Mem Cogn
DOI 10.3758/s13421-013-0350-5

The effect of newly trained verbal and nonverbal labels for the cues in probabilistic category learning

Fotis A. Fotiadis · Athanassios Protopapas

Abstract Learning in a well-established paradigm of probabilistic category learning, the *weather prediction task*, has been assumed to be mediated by a variety of strategies reflecting explicit learning processes, such as hypothesis testing, when it is administered to young healthy participants. Higher categorization accuracy has been observed in the task when explicit processes are facilitated. We hypothesized that furnishing verbal labels for the cues would boost the formation, testing, and application of verbal rules, leading to higher categorization accuracy. We manipulated the availability of cue names by training separate groups of participants for three consecutive days to associate hard-to-name artificial auditory cues to pseudowords or to hard-to-name ideograms, or to associate stimulus intensity with colors; a fourth group remained unexposed to the cues. Verbal labels, cue individuation, and exposure to the stimulus set each had an additive effect on categorization performance in a subsequent 200-trial session of the weather prediction task using these auditory cues. This study suggests that cue nameability, when controlled for cue individuation and cue familiarity, has an effect on hypothesis-testing processes underlying category learning.

Keywords Categorization · Training · Memory · Verbal labels

Categorization is a fundamental aspect of cognition underlying a broad range of human behaviors and skills, such as language acquisition, inference, concept formation, and decision making. The cognitive neuroscience of category learning has extensively tried to shed light on its mechanisms, representational contents, and neural substrates. Alternative approaches suggest that category learning is mediated either by qualitatively distinct systems (Ashby & Maddox, 2011; Poldrack & Foerde, 2008) or by a single learning mechanism (Newell, Dunn, & Kalish, 2011).

Electronic supplementary material The online version of this article (doi:10.3758/s13421-013-0350-5) contains supplementary material, which is available to authorized users.

F. A. Fotiadis (✉) · A. Protopapas
Department of Philosophy & History of Science, University of Athens, Ano Ilissia Campus, GR-157 71 Athens, Greece
e-mail: f_fotis@phs.uoa.gr

Explicit hypotheses in category learning

Multiple-systems theorists have drawn a distinction between a declarative, explicit, or verbal system or pathway, and a procedural, implicit, or nonverbal system (Ashby & Maddox, 2005, 2011; Minda & Miles, 2010; Poldrack & Foerde, 2008; Squire, 2004). The explicit system is thought to be engaged when hypothesis-testing processes—such as the formation, testing, and application of a verbalizable rule or strategy—can lead to successful performance and the knowledge acquired is accompanied by awareness. The implicit system underlies performance when no verbalizable rules exist or can easily be applied, in which case the integration of information across multiple trials occurs or perceptual learning processes are recruited. Knowledge acquired by the implicit system is considered unavailable to conscious recollection. The two systems have been suggested to compete (Ashby, Alonso-Reese, Turken, & Waldron, 1998; Poldrack et al., 2001) or to operate in parallel (Dickerson, Li, & Delgado, 2011; Minda & Miles, 2010; Shohamy, Myers, Kalanithi, & Gluck, 2008).

Single-system theorists, on the other hand, have questioned the parsimony of multiple categorization systems (Newell et al., 2011) and the validity of the methodologies (e.g., double dissociations) utilized in the past (Newell & Dunn, 2008; Newell, Dunn, & Kalish, 2010). Instead, they have suggested that human categorization is achieved through a single, general learning mechanism (Newell, Lagnado, & Shanks, 2007) and is accompanied by high levels of awareness for the learned material (Lagnado, Newell, Kahan, & Shanks, 2006). The

hypothesis of multiple memory systems or pathways remains a matter of current debate in the study of categorization (e.g., Ashby & Maddox, 2011; Newell et al., 2011).

Regardless of the existence and functional independence of discrete categorization systems, few would argue against the notion that category learning employs—in at least some task structures—hypothesis-testing processes (Ashby & Maddox, 2005), inner rehearsal (Lupyan, Rakison, & McClelland, 2007), or verbalizable strategies (Gluck, Shohamy, & Myers, 2002). Executive functioning mechanisms have been argued to contribute to category learning by means of the formulation, testing, and application of verbal rules of category membership (Price, 2009). In particular, human category learning has been argued to be influenced by verbal processes (Minda & Miles, 2010), since "humans have the potential benefit of [verbal] labels" (Lupyan et al., 2007, p. 1077).

Although language in general seems to play an important role in category learning, researchers have mainly manipulated the category structure (i.e., the availability of an easily verbalizable rule) to examine the effect of verbal processes on categorization (Ashby & Maddox, 2005; Miles & Minda, 2011). Previously, Lupyan (2006; Lupyan et al., 2007) studied the influence of category labels. He showed that verbal labels—as opposed to location cues—facilitated the categorization of artificial stimuli when paired with category classes. However, not much attention has been drawn to the existence of labels for the items to be categorized. It stands to reason that if the stimuli are accompanied by verbal labels, then hypothesis-testing or inner-rehearsal processes will be facilitated, because participants would find it easier to form, test, and apply rules such as "respond 'rain' whenever the triangle card is present" (Gluck et al., 2002, p. 416). In contrast, in the case of nonnameable stimuli, it would not be so easy to explicitly state and apply rules concerning them.

In the present study, we sought to test this idea by using hard-to-name cues in the context of a prototypical probabilistic category-learning task. Participants were first trained to learn novel nonsense verbal labels or other hard-to-name pairings for the cues. They were subsequently administered the category-learning task using these cues, in order to explore the effects of cue nameability on learning to categorize.

The weather prediction task

The prototypical weather prediction task (WPT; Knowlton, Squire, & Gluck, 1994) is a perceptual categorization task based on a paradigm developed by Gluck and Bower (1988). Participants are asked to classify combinations (patterns) of four cards with geometric shapes (cues) into one of two possible outcomes, namely "sun" and "rain." The task has a probabilistic structure, in that each cue is associated with an outcome with a fixed probability. Two of the cues are highly predictive, and the other two are less predictive of a specific outcome. Overall, throughout training a combination of cues may predict one outcome on some trials, whereas on other trials the same combination may predict the alternative outcome (see the Method section). Corrective feedback is provided after every trial. It is now well-established that both healthy and brain-damaged participants gradually improve in categorization accuracy in a variety of versions (i.e., visual stimuli serving as cues, and category classes) of this task (e.g., Hopkins, Myers, Shohamy, Grossman, & Gluck, 2004; Knowlton et al., 1994).

The WPT has been widely used by multiple-systems theorists to assess the relative contribution of explicit (declarative) and implicit (procedural)[1] learning processes to the acquisition of knowledge (Poldrack & Rodriguez, 2004). Early neuropsychological studies suggested that the task mainly taps procedural learning processes (Knowlton, Mangels, & Squire, 1996; Knowlton et al., 1994; Reber, Knowlton, & Squire, 1996). However, neuroimaging studies (Poldrack et al., 2001), mathematical modeling of healthy participants' behavior (Gluck et al., 2002), and reexamination of clinical populations' behavior (Hopkins et al., 2004; Shohamy, Myers, Onlaor, & Gluck, 2004) have indicated an engagement of both declarative and procedural processes, presumably at different periods in training.

The mathematical modeling of young healthy participants' behavior has suggested that, early in the task, participants use suboptimal verbalizable strategies (Gluck et al., 2002; Meeter, Myers, Shohamy, Hopkins, & Gluck, 2006; Meeter, Radics, Myers, Gluck, & Hopkins, 2008) that can be said to be declarative (Shohamy et al., 2008). Later in training, participants shift to optimal multicue strategies. These later strategies have also been suggested to be accompanied by high levels of self-insight (Lagnado et al., 2006) or awareness (Price, 2009), and thus can be said to reflect explicit processes as well. Newell et al. (2007) suggested that the task is mediated by a single, explicit learning mechanism. Similarly, Poldrack and Foerde (2008) suggested that normal young adults may use declarative learning strategies to solve the task. Thus, although the WPT is a legacy of the multiple-systems field, recent research has suggested that young healthy participants' behavior is mediated by explicit learning processes entailing hypothesis testing of verbal rules (Price, 2009).

Researchers have experimentally manipulated the engagement of explicit processes during the WPT. Gluck et al. (2002) tested young healthy participants in two versions of the WPT.

