Multisensory Statistical Learning: Can Associations between Perceptual Categories Be Acquired?

Anne McClure Walk (awalk@slu.edu)
Christopher M. Conway (cconway6@slu.edu)
Department of Psychology, 3511 Laclede Ave.
Saint Louis University
Saint Louis, MO 63103 USA

Abstract

Statistical learning, the process by which people learn patterns of information from their environment that they can apply to new situations, is central to the development of many higher order cognitive skills. Despite a growing research literature, little is still known about how statistical learning operates across perceptual categories. To investigate this issue we assessed college students on their ability to learn a multisensory artificial grammar containing both auditory and visual elements and both within-categorical and cross-categorical associations. The results of Experiment 1 showed that participants were sensitive to grammatically correct test items and ungrammatical test items that contained within-categorical grammatical violations, but were not sensitive to items that contained cross-categorical violations across sensory modalities. Experiment 2 showed that participants were not sensitive to items that contained cross-categorical violations within the same sensory modality. Our findings suggest that multisensory integration across perceptual categories does not occur easily during statistical learning.

Keywords: statistical learning, artificial grammar learning, multisensory processing, domain-general

Introduction

Statistical learning, the ability to detect statistical associations in the environment (Perruchet & Pacton, 2006), appears to be important across a range of cognitive domains, including language, motor skills, and event segmentation (Conway, Pisoni, Anaya, Karpicke, & Henning, 2011; Conway, Bauernschmidt, Huang, & Pisoni, 2010; Leclerq & Majerus, 2010; Zacks & Swallow, 2007). Despite a growing body of research investigating different aspects of statistical learning, little is known about how learning takes place across perceptual categories and sensory modalities.

To illustrate the importance of multisensory processing in cognition, we briefly consider its role in speech perception and production, which require the integration of material across perceptual categories. Rosenblum (2008) suggested that spoken language processing is naturally a multisensory phenomenon, pointing out that infants appear to use visual speech cues early in life to help perceive speech. Furthermore, when one sensory modality is insufficient for perceiving a speech element, the other modality can be recruited: for example, phonemes that are auditorily similar tend to be visually distinct in terms of facial and mouth movements. The importance of multisensory processing in speech perception is also seen in the well known McGurk illusion (McGurk, 1976) in which participants see a video of a person's mouth verbalizing one syllable, while an auditory track is played of a different syllable. When the auditory input does not match the visual input, participants report perceiving a hybrid syllable constructed from combining the visual and auditory information.

Clearly, multisensory processing is an important phenomenon. However, it is still unknown to what extent cross-categorical inputs can be integrated in the case of statistical learning. One possibility is that statistical learning is domain general, and therefore operates equally across all modalities and perceptual categories. Under this view, one would expect that multisensory statistical learning would be robust, and that learning would be comparable across domains. Indeed, Seitz, Kim, Wassenhoven, and Shams (2007) used a statistical learning paradigm to demonstrate that participants learned both audio and visual patterns independently when presented with audio-visual pairings, indicating equivalent levels of learning when exposed to stimuli from different sensory modalities. Several studies have also demonstrated improved performance when stimuli are presented in two rather than a single modality (Kim, Seitz, & Shams, 2008; Robinson & Sloutsky, 2007), which could indicate that stimuli in different modalities are integrated together during statistical learning tasks. Furthermore, several studies have shown transfer between sensory domains, suggesting that knowledge resulting from statistical learning processes can be easily integrated across input domains and perceptual categories (Altmann, Dienes, & Good, 1995; Manza & Reber, 1997).

On the other hand, recent research suggests that statistical learning may not be purely domain-general. For instance, modality constraints exist which bias and affect how statistical patterns are acquired (Emberson, Conway, & Christiansen, in press; Conway & Christiansen, 2005). The presence of these modality constraints suggest that although learning across perceptual domains might operate using similar computational principles, each modality may also be biased to acquire certain types of information better than others. Even so, whether people are able to learn patterns when cross-categorical dependencies are employed is a less explored issue. Conway and Christiansen (2006) showed that when learning two separate sets of regularities concurrently, participants demonstrated learning only when the two sets of stimuli were in different sensory modalities or perceptual categories. They argued that this demonstrates that statistical learning relies on stimulus-specific rather

than abstract representations since no "mixing" of the information occurred across sensory modalities. These last findings suggest that to some extent, information across sensory modalities is not easily integrated during statistical learning, raising doubts as to a completely domain-general view of statistical learning.

