis of discriminability revealed that d′ was greater for the immediate than for the delayed test (1.69 vs. 1.32), $F(1, 70) = 8.51$, $\eta_p^2 = .11$.

Replicating results from Experiments 1 and 2, repeated testing produced higher subsequent recognition performance than did repeated study (see Table 4), $F(2, 140) = 8.16$, $\eta_p^2 = .10$. Discriminability was poorer in the repeated study (1.32) than in either the intermediate testing (1.56) or the extended testing (1.62), condition, ts$(71) \geq 2.71$, ds $\geq .32$, whereas d′ did not significantly differ for the latter two conditions ($t < 1$).

Analyses of differences in criteria did not reveal a significant main effect of test delay, $F(1, 70) = 2.44$, $p = .12$, $\eta_p^2 = .03$, but did reveal a main effect of training condition: repeated study (.021), intermediate testing (.013), and extended testing (.006),

Table 4
Proportion of "Old" Responses for Recognition Memory, and Proportion of Correctly Classified Exemplars as a Function of Study, Exemplar, and Test Conditions: Experiment 3

	Study condition					
	SSSSSS		SSSTTT		STTTTT	
Test	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
Recognition (immediate)						
Studied	.71	.03	.76	.02	.80	.02
Novel	.22	.03	.18	.02	.22	.02
Recognition (delayed)						
Studied	.60	.03	.69	.03	.70	.03
Novel	.21	.02	.23	.02	.21	.02
Classification (immediate)						
Studied	.62	.03	.70	.02	.67	.03
Novel	.39	.03	.44	.03	.41	.03
Classification (delayed)						
Studied	.55	.03	.63	.03	.60	.04
Novel	.38	.03	.46	.03	.39	.03

Note. The study condition labels represent the order of study trials (S) and test trials (T). Hits and false alarms correspond to the proportion of old responses for studied and novel exemplars, respectively.

$F(2, 140) = 6.93$, $\eta_p^2 = .09$. Repeated study produced more conservative responding than did intermediate or extended testing, $ts(71) \geq 2.07$, $ds \geq .21$. The difference between the intermediate and extended testing conditions was not significant, $t(71) = 1.51$.

As in earlier experiments, effects of testing during training were largely limited to hits, producing almost no influence on false alarms. Again, this result can be explained as produced by testing enhancing memory for category-relevant information that serves as a basis for false alarms to novel exemplars as well as for features that were unique to individual, studied exemplars. Both enhanced memory for category-relevant information and for unique features would increase hits, but the two would produce opposite effects on false alarms. In agreement with results from earlier experiments, false alarms were higher for novel exemplars that were correctly classified (.29) than for novel exemplars that were not correctly classified (.16), $t(71) = 8.49$, $d = 0.91$.

Classification. Classification performance (see Table 4) was better for studied than for novel exemplars, $F(1, 70) = 340.22$, $\eta_p^2 = .83$, and this difference was greater following immediate compared to delayed testing, $F(1, 70) = 7.94$, $\eta_p^2 = .10$. Testing effects on classification accuracy were revealed by a significant main effect of training condition, $F(2, 140) = 9.50$, $\eta_p^2 = .12$, with no significant interactions between training condition and exemplar, or study condition and test delay ($Fs < 1$).

Classification performance revealed testing effects similar to those seen in Experiments 1 and 2. However, unlike results from earlier experiments, classification performance was highest in the intermediate testing condition. Classification accuracy for the intermediate testing condition (.56) was significantly greater than that for the repeated study (.48), $t(71) = 4.33$, $d = 0.52$, as well as the extended testing condition (.52), $t(71) = 2.18$, $d = 0.26$. Accuracy for the extended testing condition was greater than for the repeated study condition, $t(71) = 2.25$, $d = 0.25$.

Comparing results of Experiments 2 and 3 reveals that eliminating corrective feedback reduced performance only for the condition that had a single study trial followed by five tests. The greater opportunity for study in the other two conditions made feedback unimportant. These results suggest that repeated testing is more beneficial than additional study only when the opportunity for study is sufficient to produce a rather high probability of successful performance (cf. Gates, 1917).