[1] The terms *declarative* and *procedural* have been used to denote memory systems (e.g., Squire, 2004), whereas the terms *explicit* and *implicit* learning denote processes assessed by direct or indirect experimental tests of knowledge (e.g., Reber & Johnson, 1994). Some researchers use *declarative* and *explicit*, as well as *procedural* and *implicit*, interchangeably (Price, 2009), in an effort to reconcile the memory-systems and learning-processes approaches.

When the cue–outcome contingencies were less probabilistic (in their Exp. 2)—a manipulation thought to encourage declarative mediation (Foerde, Knolwton, & Poldrack, 2006)—performance measures increased throughout training, relative to a more probabilistic version (in their Exp. 1). Secondary task demands were introduced during WPT training in order to hamper explicit processes, resulting in the impairment of WPT categorization performance throughout (Foerde, Poldrack, & Knowlton, 2007) or during the second half of training (Foerde et al., 2006; Newell et al., 2007), as compared to single-task conditions. More recently, Price (2009, Exp. 2) reduced the time available for feedback processing, in order to impair explicit processes. Participants' performance was consistently greater in the long-feedback than in the short-feedback version. Thus, empirical data suggest that experimental manipulations favoring explicit processes result in higher categorization accuracy. Consistent with this interpretation, a reduction in WPT performance is also observed in special populations thought to be less efficient or impaired in their declarative encoding, and thus less able to form, test, and apply verbal rules, such as older healthy participants (Abu-Shaba, Myers, Shohamy, & Gluck, 2001) or hypoxic patients with medial temporal lobe lesions (Hopkins et al., 2004), respectively.

Design and rationale of the present study

In the present study, we employed a cue–response trial-and-error training paradigm modeled on the WPT. We used computer-generated auditory tones as cues because the majority of people do not possess preestablished labels for tones (Galizio & Baron, 1976). Prior to the WPT procedure, two groups of participants received extensive training to associate four novel auditory cues to pseudowords (label-training condition) or to hard-to-name ideograms (ideogram-training condition). A third group of participants were exposed to the same stimuli over the same number of trials, but learned to associate sound intensity to hard-to-name colors (intensity-training condition), disregarding cue identity. A fourth group remained unexposed to the auditory cues (no-training condition). All groups were subsequently administered an auditory version of the WPT (Fotiadis, Protopapas, & Vatakis, 2011) utilizing these cues.

The main hypothesis and motivation underlying our study were as follows: If verbal labels facilitate the formation, testing, and application of verbalizable rule-based strategies, and if facilitating explicit learning processes is accompanied by higher categorization accuracy (Price, 2009), then the label-training group should outperform the ideogram-training group in the WPT. However, the availability of verbal labels is not the sole potential facilitator of category learning, as it presupposes both familiarization and individuation, which may be partially responsible for any observed learning benefits. Cue–response training requires the formation of individuated representations for the cues, potentially causing participants to develop perceptual anchors (Ahissar, 2007). Such individuated representations may help stabilize representations in working memory and facilitate executive functions such as hypothesis testing. If this is the case, participants in the ideogram-training condition ought to have an advantage in WPT categorization accuracy relative to the intensity-training group, in which cue identity was instructed to be unattended and varied orthogonally to the intensity task. Finally, mere exposure to the stimulus features has been shown to affect subsequent categorization performance (Folstein, Palmeri, & Gauthier, 2010). We thus predicted that participants in the intensity-training group would outperform the no-training group.

Method

Participants

A group of 85 undergraduate and graduate students (19 male, 66 female; $M_{age} = 25.8$, $SD = 4.05$) of the Philosophy and History of Science Department, University of Athens, Greece, were randomly assigned to one of the three training conditions, receiving course credit for participation, or volunteered. Due to technical failures in collecting the training data or to participants' errors in following instructions, ten of the participants were excluded from the analysis. Thus, the data included 23 participants (seven male, 16 female; $M_{age} = 27.7$, $SD = 4.47$) in the label-training condition, 22 participants (six male, 16 female; $M_{age} = 24.3$, $SD = 2.55$) in the ideogram-training condition, and 30 participants (five male, 25 female; $M_{age} = 25.6$, $SD = 4.35$) in the intensity-training condition. In addition, 20 graduate students (two male, 18 female; $M_{age} = 20.3$, $SD = 3.5$) from the Psychology Department, Panteion University, Athens, Greece, were administered only the WPT (no-training group). All of the participants reported normal hearing and normal or corrected-to-normal vision, no history of neurological illness, and no dyslexia diagnosis.[2]

Materials

Cues Four 300-ms-long frequency-modulated tones, similar to those used by Holt and Lotto (2006), served as cues. The tones were created in Carnegie Mellon University using parameters listed in Table 1. A pilot study employing a two-alternative forced choice intensity discrimination task indicated that high-pitched tones were perceived as being louder than low-pitched tones. Because of the need for them to be used in intensity training, the four tones were

[2] Dyslexia was a concern because it has been linked with impaired learning of audio–visual pairing (Hulme, Goetz, Gooch, Adams, & Snowling, 2007).

Table 1 Carrier and modulation frequencies of the four tones that served as cues

Tone	Carrier frequency (Hz)	Modulation frequency (Hz)
1	790	360
2	1,060	360
3	790	198
4	1,060	198

perceptually equated in intensity. This perceptual equating of the tones (outlined in the online supplement) resulted in the tones' *adjusted* intensity levels, which were used subsequently in the training procedure.

Four intensity levels were additionally created for each tone: The *highest* intensity corresponded to the tones' adjusted levels, whereas the *high*, *low*, and *lowest* levels were created by decrements of 3, 6, and 9 dB down from the adjusted level, respectively. The 3-dB step was determined in pilot experiments that aimed to equate—to the extent possible—training performance in the three conditions.

In the WPT, the original (unadjusted) tones were used in all four groups.

Pseudowords Four Greek pseudowords were created to serve as new names for the tones: namely, σάβης (/'savis/), λίμης (/'limis/), ρήτης (/'ritis/), and δόθης (/'ðoθis/). They were equal in their numbers of letters, syllables, and phonemes, in stress position, and in orthographic typicality (the mean orthographic Levenshtein distance of the 20 nearest neighbors—OLD20—was 2.00 for all of the cues, taking stress into account, and between 2.15 and 2.85, ignoring stress; Protopapas, Tzakosta, Chalamandaris, & Tsiakoulis, 2012; Yarkoni, Balota, & Yap, 2008).

Ideograms Four Chinese characters were selected, on the basis of (a) number of strokes and (b) structure (a single component; Yan, Qiu, Zhu, & Tong, 2010): 豸(U+8C78), 赤 (U+8D64), 辛 (U+8F9B), and 辰 (U+8FB0). To equate perceptual salience, the first character was rotated to the right by 20 deg. A stroke was erased from the fourth character, resulting in seven strokes for each of the final stimuli, which are shown in Fig. 1a.

Colors Three "hard-to-name" colors (RGB: 0x649EA7, 0x583232, 0xBFBC8F) were sampled from the online version of a study used to assess the involvement of language-processing brain regions in a perceptual decision task (Tan et al., 2008; this does not imply that our stimuli were identical to those used in the previous study, due to lack of chromatic calibration). A fourth color (0xFEAD5C) was selected that was also subjectively judged to be hard to name. All of the color stimuli are shown in Fig. 1b.