Reconceptualizing Modality Differences

The previously reviewed findings raise difficulties with adopting a purely domain-general view of statistical learning. However, perhaps the problem lies in the inadequacy of using a strict dichotomous classification of either purely domain-general versus purely domain-specific (illustrations of each are depicted in Figure 1) models. In a domain-general model, all input types and modalities are treated equally, offering complete integration across perceptual categories and sensory modalities. On the other hand, in a domain-specific model, no integration occurs at all between specific sensory modalities or perceptual categories. Although there may be some theoretical usefulness out of depicting these views, sensory integration is likely more complex than either model would imply.

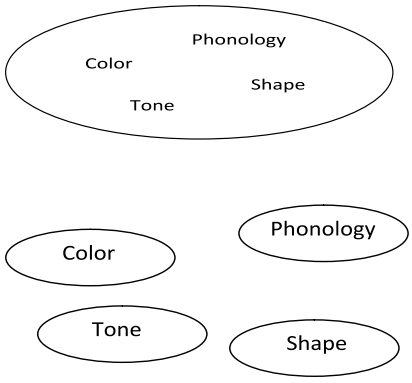

Figure 1. Domain-general model (top) versus domain-specific model (bottom) of sensory integration.

Cree and McRae (2003) investigated a similar problem in the psycholinguistic literature regarding semantic categorization. These authors reconceptualized the previously debated question as to whether semantic categorization is stored in a domain- or knowledge-specific manner, by statistically analyzing a large corpus of nouns according to various theoretical categorizational constructs, such as concept familiarity, word frequency, and visual complexity, among others. From their analyses, they found that semantic categorization can actually be conceptualized as a combination of all of the proposed constructs. Thus, they suggested a reconceptualization of the traditional domain-general/domain-specific division, into one that is more integrative (McNorgan, Reid, & McRae, 2011).

As a variation of the domain-general view, which suggests that all sensory modalities are processed within a single cognitive mechanism, McNorgan et al. (2011) proposed a shallow integration model, as depicted in Figure 2 (top). In this model, different modality features, such as shape and color for vision, enter onto different featural nodes. These nodes feed input into a central processing mechanism where the various input is integrated, producing an overall sensory experience. Importantly, in the shallow model the sensory features do not load onto a modality-specific node before moving to the central processing mechanism. Rather, various visual features, such as shape and color, and auditory features such as pitch and tone all interact once reaching the central processing mechanism. Thus, modalities are initially percept specific, but become integrated at a higher level of processing.

In addition, as an alternative to the domain-specific view, which proposes that all sensory modalities are completely isolated from each other, McNorgan et al. (2011) proposed a deep integration model (Figure 2, bottom). In this model, an additional level of nodes is introduced. Sensory input enters and is loaded onto a featural node as before, then passes onto a modality-specific sensory node, such as vision, before entering the central processing mechanism. As an example, according to this model, once a tone of a particular pitch is perceived, it loads onto the pitch node, and then is integrated with phonology and other auditory features before entering the central processor. Here the auditory information can be further integrated with information from other sensory modalities.

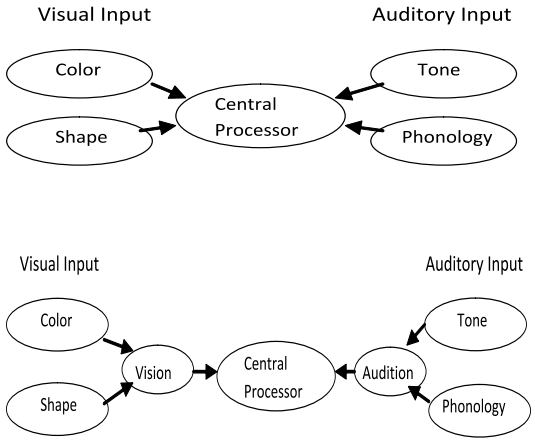

Figure 2. Shallow integration model (top) and deep integration model (bottom), adapted from McNorgan et al. (2011).

The Present Study

We believe that the perspectives offered from these shallow and deep integration models can provide insight into better understanding multisensory statistical learning. The purpose of the present study is to begin to tease apart which of these models might offer the most explanatory power for multisensory/multi-categorical processing in statistical learning. The present experiments employ an artificial grammar learning (AGL) paradigm, a common paradigm used to test such learning (Perruchet & Pacton, 2006; Reber,

