Published in final edited form as:

Psychon Bull Rev. 2014 October ; 21(5): 1346–1352. doi:10.3758/s13423-014-0612-7.

Categorical congruence facilitates multisensory associative learning

Elan Barenholtz^{*},

Department of Psychology/Center for Complex Systems and Brain Sciences, Florida Atlantic University

David J. Lewkowicz, and

Department of Psychology/ Center for Complex Systems and Brain Sciences, Florida Atlantic University

Meredith Davidson

Department of Psychology, Florida Atlantic University

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, based on 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 pairs consisting of human faces and pictures of inanimate objects (plants and rocks). Experiment 2 found that this ‘categorical congruency’ advantage extended to non-human stimuli as well, namely pairs of class-congruent animal pictures and vocalizations (e.g. dogs and barks) vs. 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.

Many objects can be identified based on multiple properties from different modalities. For example, certain foods can be identified based on 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, 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 more frequent

^{*}Correspondence concerning this article should be addressed to Elan Barenholtz, Department of Psychology, Florida Atlantic University, Boca Raton, FL, 34331 elan.barenholtz@fau.edu.
Lauren Mavica, Department of Psychology, Florida Atlantic University

than male faces and high-pitched voices. How might prior exposure to these natural statistics affect our ability to learn to novel associations among properties?

Previous studies of word-list learning provide some evidence that previously learned relations among different items can facilitate learning and memory. Heim, Watts, Bower and Hawton (1966) found that 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”, than 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 pre-existing 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 when 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 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 is because learning of categorically congruent pairs, even when 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, Bruce, & Johnston, 1990). 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 ‘reintegrates’ 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 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' ability 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, 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) while 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 vs. 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 vs. 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 the 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

Experiment 1 tested the ability to learn the association between individual human voices with different classes of visual stimuli across three, between-subject, 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 un-speeded, four-alternative forced-choice (4-AFC) 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

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 analysis on this basis but were replaced with an additional four participants to achieve an equal number of participants (20) per

condition, for a total of 60 participants. The final dataset included 33 female and 27 male participants. The ages of participants ranged from 18 to 32 with an average age of 22.

Stimuli

Stimuli in the Congruent and Incongruent conditions consisted of photographs and voice recordings of eight Caucasian females and eight Caucasian males ranging in age from 18–26 years of age. Each individual was photographed and recorded speaking the sentence: “There are clouds in the sky” in an emotionally neutral tone. Visual stimuli in the Neutral condition consisted of photographs of eight rocks and eight houseplants. 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 subject 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 while for the other half of participants the pairings were reversed.

Design and Procedure

The experimental procedure was 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 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 gender (in order to avoid confusion). Figure 1 and its accompanying caption shows the structure of a single block of trials in the Congruent/Incongruent conditions. 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. Subjects 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 described in the Supplemental Materials, each participant also ran in two additional experimental blocks testing their ability to match the faces and voices from the six learning blocks, based on novel utterances of the same voices, in the absence of feedback).

Results

Figure 2 shows the proportion 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 Neutral condition [$M = 44\%$, $SD = 9\%$]. A one-way ANOVA, with condition as the between-subjects factor, found a significant effect of condition $F(2, 57) = 15.33$, $p < .0001$,

$\eta^2 = .35$. Post-hoc analysis showed that performance in the Congruent condition was significantly better than in both the Incongruent and the Neutral conditions (p 's $< .0001$) but that the Incongruent and Neutral conditions did not differ significantly from each other (p 's $> .1$).

Figure 2. Mean performance across the six blocks in the three experimental conditions of Experiment 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 vs. Incongruent condition $F(1, 38) = 5.89$, $p = .016$, as well as between the Congruent and Neutral condition, $F(1, 38) = 4.40$, $p = .033$. These interaction effects indicate that the block-wise linear learning slope was significantly steeper in the Congruent condition than in the other two conditions. There was no similar interaction between the Incongruent and Neutral conditions, $F(1, 38) = 0.44$, $p > .1$. Finally, there were no significant higher-order (i.e. non-linear) effects for block number or for block number $\times$ condition.

