Metacognitive judgments of repetition and variability effects in natural concept learning: evidence for variability neglect

Christopher N. Wahlheim, Bridgid Finn & Larry L. Jacoby

Memory & Cognition

ISSN 0090-502X

Mem Cogn
DOI 10.3758/s13421-011-0180-2

Springer

Mem Cogn
DOI 10.3758/s13421-011-0180-2

Metacognitive judgments of repetition and variability effects in natural concept learning: evidence for variability neglect

Christopher N. Wahlheim · Bridgid Finn · Larry L. Jacoby

Abstract In four experiments, we examined the effects of repetitions and variability on the learning of bird families and metacognitive awareness of such effects. Of particular interest was the accuracy of, and bases for, predictions regarding classification of novel bird species, referred to as *category learning judgments* (CLJs). Participants studied birds in high repetitions and high variability conditions. These conditions differed in the number of presentations of each bird (repetitions) and the number of unique species from each family (variability). After study, participants made CLJs for each family and were then tested. Results from a classification test revealed repetition benefits for studied species and variability benefits for novel species. In contrast with performance, CLJs did not reflect the benefits of variability. Results showed that CLJs were susceptible to accessibility-based metacognitive illusions produced by additional repetitions of studied items.

Keywords Category learning judgments · Concept learning · Metacognition · Repetitions · Variability

Imagine that you were asked to design a training regimen to teach people to classify instances of naturally occurring categories, such as birds from various families. What would be the best way to balance the number of presentations of each instance with the number of instances in each category? A common finding in the concept learning literature is that increasing the number of presentations of instances enhances the learning of studied items, whereas increasing the number of unique instances enhances later classification of novel items (e.g., Dukes & Bevan, 1967). That repetition improves performance on studied items is not surprising since repetitions have been shown to improve rote learning in a variety of memory tasks (for a review, see Crowder, 1976). More interesting is the finding that variability enhances the learning of underlying concepts. The bulk of studies investigating these effects have used artificial concepts, such as dot patterns, to control for the similarity of instances when varying their number (e.g., Homa, Cross, Cornell, Goldman, & Shwartz, 1973; Posner & Keele, 1968). However, few studies have investigated the benefits of variability on the learning of natural concepts, and no studies to our knowledge have examined metacognitive sensitivity to such effects.

We had two primary goals in the present experiments. The first goal was to examine the effects of repetitions and variability on the learning of natural concepts. The second and more central goal was to examine the extent to which participants could predict effects of repetitions and variability on such learning. In particular, we asked whether learners are aware that variability training facilitates classification of novel items more than does repetition of instances. To answer this question, we examined the extent to which predictions made at the level of individual categories were correlated with success in classification of novel instances of studied concepts.

Investigating predictions at the category level is potentially important for theories of metacognition as well as for applied purposes. For theory, metacognition has been studied extensively in the context of memory and decision making (for a review, see Dunlosky & Metcalfe, 2009), and these investigations have typically been conducted at the level of items and at the level of lists collapsed across unrelated items (but see Shah & Oppenheimer, 2011). Consequently, little is known about

C. N. Wahlheim (✉) · B. Finn · L. L. Jacoby
Department of Psychology, Washington University,
One Brookings Drive,
St. Louis MO 63130 Missouri, USA
e-mail: cnwahlheim@gmail.com

Published online: 27 January 2012

the bases for judgments regarding the learning of concepts (see Jacoby, Wahlheim, & Coane, 2010; Kornell & Bjork, 2008; Wahlheim, Dunlosky, & Jacoby, 2011). For applied purposes, sensitivity to differences in learning across categories has implications in settings such as education and medicine. In education, for example, when students prepare for an exam in a cognitive psychology course, accurate assessments of differences in the extent to which they have learned various topics (e.g., attention, memory, problem solving) might serve to guide the allocation of additional study. As an example in medicine, physicians might be aware of which diseases are particularly difficult to identify, and such knowledge may be useful for purposes of diagnosis. For both examples, the accuracy of predictions at the topic or category level is critical for guiding performance.

Metacognitive judgments have been examined at the level of categories in recent studies that were aimed at optimizing the learning of bird families (i.e., Jacoby et al., 2010; Wahlheim et al., 2011). In those experiments, species of birds (e.g., Blue Jay) that were subordinate level instances belonging to superordinate families (e.g., Jays) were first presented for study. Following study, family names appeared individually, and participants made predictions regarding the probability that they would correctly classify novel species from studied families on a future test. These judgments were referred to as *category learning judgments* (CLJs), and they reflected participants' evaluations of the extent to which their knowledge of concepts (i.e., bird families) would generalize to novel instances of those concepts (i.e., unstudied species). Results from these studies showed that CLJs were sensitive to performance differences produced by manipulations of study conditions. Jacoby et al. (2010) found that CLJs were sensitive to testing effects in that performance on studied and novel items was better following repeated testing than repeated study, with CLJs revealing a similar pattern. In addition, Wahlheim et al. (2011) found that CLJs were sensitive to benefits in performance on studied and novel items produced by spaced as compared with massed study.

In the present experiments, we extended our investigation of CLJs by examining their sensitivity to the effects of repetitions and variability in the context of bird families. Repetitions and variability were manipulated by creating a high repetitions study condition that included two bird species from a family presented six times and a high variability study condition that included six bird species from a family presented twice. Classification performance was tested for items that were studied (e.g., a Song Sparrow) along with two types of novel items, and confidence judgments were made following each classification decision. Novel items varied in their similarity to studied items in that they were either new exemplars of studied species (e.g., an unstudied Song Sparrow) or completely new species belonging to the studied families (e.g., a Chipping Sparrow). The former were assumed to be more similar to studied items because they were the same species. The three test item types were included to replicate earlier findings showing that the benefits of repetitions increase with the similarity to studied items (e.g., Dukes & Bevan, 1967), whereas the benefits of variability decrease (e.g., Homa, Sterling, & Trepel, 1981). In addition, recognition memory was tested to establish differences in the accessibility of items between study conditions and to verify differences in the similarity of studied and novel items. Of primary interest were comparisons of CLJ magnitudes with classification performance on new species from studied families in each study condition. We expected the patterns of CLJs to reveal information about the bases on which they were made.

The two bases of interest were the accessibility of studied items produced by repetitions and the variability resulting from the number of unique species in each family. Given that CLJs were sensitive to manipulations that improved classification of both studied and novel items (Jacoby et al., 2010; Wahlheim et al., 2011), it is likely that judgments are, in part, based on the accessibility of studied items. In addition, research has shown that memory for category instances can preserve their variability (e.g., Rips, 1989; Rips & Collins, 1993), suggesting that CLJs might also be sensitive to variability differences. However, given that repetition effects are more pronounced for studied than for novel items, CLJs could potentially be inflated by the increased accessibility of studied items. Furthermore, this inflation could potentially preclude sufficient incorporation of variability effects into judgments, resulting in participants underappreciating the benefits of variability on classification of novel items (i.e., variability neglect). Finding such variability neglect would suggest that accessibility is often a preferred basis for judgments and indicate that CLJs can be susceptible to illusions of competence produced by the ease of access to studied items.