1967; Seger, 1994). The traditional AGL paradigm exploits the probability between different inputs by using a finite state grammar. Traditionally, these inputs consist of various elements in a single modality or perceptual category. Thus, a particular input sequence may be a series of pictures, tones, or letters, the order of each element being determined by the grammatical rules. Our paradigm differed from the traditional in that instead of using inputs from a single perceptual category, we used elements from multiple domains, such that both within-categorical and cross-categorical associations were present. Other studies (Robinson & Sloutsky, 2007) that have used inputs from multiple domains have bound them in such a way that when an element from one perceptual category (e.g. a visual element) appeared, it always co-occurred with an element from a different category (e.g. an auditory element). In contrast, we treated all sensory category inputs as individual units of the grammar. Thus, in Experiment 1, participants were exposed to a learning phase in which they heard tones interspersed with pictures that appeared on a screen (see Figure 3). Each auditory element could be followed by a visual or auditory element, and vice versa, creating a unique grammar consisting of three independent visual and three individual auditory elements. Importantly, because the learning phase consisted of both within-categorical and cross-categorical associations, we could test to what extent participants can acquire each, which may help us distinguish between the four possible models of multisensory integration discussed above. In Experiment 1, we employed two sets of stimuli from two different sensory modalities (visual shapes and auditory tones); in Experiment 2, we employed two categories of auditory stimuli (tones and nonwords).

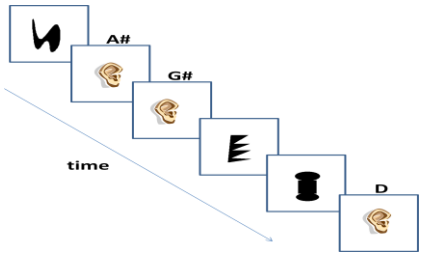

Figure 3: Example of a possible input sequence used in the present study.

Experiment 1

Method

Participants Fifteen undergraduate students from Saint Louis University participated in the study. All participants received credit toward partial fulfillment of an undergraduate course as compensation for their time. All participants reported being native speakers of English with vision and hearing at normal or corrected to normal levels.

Stimulus Materials For the learning task, we used an artificial grammar consisting of three visual elements and three auditory elements. The visual elements were abstract black shapes that were difficult to verbally label. The auditory elements were three tones generated using Audacity software having frequencies of 210, 286, and 389Hz. These frequencies were used because they neither conform to standard musical notes nor have standard musical intervals between them (Conway & Christiansen, 2005).

We used an artificial grammar with constrained probabilities to generate the input sequences (see Table 1). To generate a sequence from such a grammar, one randomly picks a starting element on the left (A-1, V-2, A-3, V-4, A-5, or V6) and then uses the listed probability to generate the next item. For instance, if V-2 is the starting element, it can be followed by either A-3 or V-4; if A-3 is the element occurring next, it can be followed by either V-4 or A-5. Thus, V-2, A-3, A-5 is an example of a short three-item input sequence that can be generated by this grammar.

In general, the grammar specifies that each auditory element has .5 probability of being followed by one other auditory element and a .5 probability of being followed by a visual element. Likewise, each visual element can be followed half of the time by one other visual element, and half of the time by a single auditory element. Thus, each element of the grammar could be followed by two other elements, one of the same modality, and one from the other modality. For Experiment 1 the within-categorical items were also within-modal (e.g., auditory-auditory and visual-visual), and the cross-categorical items were also cross-modal (e.g., auditory-visual or visual-auditory). Two types of ungrammatical items were also generated, within-modal violations and cross-modal violations. To create within-modal violation items, all within-modal dependencies were altered so that they violated the grammar; however, all cross-modal dependences remained grammatical. For cross-modal violation items, all cross-modal dependencies did not conform to the grammar; however, the within-modal dependencies remained grammatical.

Table 1: The probabilities used to formulate grammatical sequences for the learning phase and test items, which consisted of visual ("V") and auditory ("A") elements.

	A-1	V-2	A-3	V-4	A-5	V-6
A-1	0	.5	.5	0	0	0
V-2	0	0	.5	.5	0	0
A-3	0	0	0	.5	.5	0
V-4	0	0	0	0	.5	.5
A-5	.5	0	0	0	0	.5
V-6	.5	.5	0	0	0	0

Procedure All participants completed two phases of the task: a learning phase and a test phase. In the learning phase, participants were directed to put on a pair of headphones,

and pay attention to the pictures that flashed on the screen as well as any sounds they heard through the headphones. Participants were exposed to a continuous 7-8 minute sequence of pictures and tones that coincided with the grammar. In the second phase of the experiment, participants observed novel six-item sequences and had to determine if each item was grammatical (i.e., it "followed the rules" of the sequences they heard during the learning phase) or ungrammatical (i.e., it "did not follow the rules"). Participants were given 20 novel grammatical test items, 10 ungrammatical cross-modal violation items, and 10 ungrammatical within-modal violation items, in random order. Participants made their responses by pressing one of two buttons on a button box, one signifying grammatical items, the other signifying ungrammatical items. For each participant, the auditory and visual tokens were randomly assigned to the elements in the grammar; thus, for one participant, A-1 might be the 210 Hz tone, but for another participant, A-1 might be the 389 Hz tone.

