Psychon Bull Rev
DOI 10.3758/s13423-014-0612-7

BRIEF REPORT

Categorical congruence facilitates multisensory associative learning

Elan Barenholtz · David J. Lewkowicz · Meredith Davidson · Lauren Mavica

Abstract Learning about objects often requires making arbitrary associations among multisensory properties, such as the taste and appearance of a food or the face and voice of a person. However, the multisensory properties of individual objects usually are statistically constrained, such that some properties are more likely to co-occur than others, on the basis of their category. For example, male faces are more likely to co-occur with characteristically male voices than with female voices. Here, we report evidence that these natural multisensory statistics play a critical role in the learning of novel, arbitrary associative pairs. In Experiment 1, we found that learning of pairs consisting of human voices and gender-congruent faces was superior to learning of pairs consisting of human voices and gender-incongruent faces or of pairs consisting of human voices and pictures of inanimate objects (plants and rocks). In Experiment 2, we found that this "categorical congruency" advantage extended to nonhuman stimuli, as well—namely, to pairs of class-congruent animal pictures and vocalizations (e.g., dogs and barks) versus class-incongruent pairs (e.g., dogs and bird chirps). These findings suggest that associating multisensory properties that are statistically consistent with the various objects that we encounter in our daily lives is a privileged form of learning.

Keywords Human associative learning · Multisensory associative learning · Face recognition · Voice recognition

Electronic supplementary material The online version of this article (doi:10.3758/s13423-014-0612-7) contains supplementary material, which is available to authorized users.

E. Barenholtz (✉) · D. J. Lewkowicz · M. Davidson · L. Mavica
Department of Psychology, Florida Atlantic University,
Boca Raton, FL 34331, USA
e-mail: elan.barenholtz@fau.edu

E. Barenholtz · D. J. Lewkowicz
Center for Complex Systems and Brain Sciences, Florida Atlantic University, Boca Raton, FL, USA

Published online: 27 March 2014

Many objects can be identified on the basis of multiple properties from different modalities. For example, certain foods can be identified on the basis of their auditory, visual, gustatory, tactile, and olfactory properties. Similarly, "knowing" a person typically means being able to associate that person's face and voice. Obtaining such object knowledge often requires us to make arbitrary associations of multisensory stimulus properties. Of course, the properties of individual objects tend to be constrained, such that not all properties are equally likely to be encountered together. For example, male faces and low-pitched voices are associated more frequently than male faces and high-pitched voices. How might prior exposure to these natural statistics affect our ability to learn novel associations among properties?

Previous studies of word-list learning have provided some evidence that previously learned relations among different items can facilitate learning and memory. Heim, Watts, Bower, and Hawton (1966) found that the recall of word pairings is superior in a cued-associate task when the pairs consist of words that are judged to be highly semantically related, such as "bank" and "money," relative to when they consist of less related words (although note that Cofer, 1968, failed to find this effect). Similarly, Kintsch (1968) found that lists of words that can be organized into categories are recalled better than words that cannot. Importantly, the participants in these early studies had preexisting knowledge of specific word associations, which they could have used to constrain the possible set of responses or, in the case of list learning, to generate retrieval cues. Consequently, the learning principles uncovered in these studies are not applicable to situations in which we have to learn to associate attributes that have not been experienced before, such as learning the face and voice of someone we meet for the first time. In this case, we only have access to general, a priori, category-based expectations that tell us which types of faces and voices typically belong together (e.g., gender-congruent faces and voices). To date, the role of categorical congruency on the learning of novel

associations has not been investigated. Here, we compared the learning of novel visual and auditory stimulus pairs (e.g., faces and voices) that were either consistent with belonging to a single identity based on their category (e.g., same gender) or not (e.g., opposite gender).

