Proceedings of the Cognitive Science Society (2011), Vol. 33

Multisensory Associative-Pair Learning: Evidence for 'Unitization' as a specialized mechanism

Elan Barenholtz (elan.barenholtz@fau.edu)
Department of Psychology, 777 Glades Road
Boca Raton, FL 33433

Meredith Davidson (mdavid14@fau.edu)
Department of Psychology, 777 Glades Road
Boca Raton, FL 33433

David Lewkowicz (lewkowic@fau.edu)
Department of Psychology, 777 Glades Road
Boca Raton, FL 33433

Abstract

Learning about objects typically involves the association of multisensory attributes. Here, we present three experiments supporting the existence of a specialized form of associative learning that depends on 'unitization'. When multisensory pairs (e.g. faces and voices) were likely to both belong to a single object, learning was superior than when the pairs were not likely to belong to the same object. Experiment 1 found that learning of face-voice pairs was superior when the members of each pair were the same gender vs. opposite gender. Experiment 2 found a similar result when the paired associates were pictures and vocalizations of the same species vs. different species (dogs and birds). In Experiment 3, gender-incongruent video and audio stimuli were dubbed, producing an artificially unitized stimulus reducing the congruency advantage. Overall, these results suggest that unitizing multisensory attributes into a single object or identity is a specialized form of associative learning

Introduction

Learning about objects typically involves the detection and association of multisensory attributes. For example, we may be able to identify certain foods based on their visual, gustatory, tactile as well as olfactory properties. Likewise, 'knowing' a person typically means being able to associate his or her face with his or her voice. How do we encode the multisensory properties of objects? One possibility is that such "object knowledge" simply consists of a network of associations among each of an object's unisensory properties. According to this view, our knowledge about unitary objects may depend on the same learning mechanisms as other types of object memory, such as associations between different objects or between objects and other properties of the environments in which they appear. A second possibility is that multiple unisensory object properties are all linked via an intermediate 'supramodal' representation of the object (Mesulam, 1998). According to this view, associating intra-object information is a special class of associative learning, involving the creation of a 'unitized' representation (Cohen, Poldrack, & Eichenbaum, 1997; Eichenbaum, 1997; Eichenbaum & Bunsey, 1995). This view is represented in a number of theories of face recognition which hold that associating the face and voice of an individual depends on integrating distinct informational streams into a single, 'Personal Identity Node', or PIN (Bruce & Young, 1986; Burton, Bruce, & Johnston, 1990; Ellis, Jones, & Mosdell, 1997).

Unitizing multisensory properties may make multisensory object-knowledge more efficient, since each observed property of that object may be associated with all other, previously observed, properties via a single link, rather than maintaining associations among many disparate properties. An additional potential advantage to a unitized representation, implicit in the PIN model, is that it may help to organize associations that go beyond specific stimulus-stimulus pairings to more abstract properties of an underlying 'object'. For example, if one has encountered a specific auditory utterance of an individual, along with his or her face, it would be advantageous to associate a different utterance by the same individual with that face. Presumably, this depends on extracting 'invariant' properties of the underlying voice from the sample. Representing individual face and voice stimuli as properties of the same underlying individual may facilitate this process.

Despite the potential theoretical advantages to unitization, there has been no direct behavioral support for the idea that multisensory unitization is a specialized form of associative learning. In the current study, we compared associative learning of visual/auditory pairs under conditions where the members of the pair were either likely or unlikely to belong to the same object by virtue of their membership in the same or different category. Specifically, we compared face/voice learning when the members of each pair were of the same or opposite gender (Experiment 1) or the same or different species (Experiment 2). We reasoned that since only congruent pairs are consistent with belonging to the same object (for example, our experience is that people with male faces always have male voices) they would be likely to be

'unitized' into a single object or identity, while incongruent pairs would be remembered based on simple associative processes, without unitization. This difference may be reflected in better learning of the congruent pairs.