In contrast to findings that delayed testing increases positive effects of testing for paired-associates and information presented in prose passages (for a review, see Roediger & Karpicke, 2006a), testing effects did not increase after a delay for either recognition memory or classification of exemplars of natural concepts. This difference likely reflects a difference in testing effects for the different types of materials. Learning to recognize and classify exemplars of natural concepts involves processes beyond those required for simple retrieval of a response, and those additional processes might be more resistant to forgetting than are effects on retrieval.

Increasing the delay of testing decreased hits in recognition memory and decreased the accuracy of classification for studied exemplars but did not influence the probability of a false alarm or classification accuracy for novel exemplars. This pattern of results can be explained by either a prototype model (Posner & Keele, 1968) or an exemplar model (Hintzman, 1986) of concept learning.

Metacognitive judgments. Metacognitive judgments in Experiment 3 (see Table 5) were analyzed in the same manner as in Experiment 2. As found in Experiment 2, testing effects on the magnitude of JOLs for recognition memory and classification were in general agreement with effects on accuracy of responding for the two types of test. There were significant testing effects on both JOLs for recognition memory, $F(2, 142) = 13.89$, $\eta_p^2 = .16$, and JOLs for classification, $F(2, 142) = 17.27$, $\eta_p^2 = .20$. For recognition memory, differences were numerically small, but JOLs in the intermediate-testing condition were higher than those in either the repeated study or the extended testing condition, $ts(71) \geq 2.77$, $ds > .06$, and JOLs in the extended testing condition were higher than those in the repeated study condition, $t(71) = 2.76$, $d = 0.11$. For classification JOLs, the intermediate testing condition pro-

Table 5
Metacognitive Judgments for Recognition Memory and Classification Performance as a Function of Study and Exemplar Conditions, Collapsed Across Test Conditions: Experiment 3

	Study condition					
	SSSSSS		SSSTTT		STTTTT	
Measure	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
JOLs						
Recognition (Hits)	.71	.02	.75	.02	.73	.02
Classification (Studied)	.61	.02	.67	.02	.61	.02
Confidence						
Classification (Studied)	.55	.02	.60	.02	.58	.02
Classification (Novel)	.41	.02	.45	.02	.44	.02
CLJs						
Classification (Novel)	.51	.02	.60	.02	.57	.03

Note. JOL = judgment of learning; CLJ = category learning judgment. The study condition labels represent the order of study trials (S) and test trials (T).

duced higher JOLs than did either the repeated study or the extended-testing condition, $ts(71) \geq 4.78$, $ds \geq .31$. JOLs in the extended testing condition did not differ from those in the repeated study condition, $t(71) = .79$, $d = 0.05$.

Resolution for recognition JOLs could not be computed for six participants because of constant values. Gamma correlations for the remaining participants did not significantly differ across training conditions, and there was not a significant interaction between training and test conditions ($Fs < 1$). However, resolution was greater following delayed compared to immediate testing ($G = .29$ vs. $G = .10$), $F(1, 64) = 5.62$, $p = .02$, $\eta_p^2 = .08$. The analysis of resolution for classification JOLs ($G = .33$) excluded three participants because of constant values for at least one of the variables. Gamma correlations for the remaining participants did not differ across training conditions or test delay, and there was not a significant interaction between training and test ($Fs < 1$).

In sum, the general consistency between predicted performance, measured by JOLs, and the accuracy of observed recognition and classification performance suggests that participants were able to predict the beneficial effects of testing. However, testing did not increase the resolution of predictions.

Confidence judgments showed sensitivity to testing effects similar to that of JOLs. Studied exemplars were classified with higher confidence than were novel exemplars, $F(1, 71) = 245.76$, $\eta_p^2 = .78$. Testing during training produced an effect on confidence judgments that agreed with effects on accuracy, $F(2, 142) = 14.06$, $\eta_p^2 = .17$. Confidence in classification was lower in the repeated study condition than in either the intermediate testing, $t(71) = 5.33$, $d = 0.28$, or the extended testing condition, $t(71) = 3.87$, $d = 0.17$. The advantage of the intermediate testing condition over the extended testing condition in confidence of classification was numerically small and only marginally significant, $t(71) = 1.65$, $p = .10$, $d = 0.10$. There was not a significant main effect of test delay, nor was there an interaction of training, exemplars, and test ($Fs < 1$).