Critically, unlike the word-learning studies discussed above, the congruency condition in our study did not involve previously associated items, and thus did not provide any information that could be used directly to perform the learning task. For example, knowing that several face–voice pairs are all the same gender does not provide any information for remembering *which* particular congruent faces and voices belong together. We predicted, however, that categorical congruency would facilitate learning. This was because learning of categorically congruent pairs, even when they are novel, is generally more consistent with previous experience. Moreover, it is possible that encoding associations among congruent multisensory properties may be mediated by a specialized learning mechanism: the formation of a supramodal *identity*. This idea is consistent with an influential theory of person identification, which holds that the formation of face–voice associations corresponding to a particular person depends on the integration of distinct informational streams via a single "personal identity node," or PIN (Bruce & Young, 1986; Burton et al. 1990; Ellis et al. 1997). According to this view, associations of the properties indexing a single individual lead to the formation of a consolidated network in which activation of one modality "redintegrates" all of the other information associated with that individual (Thelen & Murray, 2013). Indeed, neurophysiological studies support this view, showing that face and voice areas are "functionally coupled" during associations, such that the presentation of voices previously paired with faces activates facial recognition regions of the brain (von Kriegstein & Giraud, 2006; von Kriegstein et al. 2005).

These studies suggest that learning the properties of a unified identity may be mediated by a distinct process that results in the formation of a consolidated network. If so, learning to associate categorically congruent features—which are more consistent with belonging to a unified identity—may be more efficient than learning to associate categorically incongruent properties. To test this prediction, we conducted two experiments that compared adult participants' abilities to learn categorically congruent and incongruent auditory and visual (A–V) pairs in a supervised learning task, using a between-subjects design. In Experiment 1, the participants learned to associate human voices with (a)human faces that were the same gender (congruent condition), (b)human faces that were of the opposite gender (incongruent condition), and (c)pictures of inanimate objects (neutral condition). In the congruent condition, the A–V pairs were consistent with belonging to a single object/identity, whereas in the other two conditions they were not. Furthermore, in the incongruent condition, the paired stimuli possessed *conflicting* visual and auditory characteristics (i.e., opposite-gender faces and voices), whereas in the neutral condition, the auditory characteristics did not conflict with the visual stimuli because the latter were inanimate objects. This allowed us to assess whether any observed advantage in the congruent versus the incongruent condition was due to some form of interference in the incongruent condition. If so, no similar advantage should be present for the congruent versus the neutral condition, because the stimuli in the latter condition did not possess any conflicting auditory characteristics. If, however, the congruency advantage derives from the fact that stimuli are consistent with belonging to a single object/identity per se, then a similar advantage should be present in the congruent condition, relative to the other two conditions.

Experiment 1

In Experiment 1, we tested the ability to learn the association between individual human voices with different classes of visual stimuli across three between-subjects conditions. In the *congruent* condition, each voice was paired with a static picture of a single, same-gender face. In the *incongruent* condition, each voice was paired with a static picture of a single face of the opposite gender. Finally, in the *neutral* condition, each voice was paired with a static picture of a single inanimate object, either a plant or a rock. These categories of inanimate object were chosen because, like faces, they contain members that share visual characteristics but can be readily distinguished. In all three conditions, participants were trained on a total of 16 pairings repeated across six blocks. Each pairing consisted of a unique person's voice speaking a sentence paired with a single visual stimulus. Participants performed an unspeeded, four-alternative forced choice task in which they chose, on each trial, which among four visual stimuli had been paired with the voice, and received feedback regarding their choices.

Method

Participants A group of 64 Florida Atlantic University undergraduate psychology students, who were naïve to the purposes of the experiment, participated for course credit. All participants were screened after the experiment and asked whether they personally knew any of the people whose faces/voices were presented during the experiment, and if they did, their results were not included in the analysis. Four participants were rejected from the analysis on this basis, but they were replaced with an additional four participants to achieve equal numbers of participants (20) across conditions, for a total of 60 participants. The final data set included 33 female and 27 male participants, whose ages ranged from 18 to 32 years, with an average age of 22.

