Memory & Cognition
2006, 34 (1), 17-27

Formation of category representations

A. J. WILLS, MALIA NOURY, and NICHOLAS J. MOBERLY
University of Exeter, Exeter, England

and

MATTHEW NEWPORT
Cambridge University, Cambridge, England

Many formal models of categorization assume, implicitly or explicitly, that categorization results in the formation of direct associations from representations of the presented stimuli to representations of the experimentally provided category labels. In three categorization experiments employing a polymorphous classification structure (Dennis, Hampton, & Lea, 1973) and a partial reversal, *optional shift* procedure (Kendler, Kendler, & Wells, 1960), we provide evidence consistent with the hypothesis that learning a new classification problem results in the creation of category representations that mediate between representations of the stimulus and the label. This hypothesis can be instantiated through the AMBRY model (Kruschke, 1996).

Formal models of categorization (e.g., Ashby & Gott, 1988; Gluck & Bower, 1988; Kruschke, 1992; Nosofsky, 1986; Nosofsky & Palmeri, 1997; Wills, Reimers, Stewart, Suret, & McLaren, 2000) differ in many aspects of their design, including stimulus representation, attentional mechanisms, and decision processes. Among all this variety, there is a remarkable region of agreement. All of the above models, and many more besides, appear to assume that the stimulus representations (whatever they may be) are directly linked to a representation of the appropriate category label. For example, in the ALCOVE model (Kruschke, 1992), the hidden layer of exemplar nodes connects directly to a layer of category output nodes. The activity of the category output nodes determines, via the ratio rule (Luce, 1959), the probability of a particular response. No level of representation mediates between the exemplars and the category label. The purpose of this article is to investigate whether at least one level of mediating representations is needed.

We start by briefly considering some of the evidence that has been taken to be in support of category representations, and show that it is open to alternative interpretations. We then consider one class of evidence that does seem to support the presence of category representations and describe one way in which formal categorization models could accommodate this evidence. We then derive further predictions from this account (and an alternative), and test these predictions across three experiments.

Delamater and Joseph (2000) trained undergraduates on a conditional simultaneous discrimination problem. Two stimuli appeared on the screen at the same time, and the participants had to choose one of them. Stimuli occurring immediately beforehand determined the correct choice on any given trial. Each choice stimulus had two cues (e.g., red and blue both signaled "choose black," and yellow and green both signaled "choose white"). Once this task had been mastered, the participants were transferred to a task in which the choice stimuli were changed to novel stimuli (vertical and horizontal lines). In the consistent transfer condition, cues that signaled the same outcome continued to do so. In the inconsistent transfer condition, cues that signaled the same outcome now signaled different outcomes. The participants in the consistent transfer condition learned more rapidly than those in the inconsistent transfer condition.

One explanation of such *common consequence* effects (Urcuioli, Zentall, Jackson-Smith, & Steirn, 1989) is that the organism develops a *common code* (a category representation) for cues that signal the same choice. An alternative explanation is the associatively driven activation of stimulus representations (Delamater & Joseph, 2000; Hall, 1996). In the first phase of the experiment, associations form between each of the cues and the choice stimulus that they signal should be chosen (red, blue → black; yellow, green → white). Once these associations have been formed, presentation of a cue stimulus will now associatively activate the representation of the appropriate choice stimulus. In the transfer phase, this associatively activated choice stimulus becomes, in turn, associated to the new choice stimulus (e.g., red, blue → black → vertical lines). Recent evidence supports the associative account (Hall, Mitchell, Graham, & Lavis, 2003). Associative activa-

This research was supported by U.K. BBSRC Grant 9/S17109 and by European Commission Framework 6 Grant 516542 (NEST) to the first author. The authors thank Ian Chung and Katharine Haigh for their assistance in collecting data for Experiments 1A and 1B, respectively, and Aureliu Lavric, Stephen Lea, and Ian McLaren for their helpful comments. Correspondence concerning this article should be addressed to A. J. Wills, University of Exeter, Washington Singer Laboratories, Perry Road, Exeter EX4 4QG, England (e-mail: a.j.wills@exeter.ac.uk), and related research can be found at www.willslab.co.uk.

tion of stimulus representations can also explain much of the evidence for *stimulus equivalence* (Sidman & Tailby, 1982), such as the development of symmetry (Murdock, 1956) and transitivity (Norcross & Spiker, 1958) relationships without explicit training.

Turning to other evidence, Buss (1953) taught participants to make one of two responses to presented stimuli, dependent on stimulus height. The participants found it easier to learn a subsequent reversal of the discrimination (reversal shift, RS) than to transfer to a problem in which the shape of the stimuli was the relevant dimension and height was irrelevant (extradimensional shift, EDS). The Kendlers (e.g., Kendler & Kendler, 1968) argued that the RS–EDS difference was caused by the development of representations that mediated between stimulus and label representations. However, it can also be explained by selective attention (Sutherland & Mackintosh, 1971); the initial discrimination increases attention to the stimulus dimension of height and/or decreases attention to the other stimulus dimensions. This facilitates learning of the reversal but is not helpful in learning a discrimination based on a previously irrelevant dimension.

Perhaps the clearest evidence for category representations has come from the full versus partial reversal difference. Sanders's (1971) study provides one demonstration (see also Kruschke, 1996, among others). Sanders trained second-grade children that, for example, a green T and a black outline square were "winner" cards, whereas a yellow T and a black outline triangle were "loser" cards. In the full reversal condition, the winners and losers were then swapped. In the partial reversal condition, one of the stimulus pairs was reversed, whereas the other remained the same. Second-grade children found the full reversal easier to acquire than the partial reversal. These results are consistent with the idea that the children developed category representations in response to the task. It was these category representations, rather than the individual stimulus representations, that were associated to the labels "winner" and "loser." Unlike the common consequence results discussed earlier, these results seem to require representations not specifically provided by the stimulus environment. Unlike the RS–EDS results, an alternative explanation in terms of selective attention to preexisting stimulus dimensions seems unlikely.[1]