Although learning increased more rapidly in the Congruent condition across blocks, a separate ANOVA found a significant difference in performance starting in the first block of trials, $F(2,57) = 4.48$, $p = .016$, $\eta^2 = .36$. Post-hoc analysis found a significant difference between the Congruent condition [$M = 43\%$, $SD = 10\%$] and both the Incongruent [$M = 33\%$, $SD = 13\%$, $p = .028$] and Neutral condition [$M = 34\%$, $SD = 12\%$, $p = .049$] but not between the Incongruent and Neutral conditions ($p > .1$). 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 voice-face pairs were presented in consecutive series, with the same four voices and their corresponding faces always presented in four consecutive trials. This means that once participants received feedback in the first trial of each 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 choices¹.

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 relative to both of the

¹Additional within-block analyses are provided in the Supplemental Materials

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? Human voice and face processing is thought to rely on specialized mechanisms that depend on dedicated brain regions and/or visual expertise (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 other stimuli such as cell phone pictures paired with ringtones. Thus, associating congruent human face-voice pairs may engage specialized/expertise mechanisms that do not apply to other categories. In addition, while the face-voice pairs in Experiment 1 were arbitrary and non-veridical, participants may still have *tried* to use their expectations about facial/vocal mappings in the Congruent condition to learn the pairings, potentially effecting performance.

Thus, in Experiment 2 we investigated whether there is a congruency advantage 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 (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.

Methods

Participants

Fifty 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, with an average age of 21. None of the participants self identified as a dog or bird expert.

Stimuli and Procedure

Stimuli consisted of sound recordings of eight different mid-range dog barks and eight different bird chirps and pictures of the cropped faces of eight mid-sized dogs (chosen based on subjective judgment of size) and pictures of eight mid-sized birds (audio recordings and photos were obtained from the internet). Each sound recording was approximately three seconds long, 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 (Figure 3).

The procedure was identical to 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 & 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 [$M = 44\%$, $SD = 11\%$] and a t-test indicated that this difference was statistically significant, $t(48) = 2.94$, $p < .01$; $d = 0.761$. 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 $\times$ condition interaction.

The results of Experiment 2 indicate that the categorical congruency learning advantage extends to other classes of stimuli besides human faces and voices. Nonetheless, it is still possible that expert knowledge of certain classes of stimuli such as human faces and voices may confer an additional advantage on the learning of such multisensory paired associates as opposed to non-human ones. A comparison of the congruent conditions in Experiments 1 and 2 supports this possibility: overall mean performance in the congruent human voices and faces condition in Experiment 1 was superior to performance in the congruent animal vocalizations and pictures condition in Experiment 2, $t(44) = 2.50$, $p = .016$, $d = .68$. Importantly, however, no such difference was present for the human vs. animal comparison of the incongruent conditions across the two experiments, $t(44) = 1.03$, $p > .1$. This demonstrates that there is not a general advantage for human face and voice stimuli; the advantage for this class is confined to cases where the pairs are 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 based on their categorical congruency than pairs that are not. This congruency advantage holds for human voices and faces as well as for dogs and birds and their respective vocalizations. Although several previous studies have found 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, Foxe, & Wylie, 2005; Murray et al., 2004), these studies did not measure learning or memory of multisensory associate pairs, instead focusing exclusively on memory of the unisensory stimuli. Our study is thus the first to investigate the role of categorical congruency on 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* 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 then 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) to which both the visual and auditory stimuli are both 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 condition. Regardless of mechanism, our findings suggests that the formation of consolidated, multisensory, ‘identity’ networks extends to non-human visual and auditory stimuli, something that previous theories have not proposed.

Interestingly, we found that the congruency was more pronounced for human stimuli than for non-human animal stimuli. It is important to note that there was no similar advantage for human vs. animal stimuli when they 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 it did not 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 ‘preparedness’ to learn certain associations to biological mechanisms that developed over the course of evolution. The current 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).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported in part by an NSF grant # BCS –0958615 to E.B. and NSF grant # BCS-0751888 to D.L.