Stimuli The stimuli in the congruent and incongruent conditions consisted of photographs and voice recordings of eight Caucasian females and eight Caucasian males, ranging from 18 to 26 years of age. Each individual was photographed and recorded speaking the sentence "There are clouds in the sky" in an emotionally neutral tone. The visual stimuli in the neutral condition consisted of photographs of eight visually distinct rocks and eight visually distinct houseplants.[1] Before the beginning of the experiment, each of the 16 visual images was matched with a single recorded voice as the "pair" to be learned by the participant throughout the entire experiment. In the congruent condition, each picture was uniquely paired with one randomly chosen voice of the same gender, with the constraint that it not be the true matching voice (i.e., the paired face and voice always derived from different models). This eliminated any possibility that participants could use properties of the faces and voices to correctly guess which ones went together (Mavica & Barenholtz, 2013) rather than learning on the basis of feedback. In the incongruent condition, each of the female faces was paired with a single randomly chosen male voice, and vice versa. In the neutral condition, for half of the participants female voices were paired with pictures of plants and male voices were paired with pictures of rocks, whereas for the other half of the participants, the pairings were reversed.

Design and procedure The experimental procedures were identical in the three conditions. Participants were instructed that they would be performing a task in which they must learn to match recordings of voices with pictures and that that they would receive feedback on correct or incorrect responses. In the incongruent condition, participants were additionally informed that the voices and faces would be of opposite genders (in order to avoid confusion). Figure 1 and its accompanying caption show the structure of a single block of trials in the congruent or the incongruent condition. During each trial, participants were presented with a voice recording of the spoken sentence while four different visual stimuli were presented simultaneously on the screen. Participants had to select one of the four faces as the (arbitrarily determined) "match" to the spoken sentence, and were then given feedback on the correct pairing. Participants were trained on a total of 16 face–voice pairs, all of which were repeated across six experimental blocks, for a total of 96 trials per participant. (As is described in the supplemental materials, section S1, each participant also ran in two additional experimental blocks that tested their ability to match the faces and voices from the six learning blocks, on the basis of novel utterances of the same voices, in the absence of feedback.)

[1] Section S2 of the supplemental materials reports an additional test showing that participants could discriminate the plant and rock stimuli with nearly perfect accuracy.

Results

Figure 2 shows the proportions of trials on which participants responded correctly across the six learning blocks in the three conditions. Overall performance was better in the congruent condition (M = 59%, SD = 11%) than in the incongruent condition (M = 41%, SD = 8%) and the neutral condition (M = 44%, SD = 9%). A one-way analysis of variance (ANOVA), with Condition as the between-subjects factor, revealed a significant effect of condition, $F(2, 57) = 15.33$, $p < .001$, $\eta^2 = .35$. Post-hoc analysis showed that performance was significantly better in the congruent condition than in both the incongruent and neutral conditions ($p < .001$), but that the incongruent and neutral conditions did not differ significantly from each other ($p > .1$).

To assess the linear *rate* of learning across the six blocks in the three conditions, we conducted a polynomial contrast analysis across the six learning blocks for the three conditions. This yielded a significant overall linear increase in performance as a function of block number, $F(1, 57) = 126.15$, $p < .001$. Separate paired comparisons showed an interaction in the linear components of block number and condition for the congruent versus the incongruent condition, $F(1, 38) = 5.89$, $p = .016$, as well as between the congruent and neutral conditions, $F(1, 38) = 4.40$, $p = .033$. These interaction effects indicate that the block-wise linear learning slope was significantly steeper in the congruent condition than in the other two conditions. We observed no similar interaction between the incongruent and neutral conditions, $F(1, 38) = 0.44$, $p = .511$. Finally, no significant higher-order (i.e., nonlinear) effects emerged for block number or for the Block Number × Condition interaction.