Resolution for confidence judgments did not differ across training conditions ($F < 1$) or test conditions, $F(1, 66) = 1.50$, $p = .23$, $\eta_p^2 = .02$. However, resolution was greater for studied ($G = .65$) than for novel ($G = .50$) exemplars, $F(1, 66) = 26.98$, $\eta_p^2 = .29$. The lack of an effect of training condition on the resolution of confidence judgments contrasts with the finding of such an effect in Experiment 2 and suggests that feedback is important for testing effects to enhance the resolution of confidence judgments.

Metacognitive judgments at the category level. Analyses of judgments at the category level revealed significant testing effects, $F(2, 142) = 13.43$, $\eta_p^2 = .16$, that agree with those found on the accuracy of classification of novel exemplars. Predictions for classification of novel exemplars, measured by CLJs, were higher in both the intermediate and extended testing condition than in the repeated study condition, $ts(71) \geq 3.86$, $ds \geq .29$. CLJs in the extended testing condition did not significantly differ from those in the intermediate testing condition, $t(71) = 1.38$, $d = 0.13$. The interaction of testing effects with test delay was not significant, $F(2, 140) = 2.13$, $p = .12$, $\eta_p^2 = .03$, nor was the main effect of delay on CLJs ($F < 1$).

To show that individual participants were aware of differences among families in the difficulty of classifying exemplars, we computed the Pearson product-moment correlation between CLJs and mean classification accuracy for novel exemplars from each family, collapsed across retention interval, for each participant. That correlation was significant ($r = .29$) but much smaller than was the correlation computed by averaging across individuals. Averaged across individuals, CLJs were strongly correlated with classification accuracy for novel exemplars ($r = .79$) and studied exemplars ($r = .93$) at the category level (averaged across exemplars).

General Discussion

Results from the current experiments add the learning of natural concepts to the list of tasks that benefit from testing effects. Each of the experiments revealed an advantage of testing over repeated study in recognition memory of studied exemplars and in classification accuracy for studied and novel exemplars. However, comparison of results from Experiments 2 and 3 suggests that testing is more beneficial than additional study only when the opportunity for study is sufficient to produce a rather high probability of successful performance on tests (cf. Gates, 1917). In contrast to results from experiments using other tasks and materials (e.g., Roediger & Karpicke,

2006b), testing effects on classification learning were not larger for a delayed test than for an immediate test. Also, whereas prior research has shown participants to be generally unaware of the beneficial effects of testing (for a review, see Karpicke et al., 2009), predictions of testing effects on later performance, as measured by JOLs, were in general agreement with effects on accuracy of recognition and classification performance. Confidence judgments in classification also revealed beneficial effects of testing but increases in their resolution appear to require the provision of feedback.

How does testing have its effects? Others have argued that testing effects depend on test-appropriate processing (e.g., Roediger & Karpicke, 2006a) or desirable difficulties (Bjork, 1994, 1999). To go beyond such largely circular accounts, it is necessary to gain a better understanding of what is being learned. For the learning of natural concepts, this means that testing effects must be understood in the context of theorizing about concept learning.

Rather than focusing on conditions that optimize the learning of natural concepts, investigations of concept learning have typically been aimed at choosing among theories by using artificial materials (for a review, see Murphy, 2002). The use of artificial materials is advantageous for theory testing but provides little guidance for means of optimizing the learning of natural concepts, which was the goal of our experiments. Our findings of effects of testing do not help to choose among theories of concept learning but, rather, can be accommodated by a number of theories, as could have been other outcomes. In this vein, an exemplar model of categorization (e.g., Medin & Shaffer, 1978) could accommodate our finding that testing produced parallel, beneficial effects on recognition memory and concept learning. The beneficial effects of testing on recognition memory can be explained as resulting from greater attention being given to the processing of studied exemplars so as to meet the demands of the tests. For concept learning, it can be held that exemplars have to be sufficiently well learned to serve as a useful source of generalization and that retrieval practice enhances learning.