Consistent with this notion, research has shown that other metacognitive judgments are susceptible to illusions produced by factors that enhance the ease of initial retrieval. For example, Benjamin, Bjork, and Schwartz (1998) showed that predictions of recall on a later test were susceptible to the fluency produced by the ease of retrieval on an initial test. In their experiments, they found that predictions of recall were higher on initial tests when retrieval from semantic memory was less difficult and for items in the recency portion of study lists. In contrast, later test performance revealed that recall was better for items that were more difficult to retrieve from semantic memory and for those that were not in the recency portion of study lists. These results potentially inform effects on CLJs by suggesting that the accessibility of repeatedly studied items may influence CLJs in ways that are similar to how the ease of initial retrieval influences predictions of later recall.

In the present experiments, we examined the sensitivity of CLJs to the effects of repetitions and variability. Finding that CLJs are insensitive to variability effects and susceptible to accessibility-based illusions would provide information about the interplay between the contributions of accessibility and variability to CLJs. More important perhaps, the interplay in reliance on these bases could have implications in applied settings. Returning to the earlier example of preparing for a cognitive psychology exam, fluent retrieval of a few repeatedly studied facts about attention theories could lead a student to believe incorrectly that he or she has mastered the topic. Consequently, the student might have an illusion of competence that prevents him or her from engaging in needed additional study. We further discuss the implications of such variability neglect in the General Discussion section.

Experiment 1

We examined the effects of repetitions and variability on the learning of natural concepts using bird family materials (Fig. 1) similar to those employed by Jacoby et al. (2010) and Wahlheim et al. (2011). As was described above, bird species (e.g., Blue Jay) were subordinate instances belonging to families that represent natural concepts (e.g., Jays). Effects of repetitions were examined by varying the number of presentations of studied species, and effects of variability were examined by varying the number of studied species included in each family. In the high repetitions condition, two species were repeated six times (S_2R_6), whereas in the high variability condition, six species were repeated twice (S_6R_2). Following study, and prior to test, participants made CLJs for each family. At the time of test, participants made: (a) recognition memory decisions, (b) classification decisions, and (c) confidence judgments about their classification decisions for each item, in that order. The test items included items that were studied earlier along with two types of novel items that varied in their similarity to studied items. The studied items were exact replicas of earlier studied birds (e.g., a Song Sparrow) and are referred to as old species, old exemplars (S_OE_O). The novel items that were similar to studied items consisted of new exemplars of studied species (e.g., an unstudied Song Sparrow), and are referred to as old species, new exemplars (S_OE_N). Finally, the novel items that were less similar to studied items consisted of species that were not studied earlier (e.g., a Chipping Sparrow) and are referred to as new species, new exemplars (S_NE_N).

We expected the proportion of old recognition memory decisions for S_OE_O items to be greater in the S_2R_6 condition than in the S_6R_2 condition because of the additional repetitions. We also expected a larger proportion of old responses for S_OE_N items than for S_NE_N items because items in the former condition were more similar to S_OE_O items. Finally, we expected the proportion of old responses for novel items to be greater in the S_6R_2 condition than in the S_2R_6 condition because studying more species in the former condition would increase the probability of their being similar to novel species, thus reducing the extent to which studied items could be differentiated from novel items. Regarding classification performance, consistent with previous research, we expected that (a) additional repetitions in the S_2R_6 condition would result in better classification of S_OE_O items than in the S_6R_2 condition, (b) differences in repetitions and variability would have off-setting effects for S_OE_N items resulting in little, if any, difference in performance between study

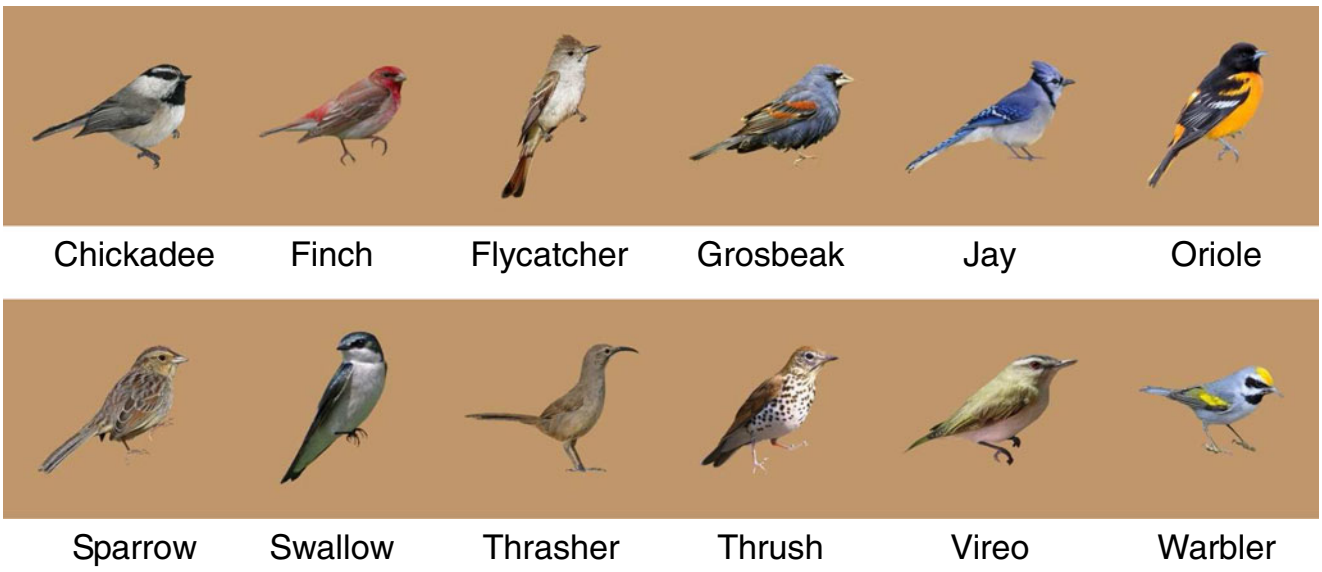

Fig. 1 Examples of species from each critical family

conditions, and (c) additional variability would facilitate classification of S_NE_N items, resulting in better performance for the S_6R_2 than the S_2R_6 condition. Finally, for metacognitive judgments, which were of primary interest, we expected that CLJs would not reflect the benefits of variability, for reasons described above, and that confidence judgments would align with performance on all item types.

Method

Participants

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

Design and materials

A 2 (study condition: high repetitions [S_2R_6] vs. high variability [S_6R_2]) × 3 (test item: old species, old exemplar [S_OE_O] vs. old species, new exemplar [S_OE_N] vs. new species, new exemplar [S_NE_N]) within-participants design was used.

Pictures of perching birds from the taxonomic order Passeriformes were chosen to represent natural concepts. The images were equally scaled and presented against a tan background. Families were selected from the same taxonomic order to provide enough between-family similarity to avoid ceiling effects. These materials were taken from a larger material set created by Wahlheim, Teune, and Jacoby (2011). The full set can be downloaded from http://psych.wustl.edu/amcclab. Critical items were chosen from the following 12 families: Chickadees, Finches, Flycatchers, Grosbeaks, Jays, Orioles, Sparrows, Swallows, Thrashers, Thrushes, Vireos, and Warblers. In addition, buffer items were chosen from the following three families: Pipits, Tanagers, and Wrens. Each of the 12 critical families included 12 species, and each of the three buffer families included three species.