Although learning increased more rapidly in the congruent condition across blocks, a separate ANOVA showed that a significant difference in performance started in the very first block of trials, $F(2, 57) = 4.48$, $p = .016$, $\eta^2 = .36$. Post-hoc analysis with Bonferroni-adjusted p values showed a significant difference between the congruent condition (M = 43%, SD = 10%) and both the incongruent (M = 33%, SD = 13%, $p = .028$) and neutral (M = 34%, SD = 12%, $p = .049$) conditions, but not between the incongruent and neutral conditions ($p = 1.00$).[2] At first blush, this congruency advantage may seem surprising, given that the face–voice pairings were arbitrary. However, the opportunity for learning was available right from the first block of trials. This is because the face–voice pairs were presented in consecutive series, with the same four voices and their corresponding faces always being presented in four consecutive trials. This meant that once participants had received feedback in the first trial of each

[2] As per SPSS analysis software convention, Bonferroni-corrected significance values of 1 or greater are reported as $p = 1.00$.

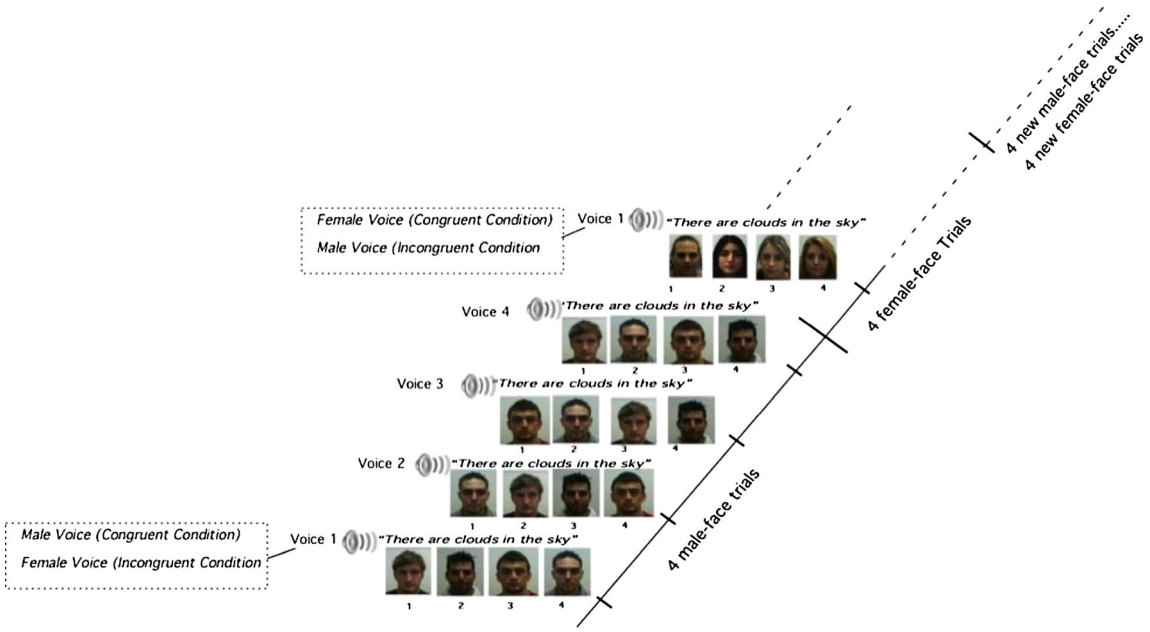

Fig. 1 Schematic of a single experimental block in the congruent or incongruent condition of Experiment 1 (see the text for details). The neutral condition is not shown. On each trial, four faces were presented with the numbers 1–4 in sequence below them. The spatial arrangement/numbers of the visual stimuli varied randomly across trials. One of the four visual stimuli was the "match" to the voice, as was determined prior to the experiment and described above, whereas the other three served as distractors. The participants were instructed to choose, by number, which of the four visual stimuli matched the voice. The correct selection was flashed once—regardless of whether or not participants had chosen it—before the stimuli were replaced by a white screen. An incorrect response resulted in a low beeping sound. Within a block, the same four faces were always shown together as a group (in a random spatial order) across four consecutive trials until each of the four "matching" voices had been presented. All four groups were repeated across six separate experimental blocks

four-person series, the subsequent pairings could be guessed at a probability level of better than chance (which was 25%), because previously learned pairs could be eliminated from consideration as the trials progressed. Categorical consistency probably facilitated learning from the beginning of the experiment by allowing participants to use their memory of previously presented pairings to eliminate future choices.[3]