Experiment 1

In Experiment 1, we compared learning of face-voice pairs of the same gender (congruent), versus learning of pairs of the opposite genders (incongruent). Importantly, because the task in both conditions was to learn arbitrarily matched faces and voices, they were—in terms of inherent task demand— equally difficult. Critically, we hypothesized that the pairs in the gender-congruent condition were more likely to be unified into a single identity and that this would result in better learning performance. We measured performance in an initial learning phase in which participants had to learn associations between pictures of specific faces and specific utterances (single sentences) using a forced-choice task with feedback. We then measured generalization of learning in a second phase where participants had to match each previously learned face with new utterances (2 novel sentences) produced by the same voices as before. All of the experiments used a between-subjects design.

Methods

Participants

Fifty undergraduate psychology students (25 assigned to each of the two experimental conditions), naïve to the purposes of the experiment, participated for course credit. Each student was screened after the experiment and asked whether they personally knew any of the people whose faces/voices were shown during the experiment. Participants who recognized one of the people used in the stimuli were not included in the analysis.

Stimuli

Stimuli consisted of photographs and voice recordings of 8 Caucasian females and 8 Caucasian males ranging in age from 18-26. Each individual was photographed and also recorded speaking three sentences: 1) "There are clouds in the sky", 2) "The boy took his sister to the park", and 3) counting from one to five. All photographs displayed the head and shoulders of the person from a frontal viewpoint. Before the beginning of the experiment, each of the 16 face images was matched with a single recorded voice as the 'pair' to be learned by the subject. In the Congruent conditions 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. In the Incongruent condition, each of the female faces was paired with a single randomly chosen male voice and vice versa.

Procedure

The procedure was identical in both conditions. Participants were instructed that they would be performing a task in which they must learn to match faces and voices and that that they would receive feedback on correct or incorrect responses. In the Incongruent condition subjects were additionally informed that the faces and voices would be of opposite gender. Each participant took part in a Learning Phase and a Generalization Phase. On each trial of the Learning Phase, participants were presented with a voice recording of one of the three recorded sentences, while four faces were presented on the screen with the numbers 1-4 below them (Figure 1). One of the four faces was the 'match' to the voice, as determined prior to the experiment as described above, while the other three served as distracters. The subjects were instructed to choose which of the four faces was matched with the voice. An incorrect response resulted in a low beeping sound. The correct selection was flashed once—regardless of whether subjects had chosen it or not— before the stimuli were replaced by a white screen. The face-voice stimuli were presented in groups, with each group containing four faces and voices; the faces within a single group were either all male or all female. There were four groups (2 male, 2 female), which were repeated, in six separate experimental blocks, for a total of 96 trials (4 trials per group X 4 groups X 6 blocks) per participant.

The Generalization Phase began immediately after the subjects completed the Learning Matching Phase. The procedure in the Generalization Phase was identical to the Learning Phase except that recordings of two new sentences, not heard in the Learning Phase, were used and that subjects did not receive feedback. The task of the participant was to match the face to the new voice recording, based on the face-voice pairs they had learned in the Learning Phase. Each participant performed two test blocks, one for each of the two new voice recordings: each test block consisted of four groups of four faces as in the Learning Phase for a total of 32 (4 groups of four faces X 2 blocks) per participant.

Results and Discussion

Figure 1a shows the results of the initial learning phase as a function of block for the two congruency conditions. While learning is apparent in both the Congruent and Incongruent conditions, it was much more efficient in the Congruent condition (peaking at 75% correct; chance performance was 25%) than in the Incongruent condition (peaking at 50% correct). A two-way ANOVA on the Learning data found a significant main effect of both block number $F(5,72) = 31.536$, $p < .00001$] and Congruency condition. [$F(1,72) = 178.962$, $p < .00001$] and no significant interaction.