Results and Discussion

The present study serves as an initial test of the domain-general and domain-specific models of sensory integration. If people process statistical information domain-generally, we expect to see no difference between performance in detecting within-modal and cross-modal violations. Under this view, what is important is that there exists a violation to the grammatical regularities, and participants should therefore be able to detect such violations, regardless if it is a cross-modal violation (e.g., detecting that A-1, V-4 is an illegal transition). However, if statistical learning is domain-specific, with learning focused solely on transitions within a sensory modality, then it might be expected that participants should fail to identify cross-modal violations.

Table 2 lists percent correct judgments for each of the three item types (grammatical, ungrammatical within-modal violations, and ungrammatical cross-modal violations). A series of single sample t-tests were run comparing the group means to chance performance (50%). A group mean significantly higher than chance would signify learning.

Table 2: Mean performance for Experiments 1 and 2. Values presented are percentage correct for each condition.

Group	Mean (SD)		
	Gram	Within-Cat	Cross-Cat
Experiment 1	59.35(11.3)*	65.3(13.0)*	50.7(17.5)
Experiment 2	60.65(9.4)*	78.7(14.6)*	51.3(22.3)

As can be seen from Table 2, learning occurred for the grammatical items ($t = 3.19$, $p < .01$) and the within-modal violation items ($t = 4.56$, $p < .001$). However, no learning was seen for the cross-modal violation items ($t = 0.15$, $p > .5$).

In other words, participants could reliably recognize a grammatical item as grammatical and could detect within-modal violations. However, they were unable to detect statistical violations that occurred between two elements from two different modalities. These results indicate that learning statistical associations between two elements may be more difficult when it takes place across two modalities compared to when it occurs within the same modality. Because no cross-modal integration was seen in Experiment 1, we can conclude that the domain-general modal is not an accurate depiction of the type of processing taking place in multisensory statistical learning.

Experiment 2

The results of Experiment 1 show that participants may be unable to use knowledge gained through statistical learning to identify sequences that contain a cross-categorical violation. However, Experiment 1 tells us only how information is integrated between sensory modalities, but nothing about how information is integrated within a single modality. Experiment 2 was conducted to further investigate to what extent different features from a single modality are integrated and learned, in order to test the shallow integration model of statistical learning.

Method

Participants Participants in this study were fifteen undergraduate students from Saint Louis University. As in Experiment 1, all participants received credit toward partial fulfillment of an undergraduate course as compensation for their time. All participants reported being native speakers of English with vision and hearing at normal or corrected to normal levels.

Stimulus Materials For Experiment 2, the stimulus materials were two different types of auditory stimuli. The same three tones used in Experiment 1 were used in this experiment with the addition of two tones, at frequencies 245 and 333 Hz, to give a total set of five tones. As in Experiment 1, the two additional tones did not conform to standard musical notes or contain intervals of any standard musical scale. In addition, five nonsense syllables were used for the second stimulus type: "vot," "pel," "dak," "jic," and "rud" (from Gómez, 2002). For each participant, three of the tones and three nonsense syllables were randomly selected and mapped onto the sequences. Thus, each participant received the same sequences (generated from the grammar in Table 1), but the actual tones and syllables used differed across participants.

The grammar used for constructing the learning and test items was the same as in Experiment 1. The learning sequence and test items used were nearly identical, except that two items from the list containing within-categorical violations and two containing cross-categorical violations were modified slightly. The test phase again consisted of three types of items: grammatical, ungrammatical within-

category violations, and ungrammatical cross-category violations.

Procedure The procedure was identical to the one undergone by the participants in Experiment 1.

Results and Discussion

If cross-categorical violations are easier to identify when presented within a single sensory modality, we would expect to see improved performance on the cross-categorical violations in Experiment 2, because the violations span perceptual categories but are within the same sensory modality (e.g., tone-syllable or syllable-tone). This finding would provide evidence in support of the shallow integration model. On the other hand, if cross-categorical violations are equally difficult to identify regardless of whether they are presented in a single or multiple sensory modality, we should see no evidence of learning for the cross-categorical items. This scenario would provide further support for domain-specific processing in statistical learning.

To test these possible outcomes, a series of t-tests were run on the data to ascertain if learning was greater than chance levels for the three types of test items. The means for each item type can be seen in Table 2. As is evident, learning was observed for the grammatical items ($t = 4.384$, $p < .001$) and for the within-categorical violation items ($t = 7.618$, $p < .001$) but not for the cross-categorical violation items ($t = 0.23$, $p > .8$).