Examples of the arrangement of species within families in each study and test item condition are displayed in Fig. 2. To create the study conditions, the 12 critical families were divided into two groups of six families matched on classification performance from earlier studies. Groups were then assigned to study conditions such that the S_2R_6 condition included one group of six families and the S_6R_2 condition included the other group of six families. Each group of families was assigned equally often to each study condition across participants. To create the test item conditions, the 12 species in each critical family were divided into two groups of six species matched on classification performance from earlier studies. This allowed for the separation of old species presented during study in the S_OE_O and S_OE_N conditions from new species that were only presented at test in the S_NE_N condition. To accommodate the S_OE_O and S_OE_N conditions, materials included two exemplars of each species (e.g., two different pictures of a Song Sparrow). Groups of species were assigned equally often to test item conditions across participants.

Six species were presented for study in the S_6R_2 condition, whereas only a subgroup of two species was presented for study in the S_2R_6 condition. Across participants, each subgroup of two species served equally often as studied items. At test, one subgroup of two species was presented for each family in each condition, resulting in an equal number of species being presented in the S_6R_2 and S_2R_6 conditions. For the S_OE_O condition, this resulted in only two of the six studied species in the S_6R_2 condition being presented at test and both of the two studied species in the S_2R_6 condition being presented at test. Similarly, for the S_OE_N and S_NE_N conditions, only two species were presented from each family in each study condition. Buffer items did not conform to the study condition manipulation and remained constant across experimental formats.

Procedure

Participants first completed the study phase. All stimuli were presented on a computer monitor against a black background with each item being presented individually for 8 s with its family name below. A blank screen appeared for 500 ms between presentations. Participants were told to read the family name aloud and to prepare for memory and classification tests that would include each of the three test item types described above. Three primacy buffers (one from each family) were presented first and in a different random order for each participant. Next, critical items were presented in random order in two blocks. In each block, six species from each of six families in the S_6R_2 condition were presented once each (36 presentations), and two species from each of six families in the S_2R_6 condition were presented three times each (36 presentations). Between blocks, this resulted in a total of two presentations of 36 species from the S_6R_2 condition (72 presentations) and six presentations of 12 species from the S_2R_6 condition (72 presentations). There were 144 total presentations of critical items. Finally, three recency buffers were presented randomly at the end of the list.

After the study phase, participants made their CLJs. The 12 family names were presented individually and in a different random order for each participant. Participants were instructed to predict the likelihood of correctly classifying new birds from studied families that had not been presented earlier in the experiment. Participants were told to make their judgments on a scale that ranged from 8% (guessing)

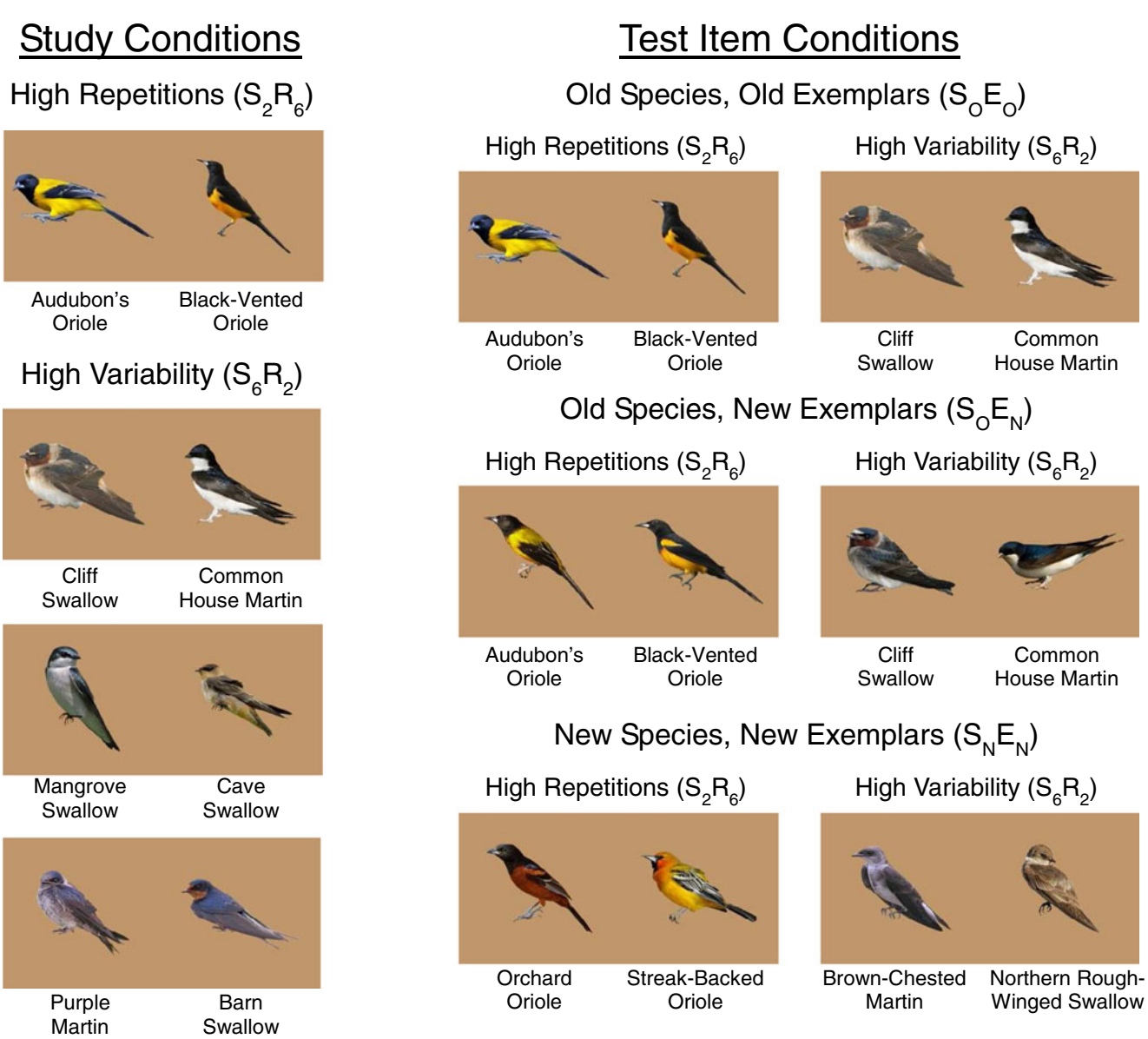

Fig. 2 Examples of species included in study and test item conditions. Species from the Oriole family are displayed for the high repetitions condition (S_2R_6), and species from the Swallow family are displayed for the high variability condition (S_6R_2)

to 100% (certain correct), and it was explained that the lower bound value on this scale was set to approximate the chance of guessing correctly when there were 12 options from which to choose. Participants moved a slider at the bottom of the screen to make their ratings and were encouraged to use the full range of the scale.

Following CLJs, participants completed recognition memory and classification tests for each item. Participants first completed a practice test that included 18 items from the buffer families. Next, participants completed the actual test, which included 72 critical items. Items were presented individually in one of 12 fixed random orders for each participant, with the restriction that no more than three items from the same condition appeared consecutively.