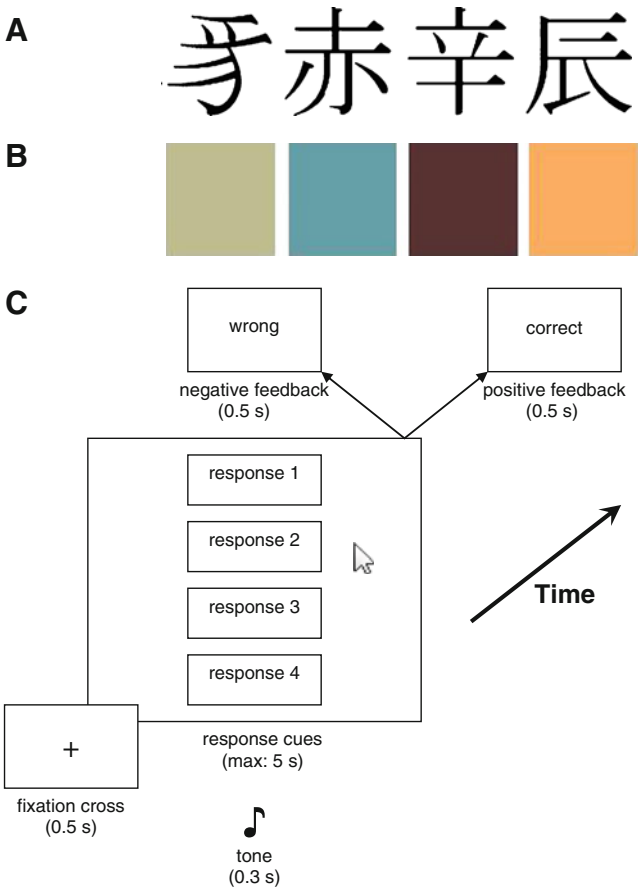

Fig. 1 Training procedure and stimuli. (**a**) Symbols used as response cues for the ideogram-training condition. (**b**) Colors used in the intensity-training condition. (**c**) Sequence of events in a training trial. Responses 1, 2, 3, and 4 are used here to depict the four available response options and were replaced with pseudowords, the stimuli depicted in panel A, and the stimuli depicted in panel B in the label-, ideogram-, and intensity-training conditions, respectively. The ♪ symbol represents the tone cue and was never presented

Procedure

Participants in the training conditions received instructions, a set of headphones, and a questionnaire (in the ideogram- and intensity-training conditions) on or before the first day of training. Moreover, each participants' computer volume was calibrated (see the online supplement for details). Training took place unsupervised at home for three consecutive days. Compliance was monitored daily by e-mail or phone and by inspection of the data. On the fourth day, the WPT was administered at the university lab. Participants used headphones during the tasks.

Training

The training tasks and all following procedures were programmed in DMDX display software (Forster & Forster, 2003). Trial randomization was done with Mix (Van Casteren & Davis, 2006).

Verbal label training In all, 192 trials were presented in each training session. Each cue was presented 12 times in each of four intensity levels. Participants heard one tone in each trial. They were instructed to guess at first, gradually learning the correct response for each tone through corrective feedback. They were explicitly told that the purpose of the task was to learn "a name for each tone" and not just to make the correct response. On the first day of training, they were asked to read aloud the word before responding. The correspondence between sounds and pseudowords was randomly selected for each participant.

The trial structure is shown in Fig. 1c. A cross appeared at the center of the screen for 500 ms. A tone lasting 300 ms followed, simultaneous with the four response options (pseudowords) being presented on the screen in a vertical configuration. On the first day of training, an additional latency period of 500 ms was included after presentation of the tone, during which participants were to pronounce the word. The pseudowords remained on screen for up to 5 s, until a mouse click on one of them. Response feedback was provided for 500 ms ("correct," "wrong," or "no response"). The intertrial interval was 1 s.

The trial order was pseudorandom and fixed for all participants, but different for each day of training. The randomization constraints precluded (a)the same configuration of response cues on two consecutive trials, (b)a lag between trials with the same tone (regardless of intensity) less than 2, and (c)a lag between trials with the same intensity less than 1. A short break occurred halfway through the procedure. Training lasted on average 18 min on the first day, and 15 min on the second and third days. The training tasks were conducted online using the DMDX remote testing mode.

Ideogram training Ideogram training was identical to the verbal-label training, except that (a)four ideograms (randomly paired with tones for each participant) replaced the four pseudowords, (b)participants were instructed to learn the ideogram that corresponded to each of the tones, and (c)no delay to pronounce the labels occurred on the first training day. Participants were instructed to fill in the sealed questionnaire received at the initial meeting on completion of the third day's training. In this questionnaire the four ideograms were printed, and participants were asked to name them using only one word.

Intensity training Participants in intensity training heard the same stimuli as in the other training conditions, but were asked to learn the color that matched each intensity level. They were explicitly instructed to ignore the identity of the tones and only pay attention to intensity. The intensity–color correspondences were randomized across participants. All other aspects of the procedure were the same as in the ideogram-training condition. Following the third day's training, participants were asked to fill in a questionnaire asking for the names of the four colors using one word (as in Sturges & Whitfield, 1995).

WPT

Participants were told that they would take part in a learning experiment and would be asked questions about it at the end. They were not informed of the probabilistic nature of the task. For those in the training conditions, we noted that this was neither a continuation nor a test of their training. Written instructions were presented on the screen (adopted from those of Lagnado et al., 2006). Five practice trials were given before the actual experiment, for familiarization and sound volume adjustment, using animal sounds as cues.

The probabilistic structure of this auditory version of the WPT followed that of Gluck et al. (2002, Exp. 2). As we already noted, each cue was independently associated with an outcome with a fixed probability. This probability can be calculated from Table 2 (as described by Shohamy et al., 2004). For example, Cue 1 is present in patterns H to N, which appeared in 100 out of the 200 trials of the experiment. In these 100 trials, the outcome of "sun" occurred 20 times, and the outcome of "rain" occurred 80 times. Thus, Cue 1 is associated with sun with a probability of 20 ÷ 100= .2, and with rain with a probability of .8. Likewise, it can be calculated that Cues 2, 3, and 4 predicted sun with probabilities of .4, .6, and .8, respectively. Cues 1 and 2 are therefore predictive of sun, Cues 3 and 4 are predictive of rain, and the highly predictive cues of the task are Cues 1 and 4 for sun and rain, respectively. The assignment of tones (Tone 1, Tone 2, etc.) to associative strengths (Cue 1, Cue 2, etc.) was counterbalanced across participants, and the relative position of a tone within a pattern was held constant for a given pattern and a given participant.

In each trial, a series of tones forming a cue pattern were delivered through the headphones sequentially, with an intercue interval of 1 s. Hence, the duration of each pattern ranged from 0.3 s (one-cue pattern) to 2.9 s (three-cue pattern). Following an additional interval of 1 s, two icons representing the outcomes (a sun and a raining cloud) appeared on the screen, for the participant to respond to by pressing the corresponding key on the keyboard. At registration of a response, the correct outcome was presented on screen for 2 s along with feedback: a happy smiley and a high tone (frequency 1000 Hz, duration 0.1 s) for correct selection, or a frowning smiley and a low tone (frequency 500 Hz, duration 0.1 s) when incorrect. If the participant did not respond within 2 s, a "Please respond now" prompt appeared at the bottom of the screen. The trial was terminated if no response was registered within a total of 5 s, counting as "incorrect" for the purpose of analysis. Following Knowlton et al. (1994), a yellow bar on the right side of the screen provided a rough estimate of performance. The intertrial interval was 500 ms. Short breaks were given every 50 trials. The complete sequence of events in a two-cue auditory pattern trial is shown in Fig. 2. The duration of the categorization task was 35 min on average.

Cue naming

Immediately after the WPT, participants were asked to write down which single cue they considered most likely for each outcome (the precise formulation of the questions was based on that of Reber et al., 1996). Participants in the three training conditions were also presented with the four tones again and were asked to denote which tone corresponded to their two previous responses.

Data analysis

The analyses reported below (except for cue naming) employed generalized mixed-effects logistic regression models for binomial distributions (Dixon, 2008) via a logit transformation (Jaeger, 2008), with participants and stimuli (or patterns of auditory stimuli for WPT) as random factors (Baayen, Davidson, & Bates, 2008), fitted with restricted maximum-likelihood estimation using the lme4 package (Bates & Sarkar, 2007) in R (R Development Core Team, 2011). Effect sizes (β) were estimated as log-odds regression coefficients, with zero corresponding to no effect.

Training

The training data were analyzed in terms of correct or erroneous responses.

WPT

Following standard procedure, participants' categorization performance was measured in terms of optimal responding (Knowlton et al., 1994). A response was marked correct if it corresponded to the most likely outcome given the task contingencies, regardless of the actual feedback presented to the participant on that particular trial. For example, throughout the task, trials incorporating Pattern A were marked as correct if and only if the response was "sun." As can be seen in Table 2, Patterns F and I were equally associated with both outcomes; hence, no optimal response could be defined for them. Responses to these patterns (12 trials overall for each participant) were not included in the analysis.

Cue naming

Answers were scored with 1 if participants responded with the tone that was highly predictive of the stated outcome, with .75 for the less predictive tone, .50 and .25 for the tones predictive of the opposite outcome (weakly or strongly, respectively), and 0 for no answering. The cue selection performance was the sum of the two outcomes, ranging from 0 to 2.