We said earlier that many formal models of categorization do not include a category representation layer. One exception to this is the AMBRY connectionist model (Kruschke, 1996), which includes a set of category nodes that mediate between exemplar-like stimulus representation nodes and category-label nodes. For brevity, we henceforth will refer to category representations that mediate between stimulus and category-label representations as an *ARMA* component.[2]

Assuming the presence of an ARMA component in the categorization process permits a striking, and perhaps counterintuitive, prediction. If one trains a categorization problem and then trains a reversal on a subset of the original stimuli, that reversal might transfer to the remainder of the stimuli without further training. This is because the reversal of the subset can be learned by reversing the category → label associations, leaving the stimulus → category associations relatively intact. When the remaining stimuli are presented, they activate the same category representation that they did previously, but that representation is now linked to the opposite label and, hence, participants make the opposite response. A related effect has been well documented within the RS–EDS paradigm (e.g., Kendler et al., 1960), where it is known as *optional shift* behavior. However, like the RS–EDS effect itself, optional shift behavior can be explained through selective attention.

There appears to be only one demonstration of optional shift in the human literature to which selective attention seems unlikely to have contributed. Wirth and Chase (2002) trained different vocal responses to different sets of three arbitrary, abstract symbols. In a subsequent phase, the appropriate response for some stimuli was reversed, and then responding to the other stimuli in the set was tested in the absence of feedback. Five of 6 participants showed optional shift behavior. No group-level statistics were performed. The optional shift behavior reported seems unlikely to have resulted from selective attention, because there was no obvious dimension that the members of these arbitrary sets shared. Nevertheless, a critic could argue that the dimensional structure of their arbitrary sets is difficult to determine, and hence, it is hard to say definitively whether or not dimensions were shared. One could also argue that three arbitrarily selected stimuli given the same label fails to capture much of what is meant by the term "category."

EXPERIMENT 1A

In this first experiment, we further investigated the evidence for nonattentional optional shift, using a stimulus set that might reasonably be described as categorical. At the same time, we wanted to use stimuli whose component features were relatively explicit and in which the formation of associations between those features could not give rise to optional shift behavior. If optional shift were found under these conditions, it would strengthen the argument for representations that mediate between stimulus representations and representations of the category label. The demonstration is important because most formal accounts of categorization do not include such a level.

The *polymorphous* category structure (Dennis et al., 1973) was chosen because it has a family resemblance structure (Rosch & Mervis, 1975) whilst having entirely uncorrelated features. The stimulus set employed in all the experiments reported in this article is illustrated in Figure 1A. Each collection of five features was a stimulus that belonged to either Category A or Category B. Category A was defined by the rule "at least three of exclamation mark, addition sign, up arrow, triangle, dollar sign" whilst Category B was defined by "at least three of question mark, multiplication sign, down arrow, square, pound

A

Category A	Category B	Reversal Exp 1A	Reversal Exp 1B
! + ↑ △ $	? × ↓ □ £		
? + ↑ △ $	! × ↓ □ £		
! × ↑ △ $	? + ↓ □ £		
! + ↓ △ $	? × ↑ □ £		
! + ↑ □ $	? × ↓ △ £		✓ NR
! + ↑ △ £	? × ↓ □ $		✓ NR
! + ↑ □ £	? × ↓ △ $	✓	
! + ↓ △ £	? × ↑ □ $	✓	✓ R
! × ↑ △ £	? + ↓ □ $	✓	✓ R
? + ↑ △ £	! × ↓ □ $		✓ NR
! + ↓ □ $	? × ↑ △ £	✓	✓ R
! × ↑ □ $	? + ↓ △ £	✓	✓ R
? + ↑ □ $	! × ↓ △ £		✓ NR
! × ↓ △ $	? + ↑ □ £	✓	
? + ↓ △ $	! × ↑ □ £		
? × ↑ △ $	! + ↓ □ £		

B

Category A	Category B
@ ♠ Ω ||| ↺	# ♥ α ≡ ↻

Figure 1. (A) The 32 stimuli employed in Experiments 1A and 1B. The leftmost column shows stimuli designated as examples of Category A. The next column shows stimuli designated as examples of Category B. The final two columns indicate which stimuli would be employed in the reversal phase of Experiments 1A and 1B, respectively. (B) The 10 symbols employed in Experiment 2.

sign.” This structure captured something of the *family resemblance* idea, in that category membership was determined by a number of characteristic but nondefining attributes. In addition, note that each of the feature pairs (e.g., question-mark/exclamation-mark) was statistically independent of each of the other feature pairs. For example, knowing that the stimulus contains a question mark did not allow one to predict what any of the other components would be. This was useful because it meant that associative accounts could not appeal to the formation of feature–feature associations (McLaren & Mackintosh, 2000; Rescorla & Durlach, 1981), a mechanism that might otherwise explain the optional shift behavior we hoped to demonstrate.

The design of our first experiment was based on earlier work with pigeons by von Fersen and Lea (1990) and is as follows. First, participants are trained to criterion on the two polymorphous categories illustrated in Figure 1A. Second, they are trained to criterion on a reversal of a subset of the original stimuli. This subset is selected on the basis that, within the context of the reversal phase, only one feature pair is predictive of category membership. Figure 1A provides an example set in which only the question-mark/exclamation-mark feature pair are diagnostic. During the reversal phase, all other features occur equally often in both categories. In the third and final phase, the participants are tested on the remaining stimuli in the absence of feedback.

The critical question is whether the participants will transfer the reversal of category labels learned in the second phase to the remaining stimuli presented in the third phase (the generalization phase). A categorization process *without* an ARMA component (i.e., without a mediating category representation) predicts between 0% and 50% reversed responses. Zero reversed responses occur where decisions are driven entirely by the first (acquisition) phase. Fifty percent reversed responses occur where decisions are driven entirely by the second (partial reversal) phase (e.g., where the participant ignores or entirely forgets the acquisition phase). This is because only one feature pair is diagnostic in the partial reversal phase and its components appear with equal frequency in both categories in the generalization phase. For example, in Figure 1A, question-mark/exclamation-mark is the diagnostic feature pair. In the generalization set, the question mark appears in five Category A stimuli and five Category B stimuli. The same is true for the exclamation mark. The relative contributions of the first and second phases will determine whether the number of reversed responses is closer to 0% or 50%. However, a system without an ARMA component cannot predict that the number of reversed responses will exceed 50%.