However, an exemplar model could also be used to accommodate an opposite result, a finding that testing reduces correct classification of new exemplars. Testing has been shown to improve recognition memory of letter strings generated from an artificial grammar while having the effect of reducing the correct classification of new exemplars (e.g., Reber & Allen, 1978; Vokey & Brooks, 1992). This result was explained within the context of an exemplar model by arguing that overdifferentiation of exemplars in memory reduces generalization and, so, has detrimental effects on concept learning (cf. Vokey & Brooks, 1992). An obvious solution for explaining opposite effects is to argue that testing effects depend on the details of the similarity space and to suggest that similarity among exemplars should be measured prior to training. However, a problem for that solution is that similarity among exemplars might change as a function of training (cf. Gauthier et al., 1998). Alternative theories of concept learning could also accommodate our findings of beneficial effects of testing as well as negative effects of testing. For example, if negative effects of testing were found, it could be argued that testing focused on memory for the details of studied exemplars at the expense of discovering features that were shared by most exemplars of a category.

Testing effects along with their underlying bases are likely to differ across tasks, materials, and what is tested (e.g., ability to retrieve the family name of studied exemplars vs. knowledge of characteristic features that are shared by studied exemplars). The form of testing might influence a participant's approach to a task along with the bases for concept learning. Further investigation of testing effects is required to examine such possibilities so as to potentially extend theorizing about concept learning and to develop a theory of testing effects in the context of concept learning. Returning to the educator's concern that testing might focus on particulars at the cost of more general or higher forms of learning, our experiments did not reveal effects of that sort. However, such tradeoffs might sometimes occur.

As described by Dunlosky and Metcalfe (2009), investigations of metacognition have typically used materials such as word lists and short text passages. In contrast to such materials, investigating metacognition in the context of learning natural concepts allowed us to examine CLJs, participants' predictions of their ability to identify novel exemplars from studied families. There was a high correlation between CLJs and classification of novel exemplars at the category level, averaged across individuals, and the correlation continued to be significant when analyzed for individual participants. These results provide evidence that participants were aware of differences across categories in the difficulty of classifying exemplars. CLJs were also correlated with the accuracy of classifying studied exemplars at the level of categories, suggesting that memory for exemplars served as a basis for CLJs. Just as memory for exemplars preserves information about variability of exemplars within a set (e.g., Rips, 1989; Rips & Collins, 1993), memory for exemplars might also preserve information about differences across categories in the difficulty of classifying exemplars. An alternative to memory for exemplars as a basis for CLJs is reliance on characteristics that are typical of members of a family. For example, participants might notice that some Thrashers have a long, slightly hooked beak (see Figure 1) and use this characteristic as a basis for predicting that novel Thrashers are likely to be relatively easy to identify. More is needed to specify the bases for category level judgments. Perhaps there are multiple bases for CLJs just as has been argued for JOLs (e.g., Koriat, Ma'ayan, & Nussinson, 2006).

CLJs can be extended to other types of material and might be as useful as or more useful than measures that are currently used for purposes of theory as well as for pedagogical purposes. As an example of the latter, prior to an exam over several chapters, participants might make predictions regarding their learning of various topics in those chapters (e.g., theories of attention). Metacognitive measures at that level (a category level) might be more useful for guiding study than are measures at an overall level (What proportion of the items on the exam will you get correct?) or at the level of individual items (e.g., Will you remember that Broadbent's model is an early selection model of attention?). What are the conditions that produce overconfidence at the category level, resulting in too little time being devoted to study of a topic? What are the bases for predictions that one can correctly respond to novel exemplars or questions regarding a particular topic or category, and what conditions are important for determining the validity of those predictions? Questions such as these are important for educating study habits as well as for purposes of theory.