Results

Training

Performance increased throughout and across the three days of training, but not all participants exhibited high performance at the end of the third day. To ensure that subsequent categorization performance (on the WPT) would be subject to the trained cue associations, we excluded participants exhibiting low performance (45% or less) in the second half of the third day of training. This included two "nonlearners" in label, two in

Table 2 Pattern and outcome frequencies in the weather prediction task

	Cue Present						
Pattern	1	2	3	4	Sun	Rain	Total
A	0	0	0	1	17	2	19
B	0	0	1	0	7	2	9
C	0	0	1	1	24	2	26
D	0	1	0	0	2	7	9
E	0	1	0	1	10	2	12
F	0	1	1	0	3	3	6
G	0	1	1	1	17	2	19
H	1	0	0	0	2	17	19
I	1	0	0	1	3	3	6
J	1	0	1	0	2	10	12
K	1	0	1	1	5	4	9
L	1	1	0	0	2	24	26
M	1	1	0	1	4	5	9
N	1	1	1	0	2	17	19
Total					100	100	200

1= cue present, 0= cue absent

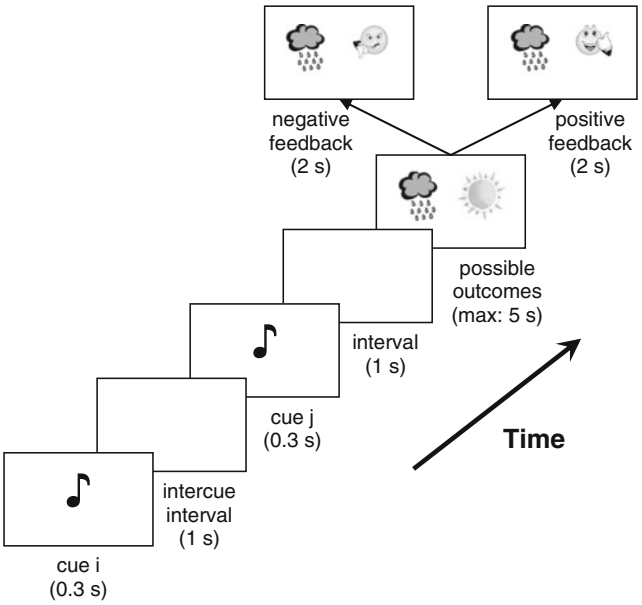

Fig. 2 Sequence of events in a two-cue auditory pattern trial of the weather prediction task (WPT), yielding the "rain" outcome, along with the two possible types of feedback. A "Please respond now!" prompt appeared on screen if the participant did not respond within 2 s of the presentation of the possible outcomes. The ♪ icon represents the tone cue and was never presented

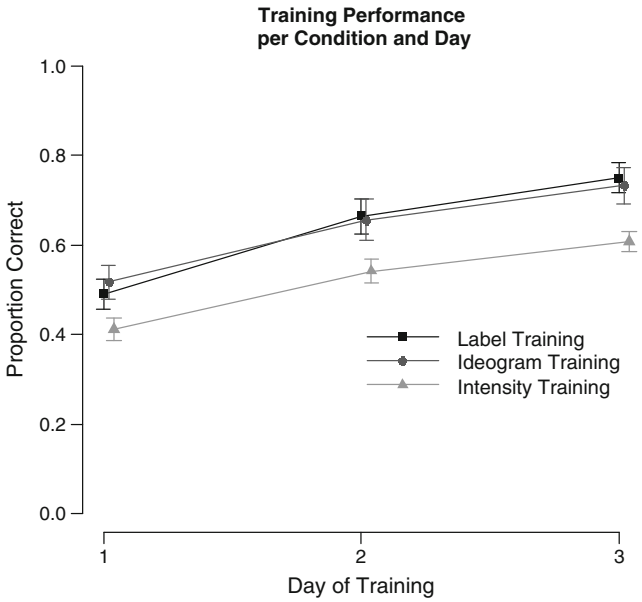

Fig. 3 Mean accuracy of 60 participants (20 in each training condition) in cue–response training. Error bars show between-subjects standard errors of the means

ideogram, and seven in intensity training. Moreover, to equate the sample sizes across conditions, we randomly excluded one participant from the label and three from the intensity condition (see Fig. S1 in the online supplement). The data shown and analyzed henceforth will correspond to the following sample: 20 participants (six male, 14 female; M_{age} = 26.8, SD = 3.47) in the label-training condition, 20 participants (six male, 14 female; M_{age} = 24.4, SD = 2.62) in ideogram training, and 20 participants (four male, 16 female; M_{age} = 25.7, SD = 3.92) in intensity training.

The mean performance in training per condition and day is shown in Fig. 3. Participants' responses were analyzed with a model including fixed effects of trial, training condition, and day of training, as well as their interactions, and random effects of participants and stimuli (4 tones × 4 intensity levels; i.e., 16 distinct stimuli). In R notation, one such model was specified as

```
accuracy ~ trial * condition * day + (1 + trial|participant) + (1|stimulus),
```

with two levels of accuracy ("correct" and "wrong") being regressed onto 192 trials, three levels of condition (intensity, ideogram, and label), and three levels of day. By-participant random slopes of trials were included in order to model participants' individual learning rates (Baayen, 2008); by-stimulus random slopes of trials did not improve the model fit and were excluded. Quadratic effects of trial were not significant and were therefore excluded from the models.

The main purpose of the analysis was to assess whether training resulted in comparable knowledge—by the end of the third day of training—of the cue–response pairings across the three groups. Therefore, the model's intercept was set at the end of training (i.e., the levels of the day predictor were ordered as Day 3, Day 2, and Day 1, and trial was specified numerically as −191, −190, …, −1, 0). The simple effect of condition indicated that the odds of correct responding at the end of Day 3 of training were comparable between the label- and ideogram-training conditions, whereas both of these groups outperformed the intensity-training group (label vs. ideogram, $\beta = -0.180, z = -0.700, p = .484$; label vs. intensity, $\beta = 0.667$, $z = 2.633$, $p = .009$; ideogram vs. intensity, $\beta = 0.847$, $z = 3.332$, $p < .001$; the last two estimates survived Bonferroni correction for three pairwise comparisons). We found a marginal interaction of trial and condition, indicating that changes in correct responding as trials progressed in Day 3 were marginally different between the label and ideogram conditions, but comparable between the other conditions (label vs. intensity, $\beta = -0.001$, $z = -1.241$, $p = .215$; ideogram vs. intensity, $\beta = 0.001, z = 1.086, p = .278$; label vs. ideogram, $\beta = -0.002$, $z = -2.185$, $p = .029$, none of which survived Bonferroni correction for three comparisons). No three-way interaction survived Bonferroni correction for multiple comparisons.[3]

Written responses on the posttraining questionnaire assessing the ideograms' names confirmed that the symbols used were

[3] Analysis of the data at the end of Day 1 indicated increased accuracy of the ideogram-training condition relative to the intensity condition, but comparable accuracy among the other conditions. An analysis of the data at the end of Day 2 indicated higher accuracy of the label-training condition relative to the intensity condition, but comparable accuracy among the other conditions. All of the analyses are available from the authors upon request.

hard to name and did not invoke common associations. The names given were mainly idiosyncratic (such as "air" or "sunset"). A few (six out of 20) of the participants named the ideograms after the sounds that they had been paired with (i.e., they gave names such as "bass" or "shrill").

In contrast, the questionnaire responses regarding colors revealed participants' tendency to give common names to Color 1 ("light blue"—a single word in Greek—by ten participants, "blue" by seven), Color 2 ("brown" by ten), Color 3 ("beige" by eight, "gray" by seven), and Color 4 ("orange" by 15).

WPT

Participants' performance is shown in Fig. 4 in blocks of 10 trials. Participants averaged 74.9% (*SD* = 8.7) optimal responses over all 200 trials in the label-training condition, 71.7% (*SD* = 8.8) in the ideogram condition, 68.5% (*SD* = 12.0) in the intensity condition, and 63.6% (*SD* = 9.0) in the no-training condition.