In contrast, a system with an ARMA component can predict up to 100% reversed responses. This is because the ARMA component can respond to the reversal phase by reversing the category → label associations, leaving the stimulus → category associations intact. However, unless the reversal of the category → label associations is very rapid, the reversal phase may lead to some disruption of what has been learned in the acquisition phase. This is because all feature pairs are diagnostic in acquisition but, in the reversal phase, one feature pair is reversed and the others are nondiagnostic. If category → label associations change very slowly, the system's predictions will be the same as those of a non-ARMA system—0% to 50% reversed responses (see above). If these associations change very rapidly, 100% reversal can be achieved. In other words, a system with an ARMA component can predict any percentage of reversed responses, whereas a non-ARMA system can predict between 0% and 50% reversed responses. Therefore, if participants show reliably more than 50% reversed responses in the generalization phase, evidence for an ARMA component will have been provided.

Method

Participants and Apparatus. Twenty-five undergraduate students participated in the experiment. They were paid at a rate of £4 an hour (approximately $6.40 an hour). The participants were tested individually in a quiet cubicle. An Acorn RISC PC computer with a 14-in. monitor controlled the experiment, and responses were collected via a standard PC-style keyboard.

Stimuli. Each stimulus was composed of five discrete components displayed in a row. Each position in the row contained one of two symbols (see Figure 1A). The stimuli were presented as large white characters (14 × 18 mm) in the center of a black background. The whole stimulus was approximately 5º of visual angle across. Figure 1A illustrates all 32 different stimuli that can be created from these five symbol pairs. For all the participants, the following symbols were arbitrarily selected as being characteristic of Category A: exclamation mark, addition sign, up arrow, triangle, and dollar sign. Under the polymorphous categorization rule employed, any stimulus containing more than two of these symbols was designated as an example of Category A. In a complementary manner, any stimulus containing fewer than three of these symbols was designated as an example of Category B. The leftmost column of Figure 1A shows all 16 Category A stimuli, and the second column shows all 16 Category B stimuli. Each column is ordered by number of characteristic symbols. The 2 stimuli in the first row have all five symbols characteristic of their category. Rows 2–6 show stimuli with four characteristic symbols. Rows 7–16 show stimuli with three characteristic symbols.

Procedure. Before commencing, the participants were informed that the experiment might take up to four 1-h sessions, each to be conducted on a different day and with no more than 2 days between each session. They were also informed that the experiment might take only a single session and that the number of sessions required depended on their performance.

Throughout the experiment, the stimuli were presented one at a time, and the participants categorized each stimulus as a member of either Category A or Category B by pressing one of two keys on the keyboard. Each stimulus was preceded by a 1-sec presentation of a small fixation cross in the center of the screen. The stimulus remained on the screen until either the participant had responded or 15 sec had elapsed. After 15 sec, the stimulus was replaced by the message "Please respond now."

The experiment comprised three phases: acquisition, partial reversal, and generalization. Each phase was subdivided into blocks, and the participants were advised to rest for a few seconds between each block. The participant initiated the next block by pressing a key on the keyboard.

In each block of the acquisition phase, all 32 stimuli were presented in a randomized order. Each categorization response was followed immediately by stimulus offset and a feedback message. A correct response produced the message "CORRECT. It was category *y*," where *y* was the appropriate category label (A or B). An incorrect response produced a short beep and the message "WRONG. It was category *y*," *y* again being the correct category label. The feedback message remained on the screen for 1 sec, after which the fixation cross was presented.

At the end of each acquisition block, the following message was displayed: "In the last block you got *x*% correct. You should be aiming for at least 90%," where *x* was the integer percentage of correct responses in the immediately preceding block. In fact, the acquisition phase continued until the participant made at least 27 correct responses in the same block (approximately 84.4% correct) or had completed a total of 48 blocks. If the participant completed 16 blocks without reaching criterion, the session was terminated, and the participant was asked to return the following day. The end of the acquisition phase was not explicitly signaled to the participant.

The partial reversal phase involved the repeated presentation of just 12 of the 32 acquisition stimuli. Category assignments in this phase were reversed, as compared with the acquisition phase. The 12 stimuli to be presented were selected on the basis that, within the context of those 12 stimuli, only one of the five symbol pairs was diagnostic of category membership. Figure 1A illustrates (ticked rows) the 12 stimuli for which only the question-mark/exclamation-mark symbol pair provides information about category membership. Across these 12 stimuli, all other symbols appear equally often in each category and, hence, are entirely nondiagnostic. Across participants, each symbol pair was selected equally often as the diagnostic pair. To summarize, for a participant who was reversed on the question-mark/exclamation-mark symbol pair, the 12 stimuli presented are illustrated by the ticked rows in Figure 1A, and the correct responses in this phase are opposite to the column headings.

In each block of the partial reversal phase, each of the 12 stimuli selected was presented three times. The participants were expected to meet or exceed a criterion of 31 correct responses in the same block (approximately 86.1% correct). The procedure was in all other respects identical to the procedure in the acquisition phase. The partial reversal phase ended when the participants reached criterion or when they had completed 16 blocks, whichever came first.

The generalization phase was signaled by the message, "From now on, you will no longer be told whether your responses are correct." The generalization phase comprised two blocks, each consisting of 20 stimuli. The stimuli presented were those from the 32 acquisition stimuli that had not been presented in the partial reversal phase. Across the two blocks, each of the 20 stimuli was presented exactly twice. Order of presentation was randomized. No feedback was given during the generalization phase. In other respects, the procedure was identical to that in the previous two phases.