Performance in the Generalization Phase was reduced in both conditions relative to performance in the learning phase (Figure 1b) but it was still well above chance for the congruent pairs [$t(28)=6.86$; $p<.001$], indicating that

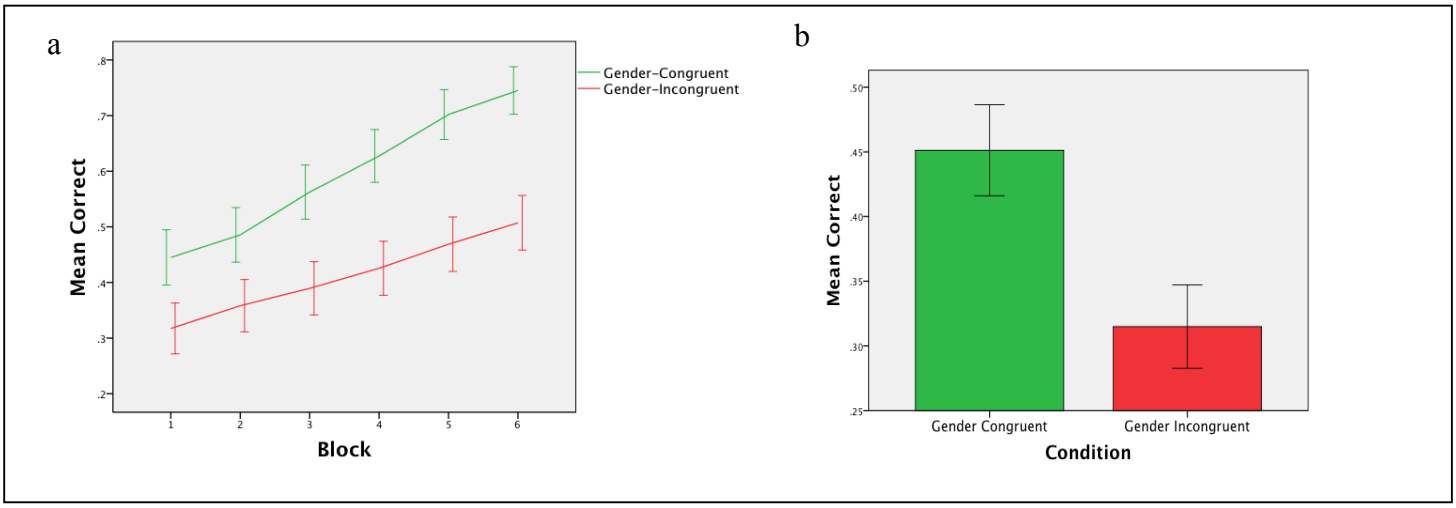

Figure 1: Experiment 1 Results. (a) Performance in the Learning phase as a function of block for the two conditions. (b) Performance in the Generalization Phase for the two conditions.

participants successfully generalized their learning to new utterances, whereas performance for incongruent pairs declined to near chance (31%; chance was 25) [*t(28)=1.23; p*<.05]. There was a significant difference in performance between the Gender-Congruent and Incongruent conditions by t-test [*t* (48) = .325, *p* = .001].

These findings indicate that generalization of learning was much more successful when the face-voice pairs were gender-congruent.

Overall, learning was more efficient and more generalized when the faces and voices making up the pairs were the same gender then when they were of the opposite gender. Since the inherent task difficulty was the same in both conditions, (i.e. the congruency did not yield any task-relevant information) the difference in performance is likely due to the fact that the incongruent pairs could not be unitized into a single identity and that learning depended on simple associative learning of the pairs.