Recognition memory judgments were made first for each item. Items appeared above boxes labeled "old" and "new." Participants were told to click "old" for studied items (i.e., S_OE_O items) and "new" for both types of novel items (i.e., S_OE_N and S_NE_N items). Classification decisions were made following each recognition memory decision. Boxes with family names were presented below items after participants made their old/new decisions. Three family names were presented during the practice test, and 12 family names were presented during the actual test. Participants were told to click on the family to which each bird belonged regardless of whether it was old or new. Finally, participants made confidence judgments regarding their classification performance on each item. Confidence judgments in the practice

phase were made on a scale ranging from 33% to 100%, and judgments for critical items were made on the same scale as CLJs (8–100%).

Results and discussion

The significance level for statistical tests was set at alpha=.05 in all experiments.

Recognition memory

We interpret greater proportions of old responses for studied items (S_OE_O) to indicate greater accessibility and greater proportions of old responses to novel items (S_OE_N and S_NE_N) to indicate greater similarity to studied items. Table 1 shows that S_OE_O items were more accessible in the S_2R_6 condition than in the S_6R_2 condition (.83 vs. .67), $t(47) = 7.60$. In addition, S_OE_N items were more similar to studied items than were S_NE_N items (.33 vs. .22), $F(1, 47)=51.09$, $\eta_p^2 = .52$. Finally, novel items were more similar to studied items in the S_6R_2 condition than in the S_2R_6 condition (.32 vs. .24), $F(1, 47) = 10.79$, $\eta_p^2 = .19$. These results show that increasing repetitions increased recognition of studied items, new exemplars of studied species were more similar to earlier studied species than were new species, and presenting additional species from a family increased the likelihood of new species from that family being similar to earlier studied species.

Classification performance

As shown in Table 2, classification results revealed a significant interaction between study condition and type of test item, $F(2, 94) = 31.01$, $\eta_p^2 = .40$. Performance on S_OE_O items was higher for the S_2R_6 condition than the S_6R_2 condition (.72 vs. .57), $t(47) = 6.12$, there was no difference between the S_2R_6 and S_6R_2 conditions for S_OE_N items (.46 vs. .46), $t < 1$, and performance on S_NE_N items was higher for the S_6R_2 condition than for the S_2R_6 condition (.42 vs. .31), $t(47) = 4.49$. These results replicate earlier findings, with artificial materials showing that repetition benefits increased with the similarity between study and test items, whereas the reverse was true for the benefits of variability (e.g., Dukes & Bevan, 1967; Homa et al., 1981). These results also extend those earlier findings to the domain of natural concepts.

Metacognitive judgments

Our primary interest was in the sensitivity of CLJs to the effects of repetitions and variability on classification of S_NE_N items. We were also interested in confidence judgments made for all test items. We examined the magnitudes of judgments and their correspondence to actual performance by means of an ANOVA.

Category learning judgments

Figure 3 shows that participants were overconfident in their CLJs as compared with their overall performance (.55 vs. .36), $F(1, 47) = 47.51$, $\eta_p^2 = .50$, and this overconfidence was greater for the S_2R_6 condition than for the S_6R_2

Table 1 Proportion of "old" recognition memory responses as a function of study condition and test item: Experiments 1–4

Study Condition	Test Item		
	S_OE_O	S_OE_N	S_NE_N
	M SEM	*M SEM*	*M SEM*
Experiment 1			
S_2R_6	.83 .02	.30 .03	.18 .02
S_6R_2	.67 .03	.37 .03	.26 .03
Experiment 2			
S_2R_6	.91 .02	.26 .03	.16 .03
S_6R_2	.61 .05	.33 .04	.24 .04
Experiment 3			
S_2R_6	.82 .03	.30 .03	.19 .02
S_6R_2	.67 .03	.35 .03	.26 .03
Experiment 4			
S_2R_6	.83 .04	.30 .04	.15 .03
S_6R_2	.65 .04	.34 .04	.22 .03

Table 2 Classification performance and confidence judgments as a function of study condition and test item: Experiments 1–4

Study Condition	Classification			Confidence		
	S_OE_O	S_OE_N	S_NE_N	S_OE_O	S_OE_N	S_NE_N
	M SEM	*M SEM*	*M SEM*	*M SEM*	*M SEM*	*M SEM*
Experiment 1						
S_2R_6	.72 .03	.46 .03	.31 .02	.73 .02	.52 .02	.48 .02
S_6R_2	.57 .03	.46 .03	.42 .02	.59 .02	.53 .02	.49 .02
Experiment 2						
S_2R_6	.78 .04	.49 .04	.32 .03	.77 .03	.54 .03	.45 .02
S_6R_2	.59 .05	.51 .05	.41 .05	.61 .04	.53 .03	.50 .02
Experiment 3						
S_2R_6	.72 .03	.48 .03	.32 .02	.74 .02	.52 .02	.46 .02
S_6R_2	.58 .03	.50 .03	.43 .03	.60 .02	.53 .02	.46 .02
Experiment 4						
S_2R_6	.73 .04	.48 .05	.25 .02	.71 .04	.51 .03	.42 .02
S_6R_2	.56 .05	.48 .04	.41 .03	.56 .03	.49 .03	.47 .02

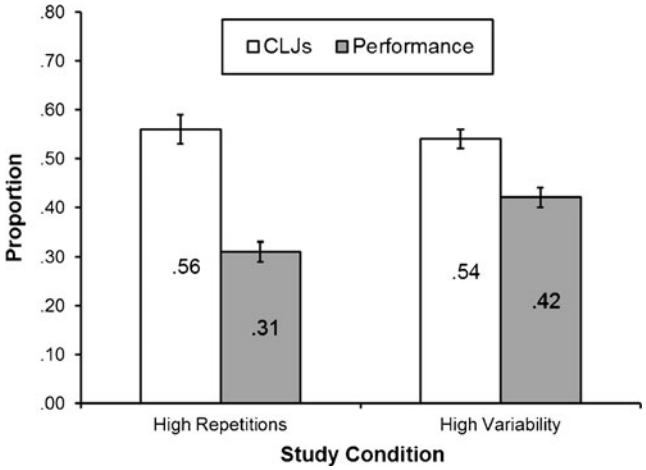

Fig. 3 Mean CLJs and classification performance on S_NE_N exemplars as a function of study condition: Experiment 1. High Repetitions = S_2R_6, High Variability = S_6R_2. *Error bars* represent standard errors of the means

condition, $F(1, 47) = 21.20$, $\eta_p^2 = .31$. The difference in overconfidence was driven by performance differences. Classification performance was greater in the S_6R_2 condition than in the S_2R_6 condition (.42 vs. .31), $t(47) = 4.49$, whereas CLJs did not differ between the S_2R_6 and S_6R_2 conditions (.56 vs. .54), $t(47) = 1.57$. These results provide evidence that participants did not adequately incorporate the benefits of variability into their CLJs.