Results

All the participants reached criterion in the acquisition phase, taking a mean of 13.76 blocks to do so ($SE = 2.35$). Performance in the final acquisition block is shown in Figure 2A. The graph shows performance for each of three stimulus types: the prototypes (Figure 1A, top row), the *one-aways* (stimuli that differ from the prototypes by one feature; Figure 1A, rows 2–6), and the *two-aways* (stimuli that differ from the prototypes by two features;

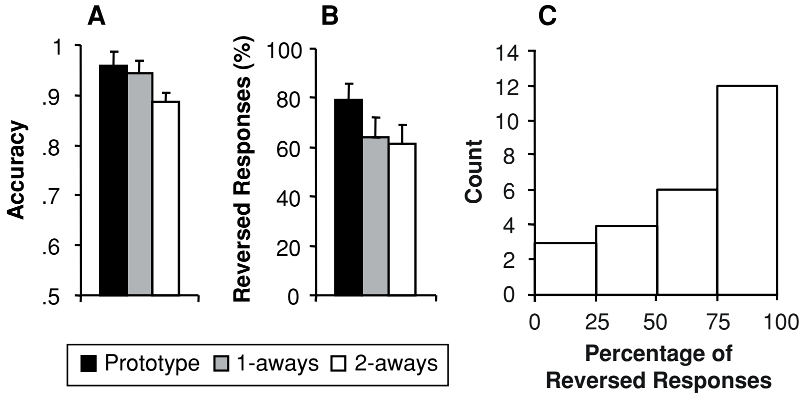

Figure 2. Results of Experiment 1A. (A) Final acquisition block accuracy by stimulus type. (B) Percentage of reversed responses in the generalization phase, by stimulus type. (C) Distribution of generalization performance (percentage of reversed responses) across participants. Error bars, where shown, represent one standard error.

Figure 1A, rows 7–16). Stimulus type significantly affects performance (Friedman test statistic = 13.68, $p <$ .001). A Friedman test is used for comparability with subsequent experiments in this article, in which the standard error for the prototype stimuli is zero and, hence, an ANOVA is inappropriate.

All the participants also reached criterion in the partial reversal phase, taking a mean of 3.28 blocks to do so (*SE* = 0.71). Responses in the generalization phase were coded as reversed or nonreversed with respect to the correct response in the acquisition phase. For example, if the correct response to a stimulus in the acquisition phase was “Category A,” a “Category B” response to that stimulus in the generalization phase was coded as a reversed response. A mean of 64.5% of responses were reversed in the generalization phase (*SE* = 6.32). This is significantly higher than 50% [$t(24) = 2.29$, $p < .05$], which can be predicted by systems with an ARMA component, but not by those without one.

Nevertheless, the mean percentage is numerically quite close to 50%, and the standard deviation is high. Might this group average be hiding some important individual differences—perhaps, even a bimodality where some participants are performing as if they were utilizing an ARMA component and others as if they were not? The distribution is shown in Figure 2C. Although there are no obvious signs of bimodality, the distribution is quite skewed, which, at this sample size, might raise slight concerns about the appropriateness of the one-sample *t* test we have employed. Fortunately, the observed skew can be corrected by a standard arcsine transformation: $2 \times \arcsin(\sqrt{p})$, where p is the number of reversals expressed as a proportion. Our critical result (a number of reversals exceeding 50%) is still reliable after this transformation [$t(24) = 2.18$, $p < .05$].

Given that accuracy in the acquisition phase is affected by stimulus type (prototype, one-away, or two-away), one might expect to see a corresponding pattern in the generalization phase. Figure 2B shows this expected trend, which is statistically reliable (Friedman test statistic = 8.78, $p < .05$). A Friedman test was used because some of the distributions for particular stimulus types were severely skewed and none of the transformations attempted fully corrected for this.

Discussion

A reversal trained to a subset of the original acquisition stimuli showed significant transfer to the remainder of the stimuli in the absence of further feedback. This *optional shift* behavior seems difficult to explain if one assumes that stimulus representations are directly linked to category labels but is straightforwardly accommodated by accounts that posit a mediating layer of category representations. Unlike some previous demonstrations of optional shift, an alternative explanation in terms of selective attention seems unlikely. More specifically, although selective attention may have occurred in our study, it seems unlikely to have resulted in significant transfer of the reversal. For example, if, during the course of the reversal phase, participants develop complete selective attention to the pair of features that are predictive of category membership, they should make equal numbers of reversed and nonreversed responses in the generalization phase. This is because these diagnostic features occur, within the context of the generalization phase, equally often in Category A and Category B.

EXPERIMENT 1B

Although selective attention cannot explain the results of Experiment 1A, there are processes other than the ARMA component that could give rise to there being more than 50% reversed responses. Experiment 1B addressed one particular alternative: generalization at a feature level. In a system with no ARMA component, it is still possible that the reversal of one feature pair in the second phase generalizes to the other feature pairs, which are nondiagnostic in that phase. The process by which this occurs cannot be feature–feature associations formed during acquisition, because our chosen category structure ensures that the feature pairs are not correlated with each other. However, one can imagine nonassociative processes that would generalize new information from one feature pair to other feature pairs whose status is uncertain (i.e., the nondiagnostic pairs).

In Experiment 1B, we changed the nature of the reversal phase so that an ARMA component and a feature generalization process would make different predictions. The acquisition phase was identical to that in Experiment 1A. Also, as in Experiment 1A, the current experiment reversed one feature pair in the reversal phase. However, two feature pairs were *non*reversed, and the remaining two were nondiagnostic. Again, as in Experiment 1A, the generalization set was simply the remainder of the stimuli. Figure 1A shows an example reversal stimulus set in which, reading left to right, the feature pairs are reversed, nonreversed, nonreversed, nondiagnostic, and nondiagnostic, respectively.

As Figure 1A illustrates, this resulted in a reversal phase in which half of the stimuli required the same response as in the acquisition phase and half required the opposite response. Therefore, unlike in Experiment 1A, the categorization presented in the reversal phase could not be solved by reversing the category → label associations. A system with an ARMA component but no feature-level generalization can predict up to 25% reversed responses in the generalization phase by responding entirely on the basis of the information in the reversal phase (in the Figure 1A example, such a system would show reversed responding in the generalization phase to the stimuli in the 7th and 17th rows). If the system ignores the reversal phase entirely but recalls the acquisition phase perfectly, it can produce 0% reversed responses. If the system ignores both acquisition and reversal phases in responding to the generalization stimuli and simply responds randomly, it can predict up to 50% reversed responses. In summary, a system with an ARMA component can predict anything from 0% to 50% reversed responses.