The current experiments are meant to be a step toward further identifying the conditions that are optimal for learning natural concepts. There are applied reasons for interest in specifying conditions that optimize the gaining of expertise in bird identification. For example, learning of natural concepts, including bird families, is a

common topic in the context of environmental studies. Means of optimizing learning to identify exemplars of bird families might also generalize to other classification tasks, such as identifying exemplars of families of disease, which is important for medical education. It is rather surprising that there has been relatively little research aimed at optimizing the learning of natural concepts. Focusing on optimizing the learning of natural concepts is an important direction for future research. Such research encourages the convergence of theorizing about memory, concept learning, and metacognition, as well as holding promise for applications to education.

References

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition, 30,* 666–677.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Bloom, B. S. (1956). *Taxonomy of educational objectives, handbook I: The cognitive domain.* New York, NY: David McKay.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Erlbaum.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135,* 553–571. doi:10.1037/0096-3445.135.4.553

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition.* Thousand Oaks, CA: Sage.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83,* 37–64. doi:10.1037/0033-295X.83.1.37

Estes, W. K. (1994). *Classification and cognition.* Oxford, England: Oxford University Press. doi:10.1093/acprof:oso/9780195073355.001.0001

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*(40).

Gauthier, I., & Tarr, M. J. (1997). Becoming a "greeble" expert: Exploring mechanisms for face recognition. *Vision Research, 37,* 1673–1682. doi:10.1016/S0042-6989(96)00286-6

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training "greeble" experts: A framework for studying expert object recognition processes. *Vision Research, 38,* 2401–2428. doi:10.1016/S0042-6989(97)00442-2

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428. doi:10.1037/0033-295X.93.4.411

Izawa, C. (1967). Function of test trials in paired-associate learning. *Journal of Experimental Psychology, 75,* 194–209. doi:10.1037/h0024971

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17,* 471–479. doi:10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L. (2008, February 15). The critical importance of retrieval for learning. *Science, 319,* 966–968. doi:10.1126/science.1152408

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationship between monitoring and control in metacognition: Lessons for the cause and effect relationship between subjective experience and behavior. *Journal of Experimental Psychology: General, 135,* 36–69. doi:10.1037/0096-3445.135.1.36

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*: New York, NY: Cambridge University Press.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19,* 494–513. doi:10.1080/09541440701326154

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207–238. doi:10.1037/0033-295X.85.3.207

Murphy, G. L. (2002). *The big book of concepts.* Cambridge, MA: MIT Press.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133. doi:10.1037/0033-2909.95.1.109

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 700–708. doi:10.1037/0278-7393.14.4.700

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363. doi:10.1037/h0025953

Reber, A. S., & Allen, R. (1978). Analogy and abstraction strategies in synthetic grammar learning: A functionalist interpretation. *Cognition, 6,* 189–221. doi:10.1016/0010-0277(78)90013-6

Reber, A. S., & Millward, R. B. (1968). Event observation in probability learning. *Journal of Experimental Psychology, 77,* 317–327. doi:10.1037/h0025760

Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511529863.004

Rips, L. J., & Collins, A. (1993). Categories and resemblance. *Journal of Experimental Psychology: General, 122,* 468–486. doi:10.1037/0096-3445.122.4.468

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Karpicke, J. D. (2006b). Test enhanced learning: Taking tests improves long-term retention. *Psychological Science, 17,* 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104,* 192–233. doi:10.1037/0096-3445.104.3.192

Tanaka, J. W., Curran, T., & Sheinberg, D. L. (2005). The training and transfer of real-world perceptual expertise. *Psychological Science, 16,* 145–151. doi:10.1111/j.0956-7976.2005.00795.x

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology, 23,* 457–482. doi:10.1016/0010-0285(91)90016-H

Vokey, J. R., & Brooks, L. R. (1992). Salience of item knowledge in learning artificial grammars. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 328–344. doi:10.1037/0278-7393.18.2.328

Received August 26, 2009
Revision received May 28, 2010
Accepted June 7, 2010 ■