Confidence judgments

Table 2 shows that confidence aligned well with classification of S_OE_O items, but judgments were significantly greater than performance on novel items (S_OE_N and S_NE_N). In addition, the difference between confidence and performance was greatest for S_NE_N items in the S_2R_6 condition, $F(2, 94) = 10.80$, $\eta_p^2 = .19$. The finding of greatest overconfidence in this condition is consistent with the pattern of overconfidence found for CLJs. These results suggest that both types of judgment neglected the benefits of variability.

Experiment 2

The lack of a difference between CLJs in each study condition in Experiment 1 suggests that participants were not aware of the benefits of variability on the classification of novel items. However, the bases on which CLJs were made is still unclear. One possibility is that CLJs reflected non-diagnostic bases that resulted in participants essentially guessing when making their judgments. For example, such guessing could have been made on the basis of the number of overall presentations from each family, which did not differ between study conditions. Alternatively, it is possible that participants did not appreciate the benefits of variability in their CLJs because the enhanced memory for studied items produced by additional repetitions precluded their doing so. In Experiment 2, we further explored participants' preferred bases for predictions by including an additional measure prior to CLJs.

After study, and prior to the CLJ phase, we presented six pairs of family names that each included one family from the S_2R_6 condition and one family from the S_6R_2 condition. We instructed participants to choose which would produce better classification of new species from studied families (S_NE_N items). Use of this paired-comparison measure eliminated the possibility of reliance on the number of presentations of family names because family names were presented equally often in each study condition. Also, by contrasting families from each study condition, we expected to increase participants' awareness of differences in the accessibility and variability of exemplars between conditions. Heightening participants' awareness of differences on these dimensions immediately prior to CLJs could have the effect of increasing participants' reliance on the more accessible basis, similar to previous studies (e.g., Benjamin et al., 1998). Given that additional repetitions should produce more accessible representations of studied items, we expected CLJs to be greater in the high repetitions condition (S_2R_6). In addition, clear evidence for overreliance on accessibility along with variability neglect would be indicated by a greater percentage of high repetitions families being chosen on the paired-comparison measure. Finally, in the interest of further understanding participants' metacognitive awareness of repetition and variability effects, we also included a post-test questionnaire at the end of the experiment to examine the extent to which participants were aware of these effects on a more theoretical level.

Method

Participants

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

Design, materials, and procedure

The design, materials, and procedure were identical to those of Experiment 1, with the following exceptions. Following the study phase, and prior to the CLJ phase, pairs of family names, including one family from each study condition, were presented individually and in random order. Families in each pair were matched as closely as possible on overall classification performance from Experiment 1. Participants

were instructed to click on the family that they thought would produce better performance on S_NE_N items.

After the test phase, participants were given a questionnaire that described the difference between the number of repetitions and species in each study condition. Participants were first asked which condition they thought produced better performance on S_NE_N items. They were then asked the same question about S_OE_O items. They were given the option to choose either of the study conditions or to indicate that neither study condition produced an advantage.

Results and discussion

Recognition Memory

Recognition memory results (Table 1) replicated Experiment 1 in showing that S_OE_O items were better recognized in the S_2R_6 condition than in the S_6R_2 condition (.91 vs. .61), $t(23) = 6.83$; S_OE_N items were more similar to studied items than were S_NE_N items (.29 vs. .20), $F(1, 23) = 17.49$, $\eta_p^2 = .43$; and both types of novel items were more similar to studied items in the S_6R_2 than S_2R_6 condition (.28 vs. .21), $F(1, 23) = 5.79$, $\eta_p^2 = .20$.

Classification performance

Classification performance (Table 2) differed across study conditions and test item types in the same manner as in Experiment 1, $F(2, 46) = 18.63$, $\eta_p^2 = .45$. Classification of S_OE_O items was better for the S_2R_6 condition than for the S_6R_2 condition (.78 vs. .59), $t(23) = 4.52$; there was no difference between S_2R_6 and S_6R_2 conditions for S_OE_N items (.49 vs. .51), $t < 1$; and S_NE_N items were classified better in the S_6R_2 condition than in the S_2R_6 condition (.41 vs. .32), $t(23) = 2.61$. These results again show that repetition benefits increased with the similarity between study and test items, whereas the benefits of variability decreased.

Paired comparisons

A nonsignificant trend showing that S_2R_6 families were chosen more often than S_6R_2 families (53% vs. 47%) indicated a lack of sensitivity to variability effects and a potential inappropriate reliance on accessibility.

Category learning judgments

Figure 4 shows that participants were more overconfident in their CLJs in the S_2R_6 condition than in the S_6R_2 condition, $F(1, 23) = 20.29$, $\eta_p^2 = .47$. However, the nature of the interaction was different than in Experiment 1. As in Experiment 1, performance was significantly higher for the S_6R_2 condition than for the S_2R_6 condition (.41 vs. .32), $t(23) = 2.61$. In contrast, CLJs were significantly higher for the S_2R_6 than S_6R_2 condition (.62 vs. .56), $t(23) = 2.79$. These results suggest that contrasting the two study conditions prior to CLJs highlighted differences in the accessibility of exemplars and increased reliance on that basis. This pattern of CLJs provides evidence of an overreliance on the accessibility of exemplars that precluded the incorporation of variability effects into judgments.

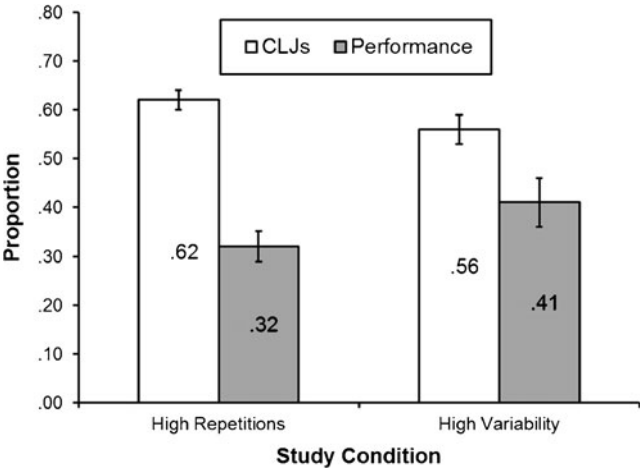

Fig. 4 Mean CLJs and classification performance on S_NE_N exemplars as a function of study condition: Experiment 2. High Repetitions = S_2R_6, High Variability = S_6R_2. *Error bars* represent standard errors of the means

Confidence judgments

Confidence judgments (Table 2) did not differ from performance, with the exception that judgments for S_NE_N items in the S_2R_6 condition were significantly greater than performance (.45 vs. .32), $t(23) = 4.65$. These results are again consistent with CLJs and potentially indicate an overreliance on accessibility as a basis for judgments.

Post-test questionnaires

In contrast with CLJs, Table 3 shows that most participants were aware of the effects of repetitions and variability on S_OE_O and S_NE_N items on a more theoretical level. The majority of participants indicated that the S_6R_2 condition produced better classification on S_NE_N items (67%) and that the S_2R_6 condition produced better classification on S_OE_O items (83%). The results suggest that even though most participants were aware of these effects, their CLJs were still primarily based on accessibility.