Responses were analyzed with a model including fixed effects of target (optimal) response, trial, and training condition, as well as their interactions, and random effects of participants and of patterns of auditory cues. In R notation, the model was specified as

```
response ~ target * trial * condition + (1 + trial|participant) + (1|pattern),
```

with two types of response ("sun" and "rain") regressed onto two types of targets ("sun" and "rain"), 188 trials (centered, thus specified numerically as –99.5, –98.5,…, 98.5, 99.5, excluding trials presenting Patterns F and I), and four types of condition (no training, intensity, ideogram, and label); there were also 12 types of pattern (A… N, excluding patterns F and I). By-participant random slopes of trials were included in order to model participants' individual learning rates.

In this model, learning effects would be evident as a significant interaction of trial and target, insofar as increases in trial would increase the probability of responding correctly. This interaction was significant (β = 0.010, z = 7.830, p < .001). A triple interaction including condition would indicate differential learning effects across training conditions; however, this interaction was not significant for any pair of conditions (all βs < 0.002, p > .3).

We observed significant interactions of condition with target, indicating significant performance differences between conditions, in the following order: label> ideogram> intensity> no training. Successive pairwise differences survived Bonferroni correction for three comparisons and were all highly significant (label vs. ideogram, β = 0.351, z = 3.205, p = .001; ideogram vs. intensity, β = 0.341, z = 3.247, p = .001; intensity vs. no training, β = 0.451, z = 4.441, p < .001).[4]

Cue naming

In response to the postcategorization questionnaire, most participants provided verbal descriptions of the tones related to their acoustical features, such as "the high-pitched one" or "the bass sound." In the label condition, 11 out of 20 participants used the trained pseudowords. In the ideogram condition, four participants gave descriptions related to the visual features of the ideograms, such as "the F" or "antenna." None of the participants in the intensity-training condition used a color name to describe the tones.

The mean cue selection scores were 1.79 (*SD* = 0.26) in the label condition, 1.76 (*SD* = 0.25) in the ideogram condition, and 1.58 (*SD* = 0.47) in the intensity condition. A one-way ANOVA revealed no effect of condition, $F(2, 57)$ = 2.328, p = .107, η^2 = .076, suggesting that participants' explicit knowledge of the highly predictive cues did not differ among training conditions.

To assess whether WPT accuracy was affected by explicit knowledge of the newly trained names for the cues as inspected through the postcategorization questionnaire, we analyzed the categorization data from the label-training group only. A modified version of the mixed-effects model included a categorical fixed effect (with two levels, "no" and "yes"), reflecting whether participants used the trained verbal labels in responding to the postcategorization questionnaire. This factor was not significant (β = −0.012, z = −.086, p = .932) and did not interact with the other predictors (all |β|s < 0.003, p > .130).

Correlation between training and categorization performance

Inspection of the individual data revealed participants with high performance during training but low performance in the WPT, and vice versa. To investigate the possibility that cue training was predictive of subsequent categorization, we regressed WPT performance onto the average performance

[4] An analysis of all of the learner participants' data (*N* = *84*) revealed qualitatively the same results—namely, significant performance differences, in the order label> ideogram> intensity> no training (all three pairwise comparisons survived Bonferroni correction). An analysis of both the learner and nonlearner data (*N* = *95*) revealed a similar—but not identical—gradation in performance across the conditions: label> ideogram= intensity> no training (significant differences survived Bonferroni correction for three comparisons). This discrepancy may be attributed to the possibility that some of the seven nonlearner participants in the intensity-training condition were unable to disregard tone identity (as suggested by their informal reports). Thus, including nonlearner data failed to test for the effect of cue individuation when exposure to the stimuli was controlled.

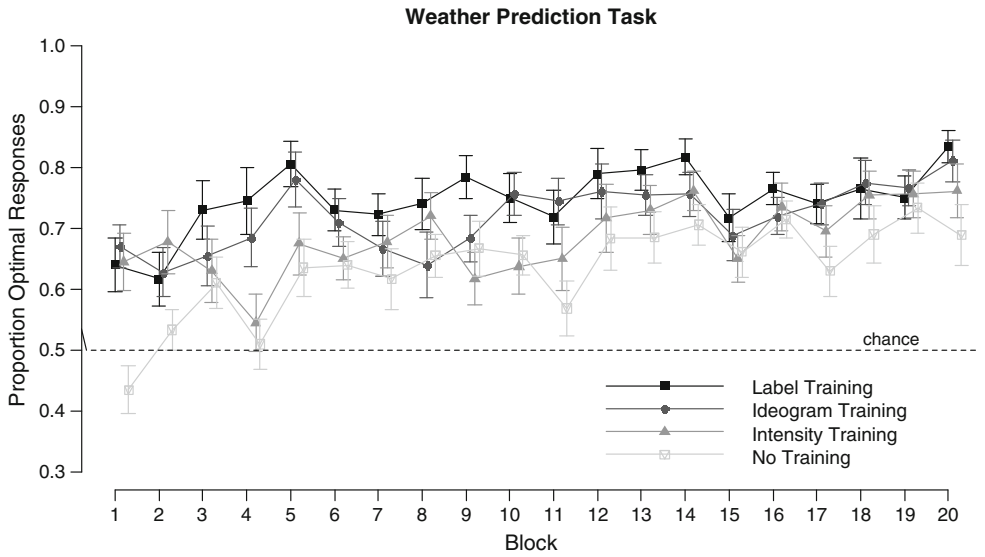

Fig. 4 Posttraining categorization performance of the four training conditions in blocks of ten trials. The dotted line denotes chance performance (50%). Error bars show between-subjects standard errors of the means

in the second half of Day 3 of training, potentially interacting with training condition. No significant effect emerged of either condition or training performance, and no significant interaction (all *p*s > .4). Figure 5 shows the scatterplot and the regression lines for the three training conditions, as well as the regression line for data pooled from all three conditions. To explore the possibility that training performance was predictive of WPT performance depending on the number of cues forming a pattern, we separately calculated average WPT performance on the one-cue, two-cue, and three-cue trials. We regressed each performance measure onto average training performance in the second half of Day 3, possibly interacting with training condition, and again found no effects or interactions in any of these analyses (all *p*s > .2).

Discussion

In this study, participants performed the WPT, a probabilistic category-learning task, using hard-to-name auditory cues. In a training phase preceding the WPT, groups of participants learned to associate the cues to verbal labels or hard-to-name ideograms, or were exposed to the cues in an intensity task orthogonal to cue identity; another group of participants received no training. Categorization performance in the WPT was significantly affected: The label-training group outperformed the ideogram group, the ideogram-training group outperformed the intensity-training group, and the intensity-training group outperformed the no-training group. Since all groups were administered the same auditory version of the WPT, the differences in performance can only be attributed to training. Therefore, (a)the availability of verbal labels, (b)cue individuation, and (c)exposure to the stimuli conferred independent benefits in the category-learning task.

Verbal labels

We assumed that the availability of the cue names would favor the formation, testing, and application of verbalizable strategies by participants in the label-training condition because these participants would have easily accessible names for the cues of the categorization task. To ensure that the availability of names was not confounded with categorical training, verbal label training was contrasted with ideogram training, which differed through the nonverbal nature of its associations. The advantage of the label-training group suggests that cue names specifically enhanced explicit processes mediating WPT performance. The lack of significant differences in learning slopes between conditions further suggests that the naming advantage was not limited to early stages in WPT learning, perhaps serving simply as initial anchors, but extended throughout training. Also, participants' identifications of the highly predictive cues, although they were a poor measure of awareness (see Lagnado et al., 2006, for a trial-by-trial assessment of task knowledge and self insight), suggest that awareness for the learned material was comparable among the training conditions, and thus precludes a potential explanation of the present results on the grounds of differential mediation of distinct memory systems in each condition.