In contrast, a system with feature-level generalization but no ARMA component can predict up to 75% reversed responses, under the assumption that the reversal of the reversed feature pair generalizes fully to the two nondiagnostic feature pairs but leaves the nonreversed feature pairs entirely intact. Such a participant would *not* reverse stimuli on the 2nd and 7th rows of the Figure 1A example but would reverse all other generalization stimuli. A system with both feature-level generalization and an ARMA component can predict anything from 0% to 75%, depending on which process dominates. A system in which neither process operates predicts less than 50% reversed responses. This is because "Category A" generalization stimuli contain more of Symbol Set 1 (!, +, up arrow, triangle, and $) than of Symbol Set 2 (?, X, down arrow, square, and £). On average, this is also true of the stimuli in the reversal phase. Assuming a positive monotonic relationship between features shared and amount of generalization, most accounts with neither process would predict fewer than 50% reversed responses in the third phase.

In summary, if the results of Experiment 1A were due to feature-level generalization, rather than to an ARMA component, one might expect to see more than 50% reversed responses in the present experiment. In contrast, if the results of Experiment 1A were due to an ARMA component, one would expect to see fewer than 50% reversed responses in the present experiment (although one could see an average of 50% responses across participants if they responded randomly in the generalization phase).

Method

Participants, Apparatus, and Stimuli. Twenty-five undergraduate students participated in the experiment. None had participated in Experiment 1A. The apparatus and stimuli employed were the same as those in Experiment 1A.

Procedure. The procedure was, in most respects, identical to that in Experiment 1A. The differences were as follows.

The partial reversal phase in the present experiment involved 16, rather than 12, stimuli. Each of the 16 stimuli was presented exactly twice in a block, and the order of presentation was randomized. The 16 stimuli were selected on the basis that, within the context of those 16 stimuli, three of the feature pairs were diagnostic of category membership and two were nondiagnostic. Figure 1A illustrates (tick marks, final column) a selection where the exclamation-mark/question-mark, plus-sign/multiply-sign, and up-arrow/down-arrow feature pairs are diagnostic and the other two pairs are not.

In terms of determining feedback in this phase, one of the three diagnostic pairs was considered to have reversed its meaning, relative to the acquisition phase. So, for example, if exclamation-mark/question-mark was selected as the reversed feature pair, a question mark was considered a feature diagnostic of Category A, and an exclamation mark diagnostic of Category B. Figure 1A illustrates the effect this would have on feedback if exclamation-mark/question-mark was reversed, plus-sign/multiply-sign and up-arrow/down-arrow were nonreversed, and the remaining two pairs were nondiagnostic. For stimuli marked "R," the correct category response is opposite to the column heading. For stimuli marked "NR," the correct category response is the same as the column heading.

Each participant experienced a different, but otherwise random, allocation of the five feature pairs to the roles of reversed diagnostic, nonreversed diagnostic, and nondiagnostic feature pairs. The criterion for the partial reversal phase was at least 28 out of 32 correct responses within the same block (87.5% correct).

As in Experiment 1A, the generalization phase involved the presentation of stimuli not presented in the partial reversal phase. Hence, each block in the generalization phase involved the presentation of 16 stimuli, rather than the 20 presented in Experiment 1A.

Results

Two participants were excluded from the following analysis because they did not reach criterion in the acquisition phase within the 48 blocks available to do so. A further 8 participants failed to reach criterion in the partial reversal phase within the 16 blocks available. All the analyses were performed on the data from the 15 participants who reached both criteria.

The participants took a mean of 22.27 blocks to reach the acquisition criterion ($SE = 5.21$). Performance in the final acquisition block is shown in Figure 3A. As in Experiment 1A, the graph shows performance for each

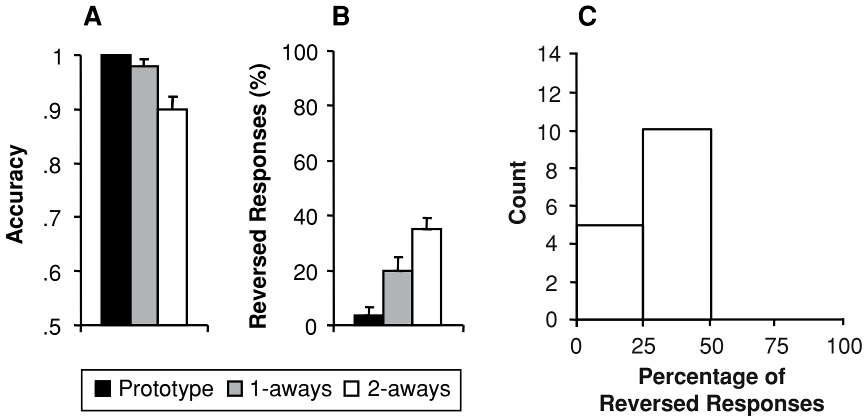

Figure 3. Results of Experiment 1B. (A) Final acquisition block accuracy by stimulus type. (B) Percentage of reversed responses in the generalization phase, by stimulus type. (C) Distribution of generalization performance across participants. Error bars, where shown, represent one standard error.

of three stimulus types: the prototypes, the one-aways, and the two-aways. Stimulus type significantly affected performance (Friedman test statistic = 9.70, $p < .01$). A Friedman test was used because the standard error for the prototype stimuli was zero (and hence, an ANOVA was inappropriate).

The participants took a mean of 5.33 blocks to reach the partial reversal criterion (SE = 1.72). In our critical measure, a mean of 25.42% of the responses (SE = 2.92) were reversed in the generalization phase, a value that is significantly lower than 50% [$t(14) = 8.41, p < .0005$]. This critical result can be predicted by an ARMA account, but not by a feature generalization account, of the results of Experiment 1A. Figure 2C shows the distribution of responses across participants. As in Experiment 1A, there is no obvious bimodality, but some evidence of skew. The skew can be corrected by the transform 2 × arcsin[$\sqrt{(1-p)}$], where p is the number of reversals expressed as a proportion. Our critical one-sample t test is still reliable after this transformation [$t(14) = 7.02, p < .0005$].