Experiment 3

The finding that a large majority of participants indicated awareness of variability effects on post-test questionnaires

Table 3 Response percentages on questionnaires: Experiments 2 and 3

	Response		
Experiment	High Variability	High Repetitions	No Difference
Experiment 2 (Post test)			
Better classification of novel species?	67%	25%	8%
Better classification of studied species?	13%	83%	4%
Experiment 3 (Post study)			
Better classification of novel species?	67%	21%	12%
Better classification of studied species?	15%	79%	6%

High Variability = S_6R_2; High Repetitions = S_2R_6

in Experiment 2 leaves open the possibility that participants may not have been aware of variability effects at the time of CLJs, but became aware of the effects as a result of the testing experience. In Experiment 3, we examined this possibility by administering a questionnaire immediately following the study phase. Doing so allowed us to determine whether participants were sensitive to variability effects prior to making their CLJs, and if so, whether those who showed such sensitivity still based their CLJs primarily on accessibility. In addition, presenting a questionnaire that contrasts differences in the variability and accessibility of studied items could have effects similar to those of presenting the paired comparisons in Experiment 2. Finding greater CLJs for high repetition families for participants who show awareness of variability effects on a theoretical level would provide strong evidence that the CLJ measure is susceptible to accessibility-based metacognitive illusions.

Method

Participants

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

Design, materials, and procedure

The design, materials, and procedure were identical to those in Experiment 1, with the following exception. Following the study phase, and prior to the CLJ phase, participants were given a questionnaire similar to that administered in Experiment 2. The questionnaire described differences between the study conditions, and participants were asked to make predictions of performance for each condition. They indicated whether one of the study conditions would produce better performance or whether there would be no difference between study conditions for S_NE_N items and then for S_OE_O items.

Results and discussion

Recognition memory

As in Experiments 1 and 2, Table 1 shows that S_OE_O items were better recognized in the S_2R_6 condition than in the S_6R_2 condition (.82 vs. .67), $t(47) = 5.64$; S_OE_N items were more similar to studied items than were S_NE_N items (.33 vs. .23), $F(1, 47) = 28.54$, $\eta_p^2 = .38$; and novel items were more similar to studied items in the S_6R_2 condition than in the S_2R_6 condition (.31 vs. .25), $F(1, 47) = 9.52$, $\eta_p^2 = .17$.

Classification performance

Classification performance (Table 2) differed across study conditions and test item types in the same manner as in Experiments 1 and 2, $F(2, 94) = 34.46$, $\eta_p^2 = .42$, showing that repetition benefits increased with the similarity between study and test items, whereas variability benefits decreased. Classification of S_OE_O items was better for the S_2R_6 condition than for the S_6R_2 condition (.72 vs. .58), $t(47) = 5.51$; there was no difference between S_2R_6 and S_6R_2 conditions for S_OE_N items (.48 vs. .50), $t < 1$; and S_NE_N items were classified better for the S_6R_2 condition than for the S_2R_6 condition (.43 vs. .32), $t(47) = 3.78$.

Post-study questionnaires

Table 3 shows that most participants were aware of the effects of repetitions and variability prior to making their CLJs. The majority of participants predicted that the S_6R_2 condition would produce better classification on S_NE_N items (67%) and that the S_2R_6 condition would produce better classification on S_OE_O items (79%). In fact, the overall pattern of responses was a near perfect replication of that observed on post-test questionnaires in Experiment 2.

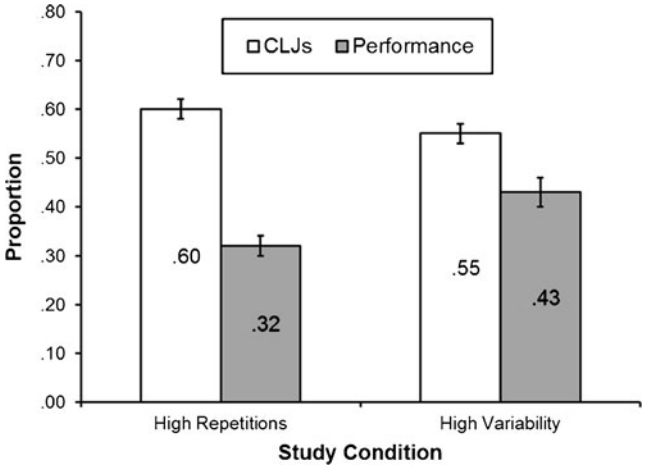

Fig. 5 Mean CLJs and classification performance on S_NE_N exemplars as a function of study condition: Experiment 3. High Repetitions = S_2R_6, High Variability = S_6R_2. *Error bars* represent standard errors of the means

Category learning judgments

Consistent with earlier experiments, Fig. 5 shows that participants' CLJs were again more overconfident for the S_2R_6 condition than for the S_6R_2 condition, $F(1, 47) = 20.10$, $\eta_p^2 = .30$. As in Experiment 2, performance was significantly higher for the S_6R_2 condition than for the S_2R_6 condition (.43 vs. .32), $t(47) = 3.78$, whereas CLJs were significantly higher for the S_2R_6 condition than for the S_6R_2 condition (.60 vs. .55), $t(47) = 2.06$. In addition, this pattern for CLJs was found for each type of questionnaire response regarding S_NE_N items. Conditional analyses of CLJs revealed no study condition × questionnaire response interaction, $F < 1$. These results show that even when participants indicated awareness of the benefits of variability immediately prior to CLJs, their judgments were still biased by the increased accessibility of studied items resulting from additional repetitions.

Confidence judgments

Confidence judgments (Table 2) did not differ from overall performance, except that judgments for S_NE_N items in the S_2R_6 condition were significantly higher than performance (.46 vs. .32), $t(47) = 6.46$. These results again suggest that confidence judgments may have also been primarily based on the accessibility of studied items.

Experiment 4

Together, results from the previous experiments show that participants neglect the benefits of variability in their CLJs and overrely on accessibility as a basis for their judgments. One possibility is that this overreliance on accessibility occurs when participants cannot remember the variability information for a family. In Experiment 4, we explored this possibility by examining CLJs as a function of whether participants could remember variability information about families. Finding that CLJs are greater for high repetition families when variability information cannot be remembered would not be surprising because the most salient basis available for CLJs would be the accessibility of studied items. Results such as these would support the earlier conclusion that greater CLJs in high repetitions families reflect reliance on accessibility as a basis for judgments. Of equal importance, finding that CLJs remain insensitive to variability effects, even when variability information can be remembered, would provide evidence for variability neglect that potentially results from an inappropriate reliance on accessibility.

Method

Participants

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

Design, materials, and procedure

The design, materials, and procedure were identical to those in Experiment 1, with the following exception. During the CLJ phase, the numbers 2 and 6 appeared in boxes below each family name, prior to the CLJ query. Participants were told to click on the number that indicated how many unique species of that family had been presented during the study phase.

Results and discussion

Recognition memory

As in Experiments 1–3, Table 1 shows that S_OE_O items were better recognized in the S_2R_6 condition than in the S_6R_2 condition (.83 vs. .65), $t(23) = 4.32$; S_OE_N items were more similar to studied items than were S_NE_N items (.32 vs. .18), $F(1, 23) = 32.14$, $\eta_p^2 = .58$; and novel items were marginally more similar to studied items in the S_6R_2 condition than in the S_2R_6 condition (.28 vs. .22), $F(1, 23) = 4.08$, $p = .055$, $\eta_p^2 = .15$.