Experiment 2

Experiment 2 investigated whether the congruency advantage observed in Experiment 1 is specific to human faces and voices. Evidence indicates that human face and voice processing are specialized processes that depend on dedicated brain regions (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Kanwisher, McDermott, & Chun, 1997; Puce, Allison, Gore, & McCarthy, 1995) and/or visual expertise (Gauthier, Skudlarski, Gore, & Anderson, 2000; Gauthier & Tarr, 1997) and that learning face-voice pairs preferentially leads to cross-activation of these unimodal selective areas(von Kriegstein et al., 2008) This raises the possibility that the multisensory unitization that we found in Experiment 1 is restricted to the learning of human faces and voices rather than a reflection of a general learning process. To test this possibility, in Experiment 2 we used the same methods as in Experiment 1 except that this time we presented pictures and vocalizations of dogs and birds and compared learning of congruent pairs (e.g. a specific dog picture and a specific bark) with incongruent pairs (a specific dog picture with a specific bird song). Then, to provide converging evidence for the concept of multisensory unitization, rather than testing for generalization of learning, we re-tested learning after a 10-minute delay to determine whether within category learning might be more robust than simple associative learning.

Methods

Participants

Sixty undergraduate psychology students (30 for each of the two experimental conditions), naïve to the purposes of the experiment participated for course credit.

Stimuli

Stimuli consisted of pictures of cropped faces of 8 ‘mid-sizes’ dogs (chosen based on subjective judgment) and pictures of 8 typically sized birds as well as sound recordings of 8 different mid-range dog barks and 8 different bird chirps (photos and audio recordings were obtained from the internet).

Procedure

As in Experiment 1, each participant first performed a Learning Phase, in which they were given feedback while learning specific picture-vocalization pairs across six blocks. After the Learning Phase, participants took a 10 minute break in which they viewed unrelated videos on the web after which they performed a final Test block consisting of the same exact task as in the Learning Phase, but without feedback.

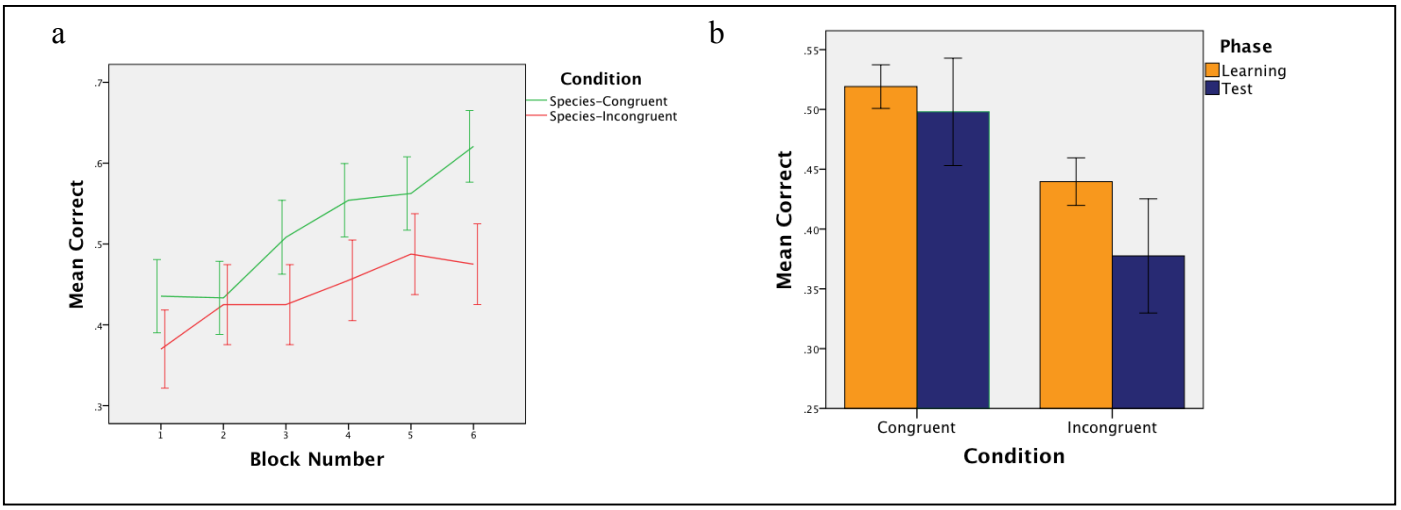

Figure 2: Experiment 2 results. (a) Mean correct in the Learning Phase as a function of block number, for the two conditions. (b) Results of the Learning Phase and the Test Phase for the two conditions.