References

- Bruce V, & Young A (1986). Understanding face recognition. *British Journal of Psychology*, 77(3), 305–327. [PubMed: 3756376]
- Burton AM, Bruce V, & Johnston RA (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, 81(3), 361–380. [PubMed: 2224396]

- Cofer CN (1968). Associative overlap and category membership as variables in paired associate learning. *Journal of Verbal Learning & Verbal Behavior*, 7(1), 230–235. doi: 10.1016/s0022-5371(68)80194-x
- Garcia J, & Koelling RA (1996). Relation of cue to consequence in avoidance learning In Drickamer LDHLC (Ed.), *Foundations of animal behavior: Classic papers with commentaries* (pp. 374–375). Chicago, IL, US: University of Chicago Press.
- Gauthier I, & Tarr MJ (1997). Becoming a “greeble” expert: Exploring mechanisms for face recognition. *Vision Research*, 37(12), 1673–1682. [PubMed: 9231232]
- Heim AW, Watts KP, Bower IB, & Hawton KE (1966). Learning and retention of word-pairs with varying degrees of association. *Q J Exp Psychol*, 18(3), 193–205. doi: 10.1080/14640746608400030 [PubMed: 5947473]
- Kanwisher N, McDermott J, & Chun MM (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 17(11), 4302–4311. [PubMed: 9151747]
- Kintsch W (1968). Recognition and free recall of organized lists. *Journal of Experimental Psychology*, 78(3, Pt.1), 481–487. doi: 10.1037/h0026462
- Lehmann S, & Murray MM (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research*, 24(2), 326–334. doi: 10.1016/j.cogbrainres.2005.02.005 [PubMed: 15993770]
- Lewkowicz DJ, & Ghazanfar AA (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in Cognitive Sciences*, 13(11), 470–478. doi: 10.1016/j.tics.2009.08.004 [PubMed: 19748305]
- Mavica LW, & Barenholtz E (2013). Matching voice and face identity from static images. *J Exp Psychol Hum Percept Perform*, 39(2), 307–312. doi: 10.1037/a0030945 [PubMed: 23276114]
- Murray MM, Foxe JJ, & Wylie GR (2005). The brain uses single-trial multisensory memories to discriminate without awareness. *NeuroImage*, 27(2), 473–478. doi: 10.1016/j.neuroimage.2005.04.016 [PubMed: 15894494]
- Murray MM, Michel CM, Grave de Peralta R, Ortigue S, Brunet D, Gonzalez Andino S, & Schnider A (2004). Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. *NeuroImage*, 21(1), 125–135. doi: 10.1016/j.neuroimage.2003.09.035 [PubMed: 14741649]
- Puce A, Allison T, Gore JC, & McCarthy G (1995). Face-sensitive regions in human extrastriate cortex studied by functional MRI. *Journal of Neurophysiology*, 74(3), 1192–1199. [PubMed: 7500143]
- Seligman ME (1970). On the generality of the laws of learning. *Psychological Review*, 77(5), 406–418. doi: 10.1037/h0029790
- Thelen A, & Murray MM (2013). The Efficacy of Single-Trial Multisensory Memories. *Multisensory Research*, 26(3), 1–18. [PubMed: 23713196]
- von Kriegstein K, & Giraud A-L (2006). Implicit Multisensory Associations Influence Voice Recognition. *PLoS Biol*, 4(10), e326. [PubMed: 17002519]
- 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(3), 367–376. [PubMed: 15813998]

Figure 1.

Schematic of a single experimental block in the Congruent/Incongruent conditions of Experiment 1 (see text for details). Neutral condition not shown. On each trial, four faces were presented with the numbers 1–4 in sequence below them. The spatial arrangement/number of the visual stimuli varied randomly across trials. One of the four visual stimuli was the ‘match’ to the voice, as determined prior to the experiment and as described above, while the other three served as distracters. 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 participants had chosen it or not—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 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.

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

Figure 4. Mean performance across the six blocks in the two conditions of Experiment 2. Error bars represent one $\pm$ one standard error of the mean.