Classification performance

Classification performance (Table 2) differed across study conditions and test item types in the same manner as in Experiments 1–3, $F(2, 46) = 19.16$, $\eta_p^2 = .45$. Performance on S_OE_O items was better for the S_2R_6 condition than for the S_6R_2 condition (.73 vs. .56), $t(23) = 3.11$; performance on S_OE_N items did not differ between the S_2R_6 and S_6R_2 conditions (.48 vs. .48), $t < 1$; and performance on S_NE_N items was better for the S_6R_2 condition than for the S_2R_6 condition (.41 vs. .25), $t(23) = 4.18$. These results again show that repetition benefits increased with the similarity between study and test items, whereas variability benefits decreased.

Category learning judgments

Figure 6 shows that participants' CLJs were again overconfident as compared with actual performance (.47 vs. .33), $F(1, 23) = 18.93$, $\eta_p^2 = .45$, and this overconfidence was greater for the S_2R_6 condition than for the S_6R_2 condition, $F(1, 23) = 20.29$, $\eta_p^2 = .47$. Performance on S_NE_N items was higher for the S_6R_2 condition than for the S_2R_6 condition, (.41 vs. .25), $t(23) = 4.18$, and CLJs revealed a non-significant advantage for the S_2R_6 over the S_6R_2 condition (.49 vs. .45), $t(23) = 1.22$. These results show a pattern consistent with those observed in earlier experiments.

To examine the sensitivity of CLJs to variability effects as a function of awareness of variability differences, we examined CLJs conditionalized on the accuracy of assessments of family size (i.e., size judgments). Analyses were conducted for 16 participants who had at least one observation in each cell. The proportion of correct size judgments did not differ between study conditions for all participants (.67 vs. .65), nor did it differ for the 16 participants included in the conditional analyses (.61 vs. .57), $ts < 1$. Figure 7 shows a significant interaction, $F(1, 15) = 8.97$, $\eta_p^2 = .37$, revealing that for correct size judgments, CLJs did not differ between the S_2R_6 and S_6R_2 conditions (.51 vs. .49), $t < 1$, whereas for incorrect size judgments, CLJs were significantly higher for the S_2R_6 condition than for the S_6R_2 condition (.46 vs. .33), $t(15) = 2.37$. These results show that even when participants were aware of variability differences, there was still a neglect of the benefits of variability. In addition, when participants were unaware of variability differences, CLJs relied heavily on the accessibility of studied items. Finally, the combination of these results suggests that variability information was incorporated to some extent when participants were aware of variability differences, but the powerful bias toward accessibility increased judgments for high repetition families to levels similar to those for high variability families. Thus, accessibility was the preferred basis for judgments, and this preference precluded adequate incorporation of variability information into participants' judgments.

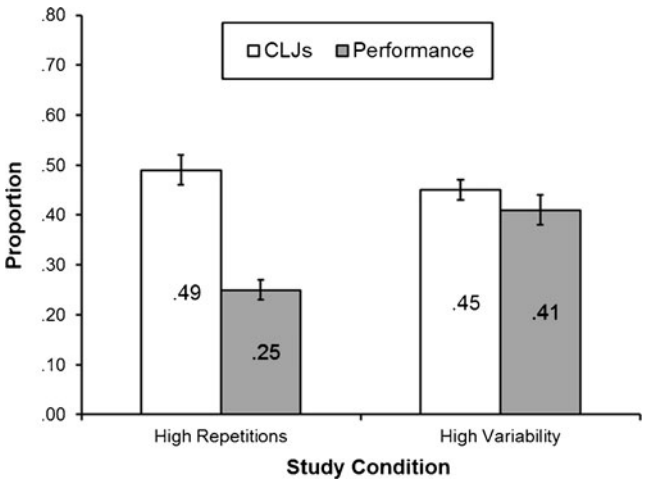

Fig. 6 Mean CLJs and classification performance on S_NE_N exemplars as a function of study condition: Experiment 4. High Repetitions = S_2R_6, High Variability = S_6R_2. *Error bars* represent standard errors of the means

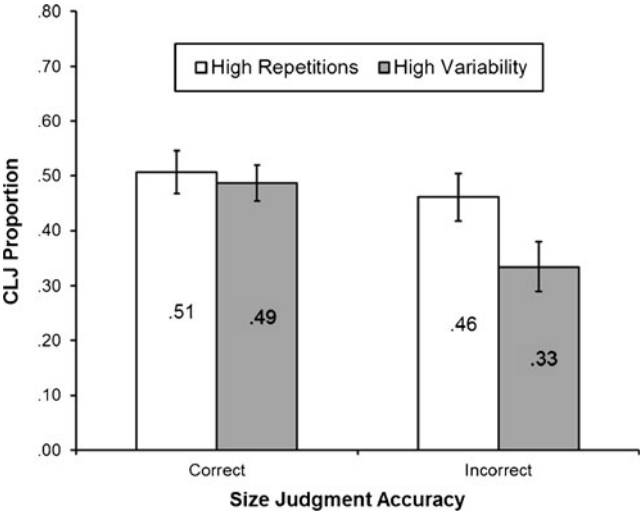

Fig. 7 Mean CLJ proportion as a function of study condition and size judgment accuracy: Experiment 4. High Repetitions = S_2R_6, High Variability = S_6R_2. *Error bars* represent standard errors of the means

Confidence judgments

Confidence judgments (Table 2) did not differ from overall performance, with the exception that judgments for S_NE_N items in the S_2R_6 condition were significantly greater than performance (.42 vs. .25), $t(23) = 5.55$. These results again suggest that confidence judgments and CLJs were made on similar bases.

General discussion

Results from the present experiments showed that in the context of natural concept learning, variability improved

classification of novel items, whereas repetition improved classification of studied items. These findings replicate and extend earlier studies in which artificial materials were used to represent concepts (e.g., Dukes & Bevan, 1967; Homa et al., 1981). A point worth noting is that differences in variability are often confounded with differences in the similarity between studied and tested items, and effects such as those obtained in the present experiments could be explained by either mechanism (e.g., Hahn, Bailey, & Elvin, 2005; Stewart & Chater, 2002). Although the recognition memory results in the present experiments provide reason to suspect that similarity played a role in producing the differences between study conditions, a full analysis of the contribution of variability and similarity mechanisms would require modeling efforts that are well outside the scope of the present article. Given that this is the first study to make use of this particular set of materials, a potential goal for future research could be to examine the similarity space of this set. Another point worth noting is that all dependent measures in the present experiments were included within participants, which may have created problems for interpretation because of the potential for carry-over effects. For recognition and classification, the orderliness of the results suggests that this was not a concern. In contrast, the inconsistency in the patterns of CLJs across experiments provides clear evidence that measures occurring prior to CLJs influenced participants' judgments.