Participants in the ideogram-training group might have developed labels for the cues due to the extended exposure (cf. Galizio & Baron, 1976; Lupyan et al., 2007). Care was taken so that the ideograms would be hard-to-name and that potential labels for the cues would not originate in them. Indeed, the posttraining questionnaire confirmed the unavailability of easily accessible names for the ideograms, and the postcategorization questionnaire showed that very few participants in the ideogram-training condition (four out of 20) gave descriptions of the tones corresponding to the ideograms' features. In contrast, in the label-

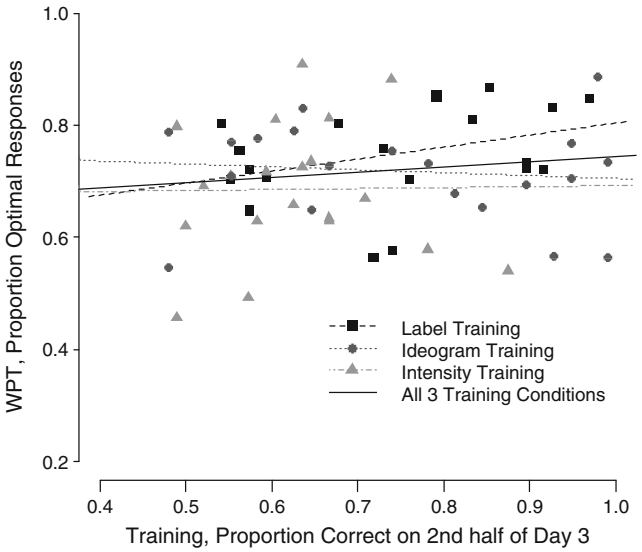

Fig. 5 Scatterplot of weather prediction task (WPT) categorization performance versus training performance on the second half of Day 3. The lines correspond to linear regression parameter estimates for the respective groups

training condition, 11 out of 20 participants used the trained pseudowords to describe the tones, a significantly larger proportion ($\chi^2 = 3.84$, $df = 1$, $p = .05$). Even if labels were developed under ideogram training, the finding that the label group outperformed the ideogram group in the WPT—given equal performance at the end of training—suggests that these purported labels were largely idiosyncratic and ineffective.

It is conceivable that the advantage in categorization of the label relative to the ideogram group might have been due to more efficient encoding of the tones under label training. The difference in encoding efficiency might have resulted in a memory benefit (easier retrieval) when categorizing the tones. The identification of auditory warning sounds has shown more robust learning using verbal than "graphic" labels (Edworthy & Hards, 1999, though with some of their sounds, graphic labels did work better, and further confounds were present in that study). However, there is little reason to assume that auditory–verbal pairings resulted in an encoding advantage in our study, given our finding of equal training performance between the label- and ideogram-training groups at the end of training.

Another possible interpretation of the categorization advantage under label training would be increased perceptual discrimination of the cues (the hypothesis of an "acquired distinctiveness of cues"; Miller & Dollard, 1941, as cited by Galizio & Baron, 1976). However, the equal performance at the end of training in the label and ideogram conditions again argues against such an interpretation. Galizio and Baron suggested that acquired distinctiveness might be manifested with label training only when the task conditions make cues difficult to discriminate. We have no reason to assume that the sequential presentation of the tones—with an interstimulus interval of 1 s—during the WPT imposes perceptual difficulty. Therefore, the acquisition of perceptual features under label training does not seem to offer a strong explanation for our results.

It could be argued that the label and ideogram group trainings differed in ways other than the verbal labels. For example, the Chinese characters might be characterized by greater visual complexity than the printed pseudowords. This difference might not affect training, but only manifest itself in a demanding task such as the WPT. The present design cannot preclude this possibility, which must be explored in further research.

To explore the mechanisms that contributed to the difference in performance between the label- and ideogram-training groups, we considered the possibility that WPT performance was driven by partial cue knowledge.[5] Given the differences in training performance across the participants and tones (e.g., not all participants were equally successful in learning the cue–response pairings for each of the four tones), we calculated each participant's individual cue knowledge—that is, the average performance for each of the four tones in the second half of the third day of training. Subsequently, we constructed a measure of "partial cue knowledge" for each pattern and each participant in the WPT by averaging the participant's cue knowledge for the tones appearing in the pattern. This was only possible for participants in the label- and ideogram-training groups (because participants in intensity training did not classify tones by their identity). The data from the two conditions were reanalyzed with a modified mixed-effects model including partial cue knowledge (centered) as a fixed effect, along with its interactions. We observed a four-way interaction involving target, trial, condition, and partial cue knowledge ($\beta = -0.034$, $z = -3.124$, $p = .002$); hence, the data from the two conditions were analyzed separately. For the label-training group, a positive effect emerged of partial cue knowledge on optimal responding (interaction of partial cue knowledge with target: $\beta = 1.722$, $z = 3.313$, $p < .001$), which did not interact with trial (interaction of partial cue knowledge with trial and target: $\beta = -0.004$, $z = -0.511$, $p = .609$), consistent with a constant influence throughout the WPT. For the ideogram-training group, an interaction of partial cue knowledge with trial and target ($\beta = 0.030$, $z = 4.480$, $p < .001$) suggested a variable effect of partial cue knowledge. Models with an alternative trial centering revealed that partial cue knowledge had a negative effect during the first half of the procedure (e.g., at Trial 50, $\beta = -2.365$, $z = -3.124$, $p = .002$; at Trial 100, $\beta = -0.851$, $z = -2.191$, $p = .029$), no effect later on (at Trial 150, $\beta = 0.633$, $p = .229$), and a positive effect at the end ($\beta = 2.144$, $z = 2.709$, $p = .007$).

[5] We thank an anonymous reviewer for suggesting this analysis.

This post-hoc analysis suggests that participants' categorization accuracy in the label-training group was driven throughout the procedure by partial knowledge of the tone–label pairings. Participants performed better on those WPT trials that employed cues for which labels were learned better during training. This is consistent with the hypothesis that explicit hypothesis-testing processes, mediated by the availability of verbal labels, are recruited during the WPT. Having names for the cues may have facilitated the verbal working memory processes that contribute to category learning (Miles & Minda, 2011). In contrast, knowledge of the tone–ideogram pairings seems to have interfered with WPT performance in the first half of the procedure. Perhaps the visual complexity of the ideograms distracted participants in the demanding WPT, impeding the formation of explicit verbal rules. Further empirical investigation will be needed to study this issue with planned comparisons in an appropriate design.

Cue individuation

The advantage in WPT performance of the ideogram-training relative to the intensity group may be attributed to the individuated representations formed for the tones during ideogram training. These representations, possibly akin to "perceptual anchors" (Ahissar, 2007), may have rendered the tones less abstract in working memory, thus facilitating the use of strategies when solving the WPT. In contrast, the participants in intensity training could perform successfully disregarding tone identity, so the task demands may not have caused the formation of individuated, concrete representations of the tones.

However, the ideogram- and intensity-training groups also differed in training performance, prior to WPT, leaving the WPT difference open to alternative interpretations that cannot be confidently rejected. For example, participants in the intensity-training group may have recruited fewer or less efficient cognitive resources during training. The lower rate of successful performance produced diminished reinforcement—through positive feedback—and may have led to less efficient processing of the auditory tones. Further research with an easier training task will be required to empirically assess this possibility.

The finding that cue individuation alone, in the absence of verbal labels, was beneficial to category learning in the WPT is important to the extent that the latter is primarily mediated by explicit processes, as it highlights the potential of individuated representations to participate flexibly in novel learning tasks. Previous research has suggested that cue characteristics are immaterial to WPT performance, as long as an isomorphic probabilistic structure is present (Hopkins et al., 2004; Knowlton et al., 1994). In contrast, cue individuation seems to affect categorization performance, necessitating an explanation from memory-systems approaches.

Prior exposure

Participants trained to associate sound intensity to colors exhibited greater categorization performance in the WPT than did participants who received no training at all. Notably, the intensity group was able to benefit from training that explicitly required that the relevant dimension for later categorization (cue identity) be disregarded. The critical manipulation in this condition required participants to form intensity "categories" orthogonal to cue identity. Our pilot experiments showed that cue identity interfered with intensity judgments, so there is reason to hypothesize that cue identity and cue intensity are "integral" dimensions (Goldstone, 1994). According to that account, it is possible that sensitization occurred along both dimensions during training, and thus that intensity training enhanced discriminability among the cues (Goldstone). That this manipulation led to increased WPT performance relative to no training therefore suggests that (a)discriminability of the cues may be crucial for their effectiveness in probabilistic category learning and (b)the exposure to stimuli is in itself beneficial for subsequent processing of these stimuli.