Results and Discussion

Figure 2a shows the results of the learning phase across the two conditions as a function of block number. Even though participants were able to learn both congruent and incongruent pairs, they exhibited a significant advantage in learning the species-congruent pairs vs. the incongruent pairs [t (58) = 2.736; p < .01]. Figure 2b shows the performance in the initial Learning Phase compared to the Test Phase for the two conditions. Performance did not drop significantly following the 10 min delay for the congruent pairs [t (29) = .61; p > .5] but did decline significantly for the incongruent pairs [t (29) = 2.23; p < .05]. These results again suggest that pairs that may be unitized into a single object lead to a different learning pattern than non-unitizable stimuli.

Experiment 3

The results from Experiments 1 & 2 indicate that learning of multisensory associations is better when the paired properties belong to the same object. However, this advantage alone does not indicate that the difference in performance is due to 'unitization' per se rather than some other effect of their congruency. In Experiment 3, we used the same method as in Experiment 1 except that here we also presented some subjects with 'dubbed' movies during the pair-learning phase. This consisted of presenting faces that could be seen and heard talking in synchrony. Because temporal audio-visual synchrony can be a powerful cue to the integration of visual and auditory stimulation (Lewkowicz, 2010), we expected that synchrony might encourage subjects to unitize the face-voice pairs even in the gender-incongruent condition. If that is the case, this, in turn, might reduce the congruency advantage.

Experiment 3 included four between-subject conditions: Gender-Congruent and Incongruent (as in Experiment 1), each with a Motion version (which included the dynamically speaking faces) and a Static version (in which only a static picture of the face was shown). This allowed us to compare the effect of motion on the Congruent and Incongruent conditions. In particular, we were interested in the possibility that motion would produce a larger benefit in the Gender-Incongruent condition because it could encourage unitization for pairs of stimuli that would otherwise not be unitized.

Methods

Participants

One hundred and twenty undergraduate psychology students (30 for each of the four experimental conditions), naïve to the purposes of the experiment participated for course credit.

Stimuli

Stimuli were movies featuring the same individuals and utterances as in Experiment 1. Each movie was created by dubbing the audio recording of one person's utterance onto the synchronized video of a different person speaking the same utterance[1]. In the Static condition only a still frame of each movie clip was shown (as described below) while in the Motion condition, the dubbed movie was shown.

Procedure

As in Experiment 1, each participant first took part in a learning phase, in which they were given feedback while learning specific face-voice pairs in groups of four. However, before performing the forced choice task, each face in the group was presented in conjunction with the recording of the matched voice. In the 'Motion' conditions, the face was a video of the person speaking, accompanied by the matched voice. In the 'Static' conditions, the face was a still-frame taken from the video sequence. This initial

[1] In order to facilitate synchronization, individuals were recorded uttering each phrase while listening on headphones to a recording of a repeated, 'standard' version of that phrase. This yielded high degrees of synchrony across different individuals' recordings with only a small amount of editing needed to bring them into a high degree of alignment.