The inconsistency in CLJs across experiments due to the measures that preceded them revealed additional information regarding bases for judgments. CLJs did not differ between study conditions in Experiments 1 and 4, whereas CLJs were greater for the high repetitions condition in Experiments 2 and 3. The lack of difference in CLJs in Experiments 1 and 4 is likely the result of accessibility and variability both contributing as bases for CLJs, with their contributions off-setting one another. In contrast, CLJs were greater in the high repetitions condition in Experiments 2 and 3 because the task that occurred in the phase prior to CLJs presumably heightened participants' attention to differences in the representations formed in each study condition, and shifted participants toward the most salient basis for judgments. Although the exact nature of the representational differences cannot be determined by the present data, recognition memory results indicate that studied items were better differentiated from novel items following high repetitions than high variability (cf. Shiffrin, Huber, & Marinelli, 1995). Consequently, directing participants' attention to representational differences may have caused them to notice that birds from some families were easier to remember. Given that the ease of retrieval is often used as a basis for metacognitive judgments (e.g., Benjamin et al., 1998), noticing that studied birds could be more easily retrieved in the high repetition families may have resulted in participants incorrectly giving those families greater CLJs. Thus, although many participants indicated awareness of variability effects, they were sufficiently captured by the accessibility of studied items when making judgments at the category level.

Similar to CLJs, confidence judgments for novel items including new species also showed insensitivity to variability effects. However, note that carry-over effects were not an issue because there was not a systematic difference in the pattern of confidence across experiments. Classification decisions about novel items are often made on the basis of comparisons of their similarity with representations formed during study (see Murphy, 2002). When assessing confidence, it is likely that participants based their judgments on the ease with which these comparisons were made, with greater ease leading to higher confidence. Reliance on the ease of comparisons can explain the overconfidence on novel items including new species for high repetition families in that comparisons made easier by repetitions produced inflated confidence for correct responses. In addition, the lack of differences in overall levels of confidence for these novel items in each study condition can be explained by off-setting effects of a greater proportion of correct responses for high variability families and greater confidence in the lower proportion of correct responses from high repetition families. Thus, confidence judgments and CLJs were both susceptible to accessibility-based illusions created by repetitions, albeit in different ways.

In contrast with CLJs and confidence judgments, responses to questionnaires revealed that participants were sensitive to variability effects when assessments were queried at a theoretical level. The inconsistency between responses made on questionnaires and judgments made at the category and item levels can be explained by considering differences in the bases for these judgments. The questionnaires were designed to tap into participants' theories about variability and repetition effects in a situation that was relatively decontextualized from processing differences produced by the arrangement of studied items in each study condition. In contrast, CLJs and confidence judgments were heavily contextualized within the situation, rendering processing differences a more salient basis for judgments. Consequently, when responding to questionnaires, participants were able to consider other variables that were important for assessing the effects of variability on natural concept learning. In contrast, participants were sufficiently captured by differences in accessibility at the category and item levels, resulting in overreliance on that basis for judgments.

Other researchers have also shown that consideration of differences in the extent to which metacognitive judgments are contextualized within a task is important for understanding the bases on which they are made. For example, Dunlosky and Hertzog (2000) found that global assessments of memory performance for items learned under imagery and repetition

conditions more accurately aligned with performance in a memory task than did assessments made for individual items. They concluded that global assessments allowed participants to make judgments on the basis of declarative knowledge about the effects of each study condition, whereas judgments made at the item level were based primarily on processing differences. Similarly, Kelley and Jacoby (1996) made the distinction between subjective experience and theories as bases for predictions of others' abilities to solve anagrams. They showed that presenting the solutions to anagrams in a phase prior to the presentation of anagrams increased the subjective ease with which anagrams could be solved. This resulted in inflated predictions of others' ability to solve those anagrams indicating that judgments were based on participants' subjective experience. However, when participants were made aware that prior presentation of a solution made anagram completion more fluent, their predictions for others became more accurate because they were able to discount their own subjective experience and rely on a more theoretical basis for predictions. Finally, Koriat, Bjork, Sheffer, and Bar (2004) described a similar distinction between experience-based and theory-based judgments in people's predictions of forgetting. They found that item-level judgments referring to one's own performance did not accurately predict rates of forgetting, presumably because they were based on participants' subjective experience. However, when a theoretical basis for judgments was elicited from participants who did not complete the memory task by asking them to make aggregate assessments about the performance of those who did complete the task, the assessments of the former group aligned well with actual rates of forgetting shown by the latter group. The results from these studies are consistent with the findings in the present experiments showing that judgments based on processing differences can sometimes mislead assessments of performance because they do not account for the influence of other variables.

Aside from the theoretical implications described above, the finding of variability neglect in CLJs in the present experiments also has important applied implications. Variability neglect occurs when people fail to appreciate the need to incorporate a sufficient variety of instances into their conceptual representation for producing a complete understanding of a concept. In a classroom setting, this could result from the belief that repeated exposure to a small number of examples is the best way to learn a concept, which is fostered by students' strategy of memorizing facts to perform well on a test. Such metacognitive errors could potentially have negative consequences for future performance both on exams and outside of the classroom if the situations requiring such knowledge have sufficiently different contexts. Research suggests that educators may play a critical role in students' understanding and appreciation of the benefits of variability. For example, students' appreciation of diverse samples for inductive reasoning has been shown to increase with age in young children (e.g., Rhodes, Gelman, & Brickman, 2008). Findings such as these highlight the importance of educating metacognition in the classroom to improve students' understanding and application of effective study strategies.

Author note The present research was supported by National Institute on Aging Grant 5T32AG000030 and by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind, and Behavior Collaborative Award. We thank Rachel Teune for her assistance with manuscript preparation and data collection. We also thank Sarah Arnspiger, Ashley Bartels, Lauren Guenther, Synneva Hagen-Lillevik, Dan Howard, and Ashim Lamichhane for their assistance with data collection.

References

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology. General, 127,* 55–68.

Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

Dukes, W. F., & Bevan, W. (1967). Stimulus variation and repetition in the acquisition of naming responses. *Journal of Experimental Psychology, 74,* 178–181.

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about strategy effectiveness: A componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging, 15,* 462–474.

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

Hahn, U., Bailey, T. M., & Elvin, L. B. (2005). Effects of category diversity on learning, memory, and generalization. *Memory and Cognition, 33,* 289–302.

Homa, D., Cross, J., Cornell, D., Goldman, D., & Shwartz, S. (1973). Prototype abstraction and classification of new instances as a function of number of instances defining the prototype. *Journal of Experimental Psychology, 101,* 116–122.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 418–439.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1441–1451.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35,* 157–175.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133,* 643–656.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19,* 585–592.

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

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

Rhodes, M., Gelman, S. A., & Brickman, D. (2008). Developmental changes in the consideration of sample diversity in inductive reasoning. *Journal of Cognition and Development, 9,* 112–143.

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.

Rips, L., & Collins, A. (1993). Categories and resemblance. *Journal of Experimental Psychology: General, 122,* 468–486.

Shah, A. K., & Oppenheimer, D. M. (2011). Grouping information for judgments. *Journal of Experimental Psychology: General, 140,* 1–13.

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 267–287.

Stewart, N., & Chater, N. (2002). The effect of category variability in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 893–907.

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory and Cognition, 39,* 750–763.

Wahlheim, C. N., Teune, R. K., & Jacoby, L. L. (2011). Birds as natural concepts: A set of pictures from the Passeriformes order. Retrieved from http://psych.wustl.edu/amcclab/AMCC%20Materials.htm