Given that accuracy in the acquisition phase is affected by stimulus type (prototype, one-away, or two-away), one might expect to see a corresponding pattern in the generalization phase. Figure 2B shows this expected trend, which is statistically reliable (Friedman test statistic = 9.70, $p < .01$). A Friedman test was used because some of the distributions for particular stimulus types were severely skewed and none of the transformations attempted fully corrected for this.

Discussion

The results of our first experiment (Experiment 1A) were consistent with the presence of an ARMA component but could alternatively be explained by generalization of the reversal at a feature level. The results of the present experiment (Experiment 1B) support the former account of Experiment 1A, although they provide no evidence for the latter. This is not to say that feature-level generalization does not occur but simply that it seems unlikely to underlie the results of Experiment 1A. Taken together, Experiments 1A and 1B support the idea that there are representations that mediate between stimulus and category label representations.

One question that remains unanswered is what this mediating layer actually represents. Although our evidence is consistent with the idea of a mediating layer of category representations, one could alternatively suggest that mediation was at the level of the response set. In other words, the mediating layer contains representations of everything that requires a specific response (e.g., "Category A"). The distinction was unimportant within the context of our first two experiments, because there was a one-to-one mapping between categories and category labels. Nevertheless, it is possible to conceive of a situation in which two distinct classes of items attract the same label. Would such a situation result in one response set mediator or two category mediators? Demonstration of a mediating layer of category representations clearly requires evidence of the latter. In the next experiment, we investigated whether such evidence could be found.

EXPERIMENT 2

Experiments 1A and 1B provided evidence that is consistent with the idea of category representations but that is equally consistent with representations that are at the level of a response set. Experiment 2 was an attempt to distinguish between these two alternative accounts by training two polymorphous classification problems simultaneously. Both classifications employed the same two category labels (A and B) but used a different set of stimulus symbols and, hence, were distinct pairs of categories. Once both classification problems had been mastered, the participant was confronted with a reversal of a subset of stimuli from just one of the two classification problems. The reversal procedure from Experiment 1A was employed, since it had previously been shown to produce reliable optional shift behavior. Stimuli from the other classification problem were not presented in this phase. Finally, optional shift behavior was tested using the remainder of the stimuli from the reversed classification *and* the corresponding subset of the other classification problem.

If the hypothesized mediating representations are at the level of a response set, a predominance of reversed responses for both classification problems would be predicted. This is because, in line with the explanation of Experiment 1A, the reversal in the second phase would be accommodated by reversing the response set → category label associations, leaving the stimulus → response set associations relatively intact. However, if the mediating representations are category specific, the prediction would be different. Experiment 2 involved two labels but four distinct categories, so a category-mediating layer would presumably develop a distinct representation for each of the four. Each of these representations would form its own associations to the category labels, and this being the case, the reversal in the second phase could be accommodated by reversing the associations between the categories present in that phase and the category labels. This would leave the category → label associations of the nonreversed categories and the stimulus → category associations of all the categories relatively intact. Hence, a predominance of reversed responses would be predicted for the reversed classification problem, whereas a predominance of nonreversed responses would be predicted for the nonreversed classification problem.

Method

Participants and Apparatus. Twenty-five adults participated in the experiment. None had participated in either of the previous two experiments. The apparatus employed was the same as that in the previous two experiments.

Stimuli. The present experiment employed the 32 stimuli from Experiment 1, plus an additional 32 stimuli. The additional stimuli formed two polymorphous categories with the same logical structure as those shown in Figure 1A. The only difference was that each of the five symbol pairs employed in Figure 1A was, for this additional stimulus set, replaced by a different symbol pair, illustrated in Figure 1B.

Procedure. The procedure was, in most respects, the same as that in Experiments 1A and 1B. The differences were as follows.

During the acquisition phase, the two symbol sets (see the Stimuli section) were employed on alternate blocks. For all the participants, the set shown in Figure 1A was employed on odd blocks, whereas the set illustrated by Figure 1B were employed on even blocks. Criterion in the acquisition phase was now 54 out of 64 correct responses across two consecutive blocks, the criterion being checked at the end of each even-numbered block. The participants now received the end-of-block feedback message on every other block, with the message modified to read, "In the last two blocks . . ." The participants were given a maximum of 80 blocks over five sessions to reach criterion.

As in Experiment 1A, the partial reversal phase involved the repeated presentation of 12 stimuli that were selected on the basis that only one feature pair was diagnostic across those 12 stimuli. For 13 participants, those 12 stimuli were drawn from those illustrated in Figure 1A. For the remaining 12 participants, they were drawn from the stimuli illustrated by Figure 1B. Across the entire partial reversal phase, any given participant saw only stimuli from the Figure 1A set *or* only stimuli from the Figure 1B set. Across participants, each of the five possible locations within the stimulus contained the diagnostic symbol pair an equal number of times.

The generalization phase comprised two blocks, each of 40 stimuli. Twenty of these stimuli were the remainder of the set presented in the partial reversal phase. The other 20 were the corresponding stimuli from the set not presented in the partial reversal phase.

Results

Two participants quit the experiment before completing it. A further 11 participants failed to reach the acquisition criterion within the time available. All the analysis was performed on the data from the remaining 12 participants.

The participants took a mean of 27.17 blocks to reach the acquisition criterion ($SE = 6.83$). Performance in the final acquisition block for each of the two stimulus sets is shown in Figure 4A ("Set 1" corresponds to Figure 1A, "Set 2" to Figure 1B). As in the previous experiments, the graphs show performance for each of three stimulus types: the prototypes, the one-aways, and the two-aways. Stimulus type significantly affected performance for both sets (Friedman test statistic = 11.38, $p < .01$, and Friedman test statistic = 12.17, $p < .01$, respectively). A Friedman test was used because the standard error for the prototype stimuli was zero (and hence, an ANOVA was inappropriate).