Overall, we found that learning of arbitrary pairings of multisensory properties is much more efficient when the members of the pair are categorically congruent than when they are incongruent or neutral. We also found that performance in the incongruent and neutral conditions was statistically equivalent, indicating that the presence of a conflicting auditory expectation did not negatively affect performance in the incongruent condition. Instead, the results suggest that performance was better in the congruent condition than in both of the other conditions because the stimuli were consistent with belonging to a single object/individual.

Experiment 2

Is the categorical congruency learning advantage observed in Experiment 1 restricted to human voices and faces? Processing of human voices and faces is thought to rely on specialized mechanisms that depend on dedicated brain regions and/or visual expertise (Gauthier et al. 2000; Gauthier & Tarr, 1997). Indeed von Kriegstein and Giraud's (2006) finding (discussed above) of functional coupling of visual and auditory areas applied *only* to human faces and voices, not to other stimuli such as cell phone pictures paired with ringtones.

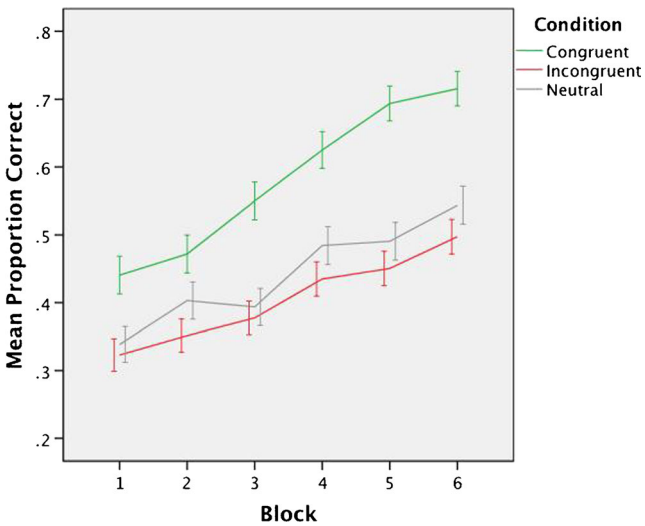

Fig. 2 Mean performance across the six blocks in the three experimental conditions of Experiment 1. Error bars represent ±1 standard error of the mean

[3] Additional within-block analyses are provided in the supplemental materials, section S3.

Thus, associating congruent human face–voice pairs may engage specialized/expertise mechanisms that do not apply to other categories. In addition, whereas the face–voice pairs in Experiment 1 were arbitrary and nonveridical, participants might still have *tried* to use their expectations about facial/vocal mappings in the congruent condition to learn the pairings, potentially affecting performance.

Thus, in Experiment 2 we investigated whether a congruency advantage would emerge for nonhuman visual–auditory properties. We used the same methodology as in Experiment 1, except that here we presented vocalizations and pictures of dogs and birds and compared learning of class-congruent pairs (e.g., a specific bark with a picture of a specific dog) with class-incongruent pairs (e.g., a specific bird song with a specific dog picture). Besides bird-watchers and dog trainers/breeders, most people have limited experience learning specific auditory–visual pairings of such stimuli.

Method

Participants A group of 50 undergraduate psychology students (25 assigned to each of the two experimental conditions), who were naïve to the purposes of the experiment, participated for course credit. Twenty-one of the participants were female, and the participants' ages ranged from 18–29 years, with an average age of 21. None of the participants self-identified as a dog or bird expert.