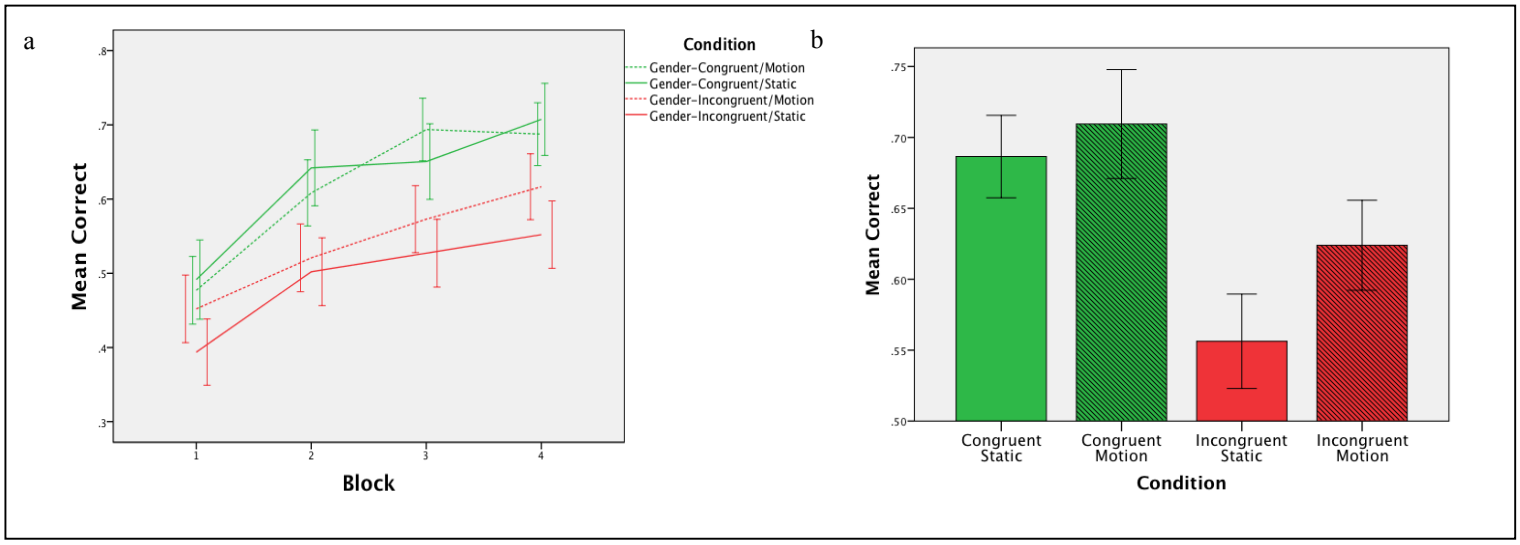

Figure 3: Experiment 3 Results. (a) Mean correct for the Learning Phase, as a function of block number, for the four conditions. (b) Mean correct in the Generalization Phase for the four conditions.

sequence of four face-voice presentations was then followed by the exact same forced-choice task as in Experiment 1. During the learning phase, participants were tested on four groups of four people (8 male face and 8 female faces) repeated across four blocks of trials for a total of 64 trials per participant.

After the learning phase, participants completed a generalization phase in which they had to try to match each learned face with the previously paired voice, now uttering a new sentence. On each trial, participants were presented with two face-voice stimuli (either static or moving, depending on condition) in succession: one in which the face was matched with the same voice it had been paired with in the learning phase and one where it was paired with one of the voices that had been paired with a *different* face in the learning phase. Participants had to chose which of the two stimuli matched the learned face-voice pairings. No feedback was given.

Results

Figure 3a shows the results of the learning phase for each of the four conditions (Gender Congruent/Incongruent in both Motion and Static Cases). Participants exhibited learning of congruent and incongruent pairs in both the dynamic and static conditions (main effects for block number [F (3, 116) = 49.89; $p < .0001$]. As in the previous experiments, the two gender-congruent conditions yielded better performance than the two gender-incongruent conditions [F (1, 116) = 77.75; $p < .0001$]. There was no significant effect of motion ($p > .1$). However, as Fig. 3a shows, learning was marginally greater for gender-mismatched pairs when the stimuli were dynamic, and thus synchronized, than when they were static [$t(48)$ = 1.675; p = .06]. However, learning was not enhanced by synchrony for gender-congruent pairs ($p > .5$). Figure 3b shows that performance in the generalization phase, where chance performance was .5, mirrored the performance in the initial learning phase. Here, response to the gender-matched pairs was equivalent regardless of whether synchrony cues were provided or not [t (48) = .964, $p > .1$], but was more robust for the moving gender-incongruent pairs than for the static ones. Thus, synchrony cues do not facilitate learning or generalization when multisensory information is easily unitized but does facilitate them when the information is not likely to be unitized.