The participants took a mean of 3.00 blocks to reach the partial reversal criterion. In our critical generalization phase tests, a mean of 76.87% reversed responses were made to stimuli from the reversal set, whereas a mean of 20% reversed responses were made to stimuli from the nonreversal set. Both differ significantly from 50% [$t(11) = 2.89, p < .05$, and $t(11) = 3.72, p < .01$, respectively]. This pattern of results is predicted by category-level mediating representations, but not by response set mediating representations.

Figures 4C and 4D show the distribution of these response percentages across participants. As in the previous experiments, there is no obvious bimodality but some evidence of skew. The skew in the reversal set can be corrected by the transform $2 \times \arcsin(\sqrt{p})$, and our critical one-sample t test is still reliable after this transformation [$t(11) = 2.59, p < .05$]. The skew in the nonreversal set can be corrected by the transform $2 \times \arcsin[\sqrt{(1-p)}]$, and our critical one-sample t test is still reliable after this transformation [$t(11) = 3.73, p < .01$].

Given that accuracy in the acquisition phase was affected by stimulus type (prototype, one-away, or two-away), one might expect to see corresponding patterns in the generalization sets. Figure 4B shows some evidence of trends in the expected direction, but these trends are not reliable in either set (Friedman test statistic = 0.04, $p >$ .5, for the reversal set, and Friedman test statistic = 3.04, $p >$ .2, for the nonreversal set).

Discussion

Unsurprisingly, the requirement to concurrently acquire two polymorphous classifications substantially increased the difficulty of the task, as evidenced by the reduced proportion of participants who reached crite-

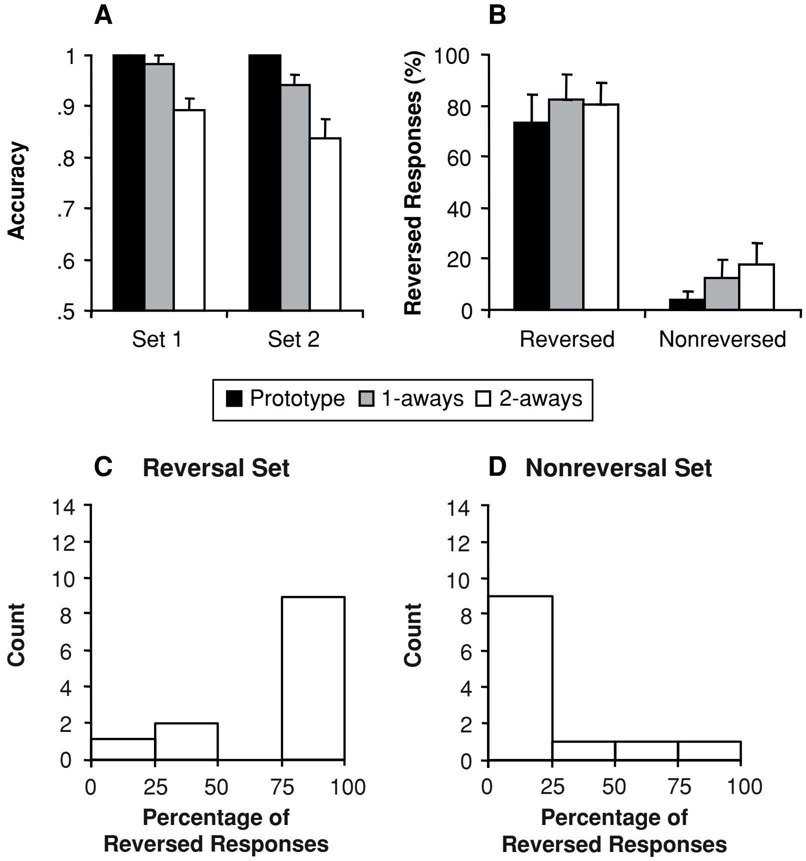

Figure 4. Results of Experiment 2. (A) Final acquisition block accuracy by stimulus type and by stimulus set. (B) Percentage of reversed responses in the generalization phase by stimulus type and training type. (C and D) Distribution of generalization performance across participants for the partially reversed stimulus set (C) and the nonreversed stimulus set (D). Error bars, where shown, represent one standard error.

rion. Nevertheless, the participants who did reach criterion behaved in the manner one would expect if category representations (rather than response set representations) mediated between their representations of the stimuli and their representations of the category labels. One might have expected to see effects of stimulus type in the generalization phase that were comparable to the stimulus type effects observed in Experiments 1A and 1B. However, although the numerical trends were approximately as expected, neither effect was statistically reliable. This seems likely to have been due to the smaller sample size in this experiment (due, in turn, to more participants failing to reach criterion).

GENERAL DISCUSSION

For at least 90 years, associative theorists have argued for representations that mediate between those of stimulus and response. The argument can be traced from Watson (1913) through the work of Hull (1930) and the Kendlers (e.g., Kendler, Kendler, & Ward, 1972) to the hidden layers of the ubiquitous back-propagation architecture (Rumelhart, Hinton, & Williams, 1986) and much of modern neuroscience (see, e.g., Gazzaniga, Ivry, & Mangun, 1998). Today, few psychologists would argue against the presence of such representations. Against this context, the argument becomes much more about the specific nature of the representations and the levels of processing at which they occur. Apart from exceptions such as the AMBRY model (Kruschke, 1996), formal models of categorization tend not to explicitly posit category-level representations. Instead, they implicitly or explicitly assume that some level of stimulus representation associates directly to a representation of a category label.

The present study suggests that category representations are not of this nature. Instead, our data are more consistent with the idea that each distinct class of objects we have encountered evokes a distinct category representation that is dissociable from the category label it is given—"a rose by any other name," rather than "a rose is a rose." The separate category and category label representations of the AMBRY model capture this conclusion neatly within a formal associative system that can be employed in many different types of formal models.