Stimuli and procedure The stimuli consisted of sound recordings of eight different mid-range dog barks and eight different bird chirps, along with pictures of the cropped faces of eight mid-sized dogs (chosen on the basis of subjective judgments of size) and pictures of eight mid-sized birds (the audio recordings and photos were obtained from the Internet). Each sound recording was approximately 3 s long, which is similar in length to the human voice recordings used in Experiment 1. Each randomly chosen sound recording was uniquely paired with a picture from the same animal category, in the congruent condition, or with a picture from the other animal category, in the incongruent condition (Fig. 3).

The procedure was identical to that of Experiment 1, except that the visual and auditory stimuli consisted of pictures of birds and dogs and recordings of their respective vocalizations. We presented either same-species or different-species pairs, depending on condition (congruent or incongruent).

Results and discussion

Figure 4 shows mean performance across the six blocks in the two conditions. Overall performance was better in the congruent condition (M = 52%, SD = 10%) than in the incongruent condition (M = 44%, SD = 11%), a significant difference by t test, $t(48) = 2.94$, $p < .01$, $d = 0.761$. As in Experiment 1, we conducted a polynomial contrast analysis across the six learning blocks for the two conditions. This yielded a significant linear increase in performance as a function of block number, $F(1, 49) = 46.06$, $p < .001$, and a significant interaction between the linear components of block number and condition, $F(1, 49) = 5.29$, $p = .037$. No significant higher-order effects were found for block number or for the Block Number × Condition interaction.

The results of Experiment 2 indicated that the categorical congruency learning advantage extends to other classes of stimuli besides human faces and voices. Nonetheless, it was still possible that expert knowledge of human faces and voices may confer an additional advantage, as compared with nonhuman multisensory paired associates. To test this, we conducted an additional 2 (congruent vs. incongruent) × 2 (human vs. animal) ANOVA that yielded a significant main effect of congruency condition, $F(1, 86) = 41.71$, $p < .001$, no main effect of human versus animal, $F(1, 86) = 1.35$, $p = .248$, and a significant interaction, $F(1, 86) = 6.551$, $p = .012$. Planned comparisons revealed better performance for the human than for the animal pairs when the pairings were congruent, $t(44) = 2.552$, Bonferroni-adjusted $p = .028$, but not when they were incongruent, $t(44) = 1.030$, Bonferroni-adjusted $p = .62$. Thus, we found no general advantage for human face and voice stimuli; the advantage for this class was confined to cases in which the pairs were congruent.

General discussion

The findings from this study suggest a novel principle in paired-associate learning: People are better at learning pairs of multisensory stimulus properties that are consistent with belonging to the same object, on the basis of their categorical congruency, than at learning pairs that are not. This congruency advantage holds for human voices and faces, as well as for dogs and birds and their respective vocalizations. Several previous studies have shown that memory for a unisensory stimulus in one modality is enhanced when it is encoded together with a congruent stimulus in a different modality (Lehmann & Murray, 2005; Murray et al. 2005; Murray et al. 2004). Similarly, several studies have shown that congruent auditory information can facilitate perceptual learning in a visual-motion categorization task (Kim et al. 2008; Seitz et al. 2006). However, these studies did not measure learning of multisensory associate pairs, but instead focused exclusively on learning or memory of unisensory stimuli. Our study is thus the first to investigate the role of categorical congruency on the learning of novel multisensory pairs.

What underlies the congruency advantage reported here? Unlike previous studies reporting the effects of semantic congruency on associative learning (Heim et al. 1966; Kintsch, 1968), participants in our study had to learn *novel*