General Discussion

The current results demonstrate a previously unreported phenomenon in associative-pair learning. We found that learning to pair multisensory stimulus properties was much more efficient, robust, and general when the paired properties were members of the same category vs. when they were not. This advantage is likely due, at least in part, to the ability to *unitize* the pairs in the congruent category conditions since artificially encouraging unitization —as in Experiment 3—significantly decreased the congruency differential. The current results with regard to faces and voices are consistent with earlier theories of personal identity representation, such as Bruce and Young's (1986) theory in which multiple properties are integrated via a single node. However, the extension of the congruency advantage to visual and auditory pairs of other species—as in Experiment 2— suggests that unitization may be a general mechanism, that extends to other kinds of objects. If so, these results may suggest a fundamental dichotomy between 'simple associative learning'— which applies to associations among properties of different objects—and unitization— which applies to associations of stimulus properties corresponding to a single object. Indeed, the current behavioral results bear interesting relations to previous findings in both the neuropsychology and neuroimaging literatures suggesting that "intra-item" and "inter-item" memories are encoded in distinct neural substrates (Cohen et al., 1997; Eichenbaum, 1997; Eichenbaum & Bunsey, 1995). This raises the intriguing possibility that the different learning patterns observed in our study for congruent vs. incongruent pairs may represent neurally separable mechanisms.

The process of unitization discussed here has clear relations to the concept of 'binding' in attention and short-term memory. The so-called 'Binding Problem' refers to the process by which different properties—typically visual

properties such as shape and color—are identified and remembered as belonging to a single object during a task such as visual search or identification. Generally, this process is thought to involve a specialized process, requiring attentional mechanisms, in order to integrate the separate properties into a single 'object-file' (Treisman & Gelade, 1980). This mechanism is also thought to underlie the capacity limitations of working memory (Luck and Vogel, 1997). The object-files formed in these cases are assumed to be inherently short-lived, lasting perhaps only as long as the stimulus remains in working memory (Wheeler and Treisman, 2002. However, the current results suggest the existence of a long-term object-file mechanism as well.

Acknowledgments

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

References

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *NAture, 403*, 309-312.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305-327.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology, 81*(3), 361-380.

Cohen, N. J., Poldrack, R. A., & Eichenbaum, H. (Eds.). (1997). *Memory for items and memory for relations in the procedural/declarative memory framework*: Mayes, Andrew R.; Downes, John Joseph (1997). Theories of organic amnesia.

Eichenbaum, H. (1997). Declarative memory: Insights from cognitive neurobiology. *Annual Review of Psychology, 48*, 547-572.

Eichenbaum, H., & Bunsey, M. (1995). On the binding of associations in memory: Clues from studies on the role of the hippocampal region in paired-associate learning. *Current Directions in Psychological Science, 4*(1), 19-23.

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*(1), 143-156.

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*(2), 191-197.

Gauthier, I., & Tarr, M. J. (1997). Becoming a "greeble" expert: Exploring mechanisms for face recognition. *Vision Research, 37*(12), 1673-1682.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience, 17*(11), 4302-4311.

Lewkowicz, D. (2010). The Ontogeny of Human Multisensory Object Perception: A Constructivist Account. In *Multisensory Object Perception in the Primate Brain*: Springer Press.

Mesulam, M. M. (1998). From sensation to cognition. *Brain, 121*, 1013–1052.

Puce, A., Allison, T., Gore, J. C., & McCarthy, G. (1995). Face-sensitive regions in human extrastriate cortex studied by functional MRI. *Journal of Neurophysiology, 74*(3), 1192-1199.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97-136.

von Kriegstein, K., Dogan, z. r., Grüter, M., Giraud, A.-L., Kell, C. A., Grüter, T., et al. (2008). Simulation of talking faces in the human brain improves auditory speech recognition. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 105*(18), 6747-6752.