Our stimuli were designed under the assumption that stimulus representations involve either the component features of the stimulus (Gluck & Bower, 1988) or the specific configurations (Gluck, 1991) or exemplars (Medin & Schaffer, 1978) they make up, or the stimulus's position or distribution in psychological space (Ashby & Gott, 1988; Nosofsky, 1986), or some combination of these. We have also assumed a relatively orderly relationship between the "physical" similarity of stimuli (in terms of number of features shared) and their psychological similarity as perceived by our participants. Such assumptions are not universally adopted in the formal modeling of categorization, but they are widely used. Hence, although further research is required, we feel that the present experiments make progress toward the conclusion that learning a novel classification problem can result in the development of category representations that are dissociable from representations of the category labels. One caveat is that participants may have created these category representations in response to the unexpected reversal, rather than in response to the original presentation, of the categories. Given that the methodology we have employed here critically relies on reversal, this possibility cannot be discounted on the basis of the present evidence. It is, however, an important issue for future research.

The idea that participants might create category representations on demand has a number of resonances with other phenomena. Work on free classification (see Wills & McLaren, 1998, for a brief review) indicates that category representations can be formed even in the absence of feedback. Markman and Ross (2003) have argued that different category-related learning tasks feed into a common category representation. Johansen and Palmeri (2002) have proposed that representational shifts occur during category learning, with prototype-like representations emerging as training continues. Schyns and Rodet (1997) have claimed that categorization can drive the creation of new *feature-level* representations. An important issue for the future is to determine how these phenomena relate to each other and to what extent they are produced by a common or different underlying processes.

REFERENCES

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **14**, 33-53.

Buss, A. H. (1953). Rigidity as a function of reversal and non-reversal shifts in the learning of successive discriminations. *Journal of Experimental Psychology*, **45**, 75-81.

Delamater, A. R., & Joseph, P. (2000). Common coding in symbolic matching tasks with humans: Training with a common consequence or antecedent. *Quarterly Journal of Experimental Psychology*, **53B**, 255-273.

Dennis, I., Hampton, J. A., & Lea, S. E. G. (1973). New problem in concept formation. *Nature*, **243**, 101-102.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (1998). *Cognitive neuroscience: The biology of the mind*. London: Norton.

Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science*, **2**, 50-55.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, **117**, 227-247.

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General*, **130**, 116-139.

Hall, G. (1996). Learning about associatively activated stimulus representations: Implications for acquired equivalence and perceptual learning. *Animal Learning & Behavior*, **24**, 233-255.

Hall, G., Mitchell, C., Graham, S., & Lavis, Y. (2003). Acquired equivalence and distinctiveness in human discrimination learning: Evidence for associative mediation. *Journal of Experimental Psychology: General*, **132**, 266-276.

Hull, C. L. (1930). Knowledge and purpose as habit mechanisms. *Psychological Review*, **37**, 511-525.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, **45**, 482-553.

Kendler, H. H., & Kendler, T. S. (1968). Mediation and conceptual behavior. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 197-244). New York: Academic Press.

Kendler, H. H., Kendler, T. S., & Ward, J. W. (1972). An ontogenetic analysis of optional intradimensional and extradimensional shifts. *Journal of Experimental Psychology*, **95**, 102-109.

Kendler, H. H., Kendler, T. S., & Wells, D. (1960). Reversal and nonreversal shifts in nursery school children. *Journal of Comparative & Physiological Psychology*, **53**, 83-87.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22-44.

Kruschke, J. K. (1996). Dimensional relevance shifts in category learning. *Connection Science*, **8**, 225-247.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, **129**, 592-613.

McLaren, I. P. L., & Mackintosh, N. J. (2000). An elemental model of associative learning: I. Latent inhibition and perceptual learning. *Animal Learning & Behavior*, **28**, 211-246.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

Murdock, B. B., Jr. (1956). "Backward" learning in paired associates. *Journal of Experimental Psychology*, **51**, 213-215.

Norcross, K. J., & Spiker, C. C. (1958). Effects of mediated associations on transfer in paired-associate learning. *Journal of Experimental Psychology*, **55**, 129-134.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-61.

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

Rescorla, R. A., & Durlach, P. J. (1981). Within-event learning in Pavlovian conditioning. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 81-111). Hillsdale, NJ: Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **7**, 573-605.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.

Sanders, B. (1971). Factors affecting reversal and nonreversal shifts in rats and children. *Journal of Comparative & Physiological Psychology*, **74**, 192-202.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 681-696.

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior*, **37**, 5-22.

Sutherland, N. S., & Mackintosh, N. J. (1971). *Mechanisms of animal discrimination learning*. New York: Academic Press.

Urcuioli, P. J., Zentall, T. R., Jackson-Smith, P., & Steirn, J. N. (1989). Evidence for common coding in many-to-one matching: Retention, intertrial interference, and transfer. *Journal of Experimental Psychology: Animal Behavior Processes*, **15**, 264-273.

von Fersen, L., & Lea, S. E. G. (1990). Category discrimination by pigeons using five polymorphous features. *Journal of the Experimental Analysis of Behavior*, **54**, 69-84.

Watson, J. B. (1913). Psychology as the behaviorist sees it. *Psychological Review*, **20**, 158-177.

Wills, A. J., & McLaren, I. P. L. (1998). Perceptual learning and free classification. *Quarterly Journal of Experimental Psychology*, **51B**, 235-270.

Wills, A. J., Reimers, S., Stewart, N., Suret, M., & McLaren, I. P. L. (2000). Tests of the ratio rule in categorization. *Quarterly Journal of Experimental Psychology*, **53A**, 983-1011.

Wirth, O., & Chase, P. N. (2002). Stability of functional equivalence and stimulus equivalence: Effects of baseline reversals. *Journal of the Experimental Analysis of Behavior*, **77**, 29-47.

NOTES

[1] One might suggest that the training resulted in the children constructing an artificial dimension with *green square* as one end point and *yellow triangle* as the other end point. Selective attention was then directed toward this new dimension, (cf. Goldstone & Steyvers, 2001), which would have facilitated a full reversal but would have interfered with a partial reversal. However, such an account does not seem very different from the category representation development account we are offering. If a participant develops a dimension upon which each category occupies a distinct region, it does not seem inappropriate to describe this as category representation development.

[2] The name was chosen to acknowledge the important contribution of the AMBRY model, while underlining that the idea it contains is a generally useful theoretical tool that could be incorporated into many formal accounts (*ambry* is derived from the Latin *arma*, meaning *utensil*).

(Manuscript received June 13, 2003;
revision accepted for publication February 3, 2005.)