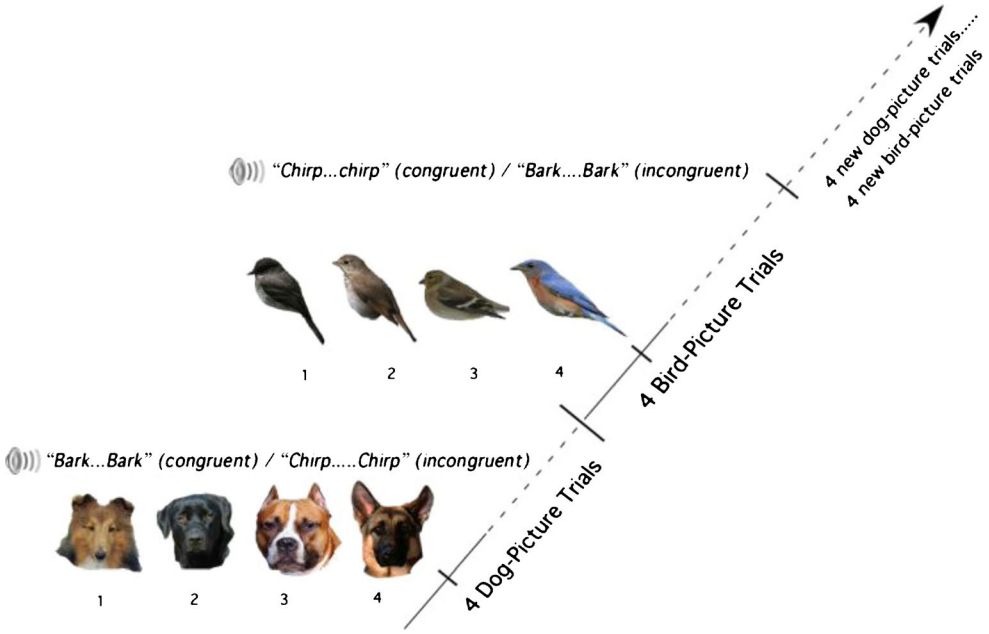

Fig. 3 Schematic of a single experimental block in Experiment 2

and arbitrary associations, and the congruency conveyed no task-relevant information. Importantly, the congruency advantage was not the result of interference in the incongruent condition; if it were, the neutral condition should have yielded a similar advantage. In fact, performance in the neutral condition and the incongruent conditions in both Experiments 1 and 2 was strikingly similar.

Overall, our findings suggest that the congruency advantage was mediated by a specialized learning mechanism that integrates multisensory properties into a single object or identity. This integration may depend on generating a novel, supramodal representation in the form of a PIN (Bruce & Young, 1986; Burton et al. 1990; Ellis et al. 1997), to which both the visual and auditory stimuli are linked. Alternatively, it is possible that face–voice associations may be based on direct connections between modality-specific face and voice areas (von Kriegstein & Giraud, 2006; von Kriegstein et al. 2005), with greater coupling in the congruent conditions. Regardless of the mechanism, our findings suggest that the formation of consolidated, multisensory "identity" networks extends to nonhuman visual and auditory stimuli, something that previous theories have not proposed.

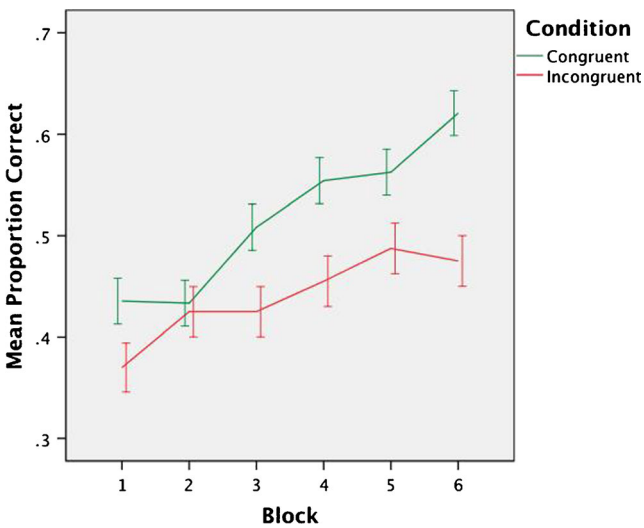

Fig. 4 Mean performance across the six blocks in the two conditions of Experiment 2. Error bars represent ±1 standard error of the mean

Interestingly, we found that the congruency was more pronounced for human stimuli than for nonhuman animal stimuli. It is important to note that we observed no similar advantage for human versus animal stimuli when the stimuli were incongruent. Thus, the heightened advantage for congruent human faces and voices was probably not due to better processing/memory of the individual human faces and voices, which should have resulted in better performance in the incongruent condition as well. The fact that this did not occur suggests that the advantage applies only to learning *relations* between congruent human faces and voices.