The beneficial effect of intensity training was especially apparent early in the WPT, since participants in the no-training condition exhibited near-chance performance in the first two blocks of ten trials (see Fig. 4), perhaps reflecting an initial difficulty with identifying the four tones. Generally, familiarity with the stimulus set is known to affect subsequent performance (e.g., Goldstone & Steyvers, 2001). More specifically, Folstein et al. (2010) exposed participants to artificial stimuli prior to a categorization task utilizing categorizing stimuli that were novel but had a configuration similar to the exposure stimuli. Even when the dimensions of the exposure stimuli were uncorrelated, and thus provided no diagnostic value for later categorization, these participants displayed a clear advantage in categorization performance relative to a group that remained unexposed to the stimuli. Perhaps participants were able to learn the structure of the stimuli, and thus had an advantage in hypothesis testing or resource allocation. In our experiment, participants received feedback for associating sound intensity to colors. However, the relevant dimension for training was absent in later categorization, as in Folstein et al.'s study, allowing an explanation of the beneficial effect of exposure to stimuli in later categorization performance along the same lines.

Concerns and limitations

It is notable that average performance on the second half of Day 3 of training was not correlated with average WPT categorization performance for any of the training conditions. This may be interpreted as supporting the existence of discrete learning systems: Training required a gradual acquisition of cue–response pairings, whereas the WPT

presumably required explicit hypothesis testing. At the moment, the differences in task demands between the training and categorization tasks in our study do not allow us to draw firm conclusions in this matter (cf. Dunn & Kirsner, 2003). On the other hand, a more refined, by-cue measure of training performance was found to be predictive of between-trials differences in WPT performance. Partial knowledge of the cue–label pairings acquired during training was found to facilitate posttraining categorization, whereas partial knowledge of the cue–ideogram pairings initially interfered with, and later facilitated, categorization. This connection between training and categorization provides no evidence in favor of a multiple-systems account.

The observed differences in training performance between the groups may cause some concern regarding the interpretations. Verbal-label- and ideogram-training performance did not differ at the end of training, yet participants in the label-training group probably achieved plateau performance (as evidenced by the lack of an effect of trial on Day 3) earlier than did the ideogram-training group (which kept on learning the cue–response pairings during Day 3, as evidenced by an effect of trial). We believe that this discrepancy between the two conditions does not pose a significant limitation on the interpretation of our results, insofar as both groups' knowledge of the cue–response pairings was comparable at the end of the training procedure.

Another concern stems from the fact that the ideogram group outperformed the intensity group in training performance. Although similar performance in all three training conditions was desirable, the intensity-training condition was primarily designed to equate exposure to the stimulus set and recruitment of attentional resources. The design constraint that tone identity be disregarded led to a significant difference in training performance at the end of training, leaving our results regarding individuation open to alternative interpretations.

Finally, we acknowledge that care should taken when interpreting the difference in WPT performance between the intensity and no-training groups. Participants in these conditions were—due to recruiting difficulties—sampled from different pools, and hence no strong inferences can be made. This confound does not undermine the comparison of prime interest in our study—that is, between label and ideogram training.

Implications and conclusion

This has been the first detailed report of gradual learning in an auditory version of the WPT. Two procedural discrepancies between this version and the prototypical task (Knowlton et al., 1994) were imposed by the auditory nature of the cues: First, the cues were presented sequentially, and second, feedback was delivered in the absence of the cues. There is evidence that both of these factors modulate the involvement of distinct memory systems during visual category learning (Foerde & Shohamy, 2011; Maddox, Ashby, & Bohil, 2003; Maddox & Ing, 2005; Worthy, Markman, & Maddox, 2013; but see Dunn, Newell, & Kalish, 2012, for an alternative interpretation). However, it has been demonstrated that, for learning to take place, the appropriate mode of presentation for auditory stimuli is sequential and not concurrent, as in the visual modality (Conway & Christiansen, 2009; Saffran, 2002). Further research will be required in order to examine whether sequential presentation resulted in different memory-system involvement relative to the prototypical WPT. Importantly, all our participants were administered the exact same version of the WPT. Therefore, the procedural discrepancies between this auditory version and the prototypical WPT do not undermine the between-groups comparison that supports the idea that verbal labels facilitate explicit hypothesis testing.

The WPT has been used extensively as a tool by multiple-systems (e.g., Knowlton et al., 1996; Poldrack & Foerde, 2008) and single-system theorists (e.g., Newell et al., 2011) to assess the existence and relative contributions of discrete memory systems during categorization learning. It has been suggested that the majority of young, healthy participants (Gluck et al., 2002; Poldrack & Foerde, 2008) initially approach the task via suboptimal strategies that can be said to be declarative (Shohamy et al., 2008), but later on engage multiple-cue (or integrative) strategies. These later strategies may be mediated by the procedural system (Shohamy et al., 2008), or they may be supported by declarative learning processes, since they are accompanied by high levels of awareness (Price, 2009) or self-insight (Lagnado et al., 2006; Newell et al., 2007). Our results are consistent with the latter assumption. If the WPT is mediated by a procedural system and not by explicit hypothesis testing later in training, then having names for the cues should not affect later categorization performance. The fact that the label-training group outperformed the ideogram-training group throughout the task suggests that the declarative–procedural distinction does not explain healthy participants' behavior in the WPT. Instead, a general learning mechanism may support performance throughout the task (Newell et al., 2007).

To conclude, we have shown that newly trained verbal labels for the cues provide an advantage in probabilistic category-learning performance. We based our hypothesis on the assumption that explicit hypothesis testing of verbal rules would be facilitated when participants had names for the cues, as opposed to associating the cues to difficult-to-name ideograms. The present results extend previous studies that have suggested that language is not just for talking (Lupyan, 2008; Lupyan et al., 2007) and that verbal processes are important for categorization (Ashby & Maddox, 2005; Miles & Minda, 2011). Future research should examine in

more detail the intuitive (but perhaps simplistic; see Newell et al., 2011) notion that humans may benefit from linguistic faculties during categorization, with a new focus on verbal labels for categorizing items.

Author note We thank George Gyftodimos for suggesting the idea of new names for the auditory cues, Eleni Vlahou for help with DMDX programming and preprocessing of the data, Lori Holt and Sung-joo Lim for providing the frequency-modulated tones, Martijn Meeter for providing the order of the trials in the WPT and for comments on an earlier draft, Catherine Myers for help with the literature, Maarten Van Casteren for adapting the MIX program to facilitate advanced trial randomization, Jonathan C. Forster for technical advice on DMDX remote-testing mode, and all of the ling-r-lang-L mailing list subscribers (especially Florian Jaeger) for advice on the statistical analyses. We also thank Argiro Vatakis and all of the members of the Language and Learning Lab at the University of Athens for help with recruiting participants.

References

Abu-Shaba, N., Myers, C. E., Shohamy, D., & Gluck, M. A. (2001). *Age effects in probabilistic category learning. Unpublished manuscript*. Newark: Rutgers University.

Ahissar, M. (2007). Dyslexia and the anchoring-deficit hypothesis. *Trends in Cognitive Sciences, 11,* 458–465. doi:10.1016/j.tics.2007.08.015

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442–481. doi:10.1037/0033-295X.105.3.442

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149–178. doi:10.1146/annurev.psych.56.091103.070217

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences, 1224,* 147–161. doi:10.1111/j.1749-6632.2010.05874.x

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. doi:10.1016/j.jml.2007.12.005

Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes (R package version 0.99875-6). Retrieved from http://cran.rproject.org/web/packages/lme4/index.html

Conway, C. M., & Christiansen, M. H. (2009). Seeing and hearing in space and time: Effects of modality and presentation rate on implicit statistical learning. *European Journal of Cognitive Psychology, 21,* 561–580. doi:10.1080/09541440802097951

Development Core Team, R. (2011). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from www.R-project.org.

Dickerson, K. C., Li, J., & Delgado, M. R. (2011). Parallel contributions of distinct human memory systems during probabilistic learning. *NeuroImage, 55,* 266–276. doi:10.1016/j.neuroimage.2010.10.080

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59,* 447–456. doi:10.1016/j.jml.2007.11.004

Dunn, J. C., & Kirsner, K. (2003). What can we infer from double dissociations? *Cortex, 39,* 1–7. doi:10.1016/S0010-9452(08)70070-4

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 840–859. doi:10.1037/a0027867

Edworthy, J., & Hards, R. (1999). Learning auditory warnings: The effects of sound type verbal labelling and imagery on the identification of alarm sounds. *International Journal of Industrial Ergonomics, 24,* 603–618.

Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Science, 103,* 11778–11783. doi:10.1073/pnas.0602659103

Foerde, K., Poldrack, R. A., & Knowlton, B. J. (2007). Secondary task effects on classification learning. *Memory & Cognition, 35,* 864–874. doi:10.3758/BF03193461

Foerde, K., & Shohamy, D. (2011). Feedback timing modulates brain systems for learning in humans. *Journal of Neuroscience, 31,* 13157–13167. doi:10.1523/JNEUROSCI.2701-11.2011

Folstein, J., Palmeri, T. J., & Gauthier, I. (2010). Mere exposure alters category learning of novel objects. *Frontiers in Psychology, 1,* 1–6. doi:10.3389/fpsyg.2010.00040

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35,* 116–124. doi:10.3758/BF03195503

Fotiadis, F. A., Protopapas, A., & Vatakis, A. (2011). The effect of cue naming in probabilistic category learning. In B. Kokinov, A. Karmiloff-Smith, & N. J. Nersessian (Eds.), *European perspectives on cognitive science: Proceedings of the European Conference on Cognitive Science—EuroCogSci 2011*. Sofia: New Bulgarian University Press.

Galizio, M., & Baron, A. (1976). Label training and auditory categorization. *Learning and Motivation, 7,* 591–602. doi:10.1016/0023-9690(76)90009-6

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology. General, 117,* 227–247. doi:10.1037/0096-3445.117.3.227

Gluck, M. A., Shohamy, D., & Myers, C. E. (2002). How do people solve the "weather prediction" task? Individual variability in strategies for probabilistic classification learning. *Learning and Memory, 9,* 408–418. doi:10.1101/lm.45202

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology. General, 123,* 178–200. doi:10.1037/0096-3445.123.2.178

Goldstone, R. L., & Steyvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology. General, 130,* 116–139. doi:10.1037110096-3445.130.1.116

Holt, L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implication for first and second language acquisition. *Journal of the Acoustical Society of America, 119,* 3059–3071. doi:10.1121/1.2188377

Hopkins, R. O., Myers, C. E., Shohamy, D., Grossman, S., & Gluck, M. A. (2004). Impaired probabilistic category learning in hypoxic subjects with hippocampal damage. *Neuropsychologia, 42,* 524–535. doi:10.1016/j.neuropsychologia.2003.09.005

Hulme, C., Goetz, K., Gooch, D., Adams, J., & Snowling, M. J. (2007). Paired-associate learning, phoneme awareness, and learning to read. *Journal of Experimental Child Psychology, 96,* 150–166. doi:10.1016/j.jecp.2006.09.002

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446. doi:10.1016/j.jml.2007.11.007

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science, 273,* 1399–1402. doi:10.1126/science.273.5280.1399

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning and Memory, 1,* 106–120. doi:10.1101/lm.1.2.106

Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology. General, 135,* 162–183. doi:10.1037/0096-3445.135.2.162

Lupyan, G. (2006). Labels facilitate learning of novel categories. In A. Cangelosi, A. D. M. Smith, & K. Smith (Eds.), *The Sixth International Conference on the Evolution of Language* (pp. 190–197). Singapore: World Scientific.

Lupyan, G. (2008). From chair to "chair": A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology. General, 137,* 348–369. doi:10.1037/0096-3445.137.2.348

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science, 18,* 1077–1083. doi:10.1111/j.1467-9280.2007.02028.x

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 650–662. doi:10.1037/0278-7393.29.4.650

Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 100–107. doi:10.1037/0278- 7393.31.1.100

Meeter, M., Myers, C. E., Shohamy, D., Hopkins, R. O., & Gluck, M. A. (2006). Strategies in probabilistic categorization: Results from a new way of analyzing performance. *Learning and Memory, 13,* 230–239. doi:10.1101/lm.43006

Meeter, M., Radics, G., Myers, C. E., Gluck, M. A., & Hopkins, R. O. (2008). Probabilistic categorization: How do normal participants and amnesic patients do it? *Neuroscience and Biobehavioural Reviews, 32,* 237–248. doi:10.1016/j.neubiorev.2007.11.001

Miles, S. J., & Minda, J. P. (2011). The effect of concurrent verbal and visual tasks on category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 588–607. doi:10.1037/a0022309

Minda, J. P., & Miles, S. J. (2010). The influence of verbal and nonverbal processing on category learning. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 52, pp. 117–162). San Diego: Academic Press. doi:10.1016/S0079-7421(10)52003-6

Newell, B. R., & Dunn, J. C. (2008). Dimensions in data: Testing psychological models using state-trace analysis. *Trends in Cognitive Sciences, 12,* 285–290. doi:10.1016/j.tics.2008.04.009

Newell, B. R., Dunn, J. C., & Kalish, M. (2010). The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition, 38,* 563–581. doi:10.3758/MC.38.5.563

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning: Fact or fantasy? In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 54, pp. 167–215). San Diego: Academic Press. doi:10.1016/B978-0-12-385527-5.00006-1

Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2007). Challenging the role of implicit processes in probabilistic category learning. *Psychonomic Bulletin and Review, 14,* 505–511. doi:10.3758/BF03194098

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso, J., Myers, C. E., & Gluck, M. A. (2001). Interactive memory systems in the human brain. *Nature, 414,* 546–550. doi:10.1038/35107080

Poldrack, R. A., & Foerde, K. (2008). Category learning and the memory system debate. *Neuroscience and Biobehavioral Reviews, 32,* 197–205. doi:10.1016/j.neubiorev.2007.07.007

Poldrack, R. A., & Rodriguez, P. (2004). How do memory systems interact? Evidence from human classification learning. *Neurobiology of Learning and Memory, 82,* 324–332. doi:10.1016/j.nlm.2004.05.003

Price, A. L. (2009). Distinguishing the contributions of implicit and explicit processes to performance of the weather prediction task. *Memory & Cognition, 37,* 210–222. doi:10.3758/MC.37.2.210

Protopapas, A., Tzakosta, M., Chalamandaris, A., & Tsiakoulis, P. (2012). IPLR: An online resource for Greek word-level and sublexical information. *Language Resources and Evaluation, 46,* 449–459. doi:10.1007/s10579-010-9130-z

Reber, P. J., Knowlton, B. J., & Squire, L. R. (1996). Dissociable properties of memory systems: Differences in the flexibility of declarative and nondeclarative knowledge. *Behavioral Neuroscience, 110,* 861–871. doi:10.1037/0735-7044.110.5.861

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 585–594. doi:10.1037/0278-7393.20.3.585

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language, 47,* 172–196. doi:10.1006/jmla.2001.2839

Shohamy, D., Myers, C. E., Kalanithi, J., & Gluck, M. A. (2008). Basal ganglia and dopamine contributions to probabilistic category learning. *Neuroscience and Biobehavioral Reviews, 32,* 219–236. doi:10.1016/j.neubiorev.2007.07.008

Shohamy, D., Myers, C. E., Onlaor, S., & Gluck, M. A. (2004). Role of the basal ganglia in category learning: How do patients with Parkinson's disease learn? *Behavioral Neuroscience, 118,* 676–686. doi:10.1037/0735-7044.118.4.676

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory, 82,* 171–177. doi:10.1016/j.nlm.2004.06.005

Sturges, J., & Whitfield, T. W. A. (1995). Locating basic colours in the Munsell space. *Color Research and Application, 20,* 364–376. doi:10.1002/col.5080200605

Tan, L. H., Chan, A. H. D., Kay, P., Khong, P.-L., Yip, L. K. C., & Luke, K.-K. (2008). Language affects patterns of brain activation associated with perceptual decision. *Proceedings of the National Academy of Sciences, 105,* 4004–4009. doi:10.1073/pnas.0800055105

Van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods, 38,* 584–589. doi:10.3758/BF03193889

Worthy, D. A., Markman, A. B., & Maddox, W. T. (2013). Feedback and stimulus-offset timing effects in perceptual category learning. *Brain and Cognition, 81,* 283–293. doi:10.1016/j.bandc.2012.11.006

Yan, J., Qiu, Y., Zhu, Y., & Tong, S. (2010). Mental rotation differences between Chinese characters and English letters. *Neuroscience Letters, 479,* 146–151. doi:10.1016/j.neulet.2010.05.051

Yarkoni, T., Balota, D. A., & Yap, M. J. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin and Review, 15,* 971–979. doi:10.3758/PBR.15.5.971