In conclusion, our results demonstrate that learning of novel multisensory associations is highly influenced by prior experience. Research in the classical conditioning literature has shown that some types of associations are easier to form than others. For example, Garcia and Koelling (1996) found that rats learned to associate lights and sounds—but not taste—with an electric shock. In contrast, they learned to associate poison with taste but not with sound or light. Seligman (1970) ascribed this enhanced capacity to learn certain associations—which he termed "preparedness"—to biological mechanisms that developed

over the course of evolution. The present results suggest a form of preparedness that is developmental and experience-dependent in nature, whereby novel associations that are consistent with previously experienced categorical pairings are privileged for association (Lewkowicz & Ghazanfar, 2009).

Author note This work was supported in part by NSF Grant No. BCS-0958615 to E.B., and by NSF Grant No. BCS-0751888 to D.J.L.

References

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305–327.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology, 81*, 361–380.

Cofer, C. N. (1968). Associative overlap and category membership as variables in paired associate learning. *Journal of Verbal Learning and Verbal Behavior, 7*, 230–235. doi:10.1016/s0022-5371(68)80194-x

Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology, 88*, 143–156.

Garcia, J., & Koelling, R. A. (1996). Relation of cue to consequence in avoidance learning. In L. D. Houck, L. C. Drickamer, & Animal Behavior Society (Eds.), *Foundations of animal behavior: Classic papers with commentaries* (pp. pp. 374–pp. 375). Chicago: University of Chicago Press.

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

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3*, 191–197.

Heim, A. W., Watts, K. P., Bower, I. B., & Hawton, K. E. (1966). Learning and retention of word-pairs with varying degrees of association. *Quarterly Journal of Experimental Psychology, 18*, 193–205. doi:10.1080/14640746608400030

Kim, R. S., Seitz, A. R., & Shams, L. (2008). Benefits of stimulus congruency for multisensory facilitation of visual learning. *PLoS ONE, 3*, e1532. doi:10.1371/journal.pone.0001532

Kintsch, W. (1968). Recognition and free recall of organized lists. *Journal of Experimental Psychology, 78*, 481–487. doi:10.1037/h0026462

Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research, 24*, 326–334. doi:10.1016/j.cogbrainres.2005.02.005

Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in Cognitive Sciences, 13*, 470–478. doi:10.1016/j.tics.2009.08.004

Mavica, L. W., & Barenholtz, E. (2013). Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 307–312. doi:10.1037/a0030945

Murray, M. M., Michel, C. M., Grave de Peralta, R., Ortigue, S., Brunet, D., Gonzalez Andino, S., et al. (2004). Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. *NeuroImage, 21*, 125–135. doi:10.1016/j.neuroimage.2003.09.035

Murray, M. M., Foxe, J. J., & Wylie, G. R. (2005). The brain uses single-trial multisensory memories to discriminate without awareness. *NeuroImage, 27*, 473–478. doi:10.1016/j.neuroimage.2005.04.016

Seitz, A. R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Current Biology, 16*, 1422–1427. doi:10.1016/j.cub.2006.05.048

Seligman, M. E. (1970). On the generality of the laws of learning. *Psychological Review, 77*, 406–418. doi:10.1037/h0029790

Thelen, A., & Murray, M. M. (2013). The efficacy of single-trial multisensory memories. *Multisensory Research, 26*, 483–502. doi:10.1163/22134808-00002426

von Kriegstein, K., & Giraud, A.-L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology, 4*, e326. doi:10.1371/journal.pbio.0040326

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience, 17*, 367–376. doi:10.1162/0898929053279577