The data from Experiment 2 replicate and extend the results seen in the previous experiment. Once again, learning was robust for grammatical items and ungrammatical items when the grammatical violation was present between two units of the same feature type (i.e., two tones or two syllables). However, when the violation appeared between a tone and a syllable, participants were unable to identify it as ungrammatical at levels above chance. Thus, the difficulty seen in Experiment 1 for individuals identifying grammatical violations in cross-modal situations extends to instances where the grammatical elements are in the same sensory modality, but in different perceptual categories.

General Discussion

The present studies investigated categorical integration in a statistical learning paradigm. Experiment 1 used visual and auditory elements in a single artificial grammar to investigate within-modal and cross-modal processing. Experiment 2 investigated how learning takes place when two distinct features within a single modality are employed. The findings were used to evaluate four models of multisensory integration, based on those recently applied to linguistic processing (McNorgan, Reid & McRae, 2011).

Taken together, the studies demonstrate that participants are capable of learning grammatical and within-categorical violations, but have difficulty with cross-categorical violations. The discrepancy in performance between within- and cross-category violations may be due to a tendency to focus first on within-category patterns, which may be adaptive. That is, it may be more useful to learn within-category associations at the expense of cross-category ones, assuming that only a limited amount of cognitive resources are available to detect violations. The reasons for this are currently unexplored, though several possible explanations exist. It is a possible that it is more cognitively efficient to look for patterns in stimuli that are more similar before trying to find rules in patterns that exist across domains. Perhaps participants would have shown learning if they had greater exposure to the cross-categorical patterns in the learning phase, which would support this claim. It is also possible that within-category associations are encountered more frequently or are more informative, though this possibility seems less likely given the infant literature showing that learning is enhanced when infants are given stimuli in multiple modalities (Lewkowitcz, 2004).

The two studies presented here provide initial evidence in support of a domain-specific model of multisensory integration, suggesting that people have difficulty integrating sensory input across perceptual domains. However, this finding is preliminary. Interestingly, this conclusion does not correspond to the conclusions in McNorgan et al.'s (2011) initial test of their linguistic model, in which they determined that the deep model of processing best accounts for linguistic categorization. Several reasons for this discrepancy may exist. First, it is possible that statistical learning is a functionally different process than linguistic processing, at least as assessed by the two different tasks used in our study and theirs. One major difference between our statistical learning task and their linguistic task is that in the McNogran et al. (2011) study, participants did not actually perceive stimuli in different modalities. Instead, they were presented with words that theoretically appealed to different sensory modalities. If processing operates differently in these two domains (linguistic and statistical learning), it is not unreasonable to assume that a test of linguistic categorization would yield a different pattern of results than a test of statistical learning.

A second explanation deals with the previously mentioned issue of exposure time. It is possible that learning would have occurred if participants had been given more exposure to the cross-categorical dependencies in the learning phase. If this were the case, then the shallow and deep integration models could be directly tested against each other by integrating multiple features of each sensory modality into a single grammar. By varying the amount of exposure time with such a grammar, it could be possible to determine whether learning associations across different sensory modalities differs in comparison to learning associations across different perceptual categories within the same modality.

An important issue for further study is how these processes work in infants and children, as it has implications for multisensory aspects of cognitive processing such as speech perception. Since speech processing is one of many

cognitive skills that is considered multisensory, especially for young infants (Rosenblum, 2008), it is necessary to determine if they are capable of detecting cross-category violations. Little is currently known about the developmental trajectory of multisensory learning in children. Other cognitive processes not systematically studied in these experiments may also be involved, such as to what extent attention is specifically deployed in the learning phase toward learning the within-modal versus the cross-modal associations. These are important issues for future study.

In summary, the present experiments indicate that statistical learning is a complex process with constraints present in categorization. Though people are capable of correctly identifying grammatical information and within-categorical violations, they have difficulty learning grammatical violations when the violation appears between elements from two different categories of information. These categories may be different sense modalities, or different features within the same modality. On the other hand, people are very skilled at identifying violations that occur within a single perceptual category. On the one hand these findings would appear to suggest a purely domain-specific view of multisensory statistical learning, in which sensory integration does not occur at all. On the other hand, there may be other factors not explicitly explored in the current experiment (e.g., exposure time, attention) that could instead make cross-modal statistical learning more amenable.

References

Altmann, G.T.M., Dienes, Z., & Goode, A. (1995). Modality independence of implicitly learned grammatical knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 899-912.

Conway, C.M., Bauernschmidt, A., Huang, S.S., & Pisoni, D.B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*, 356-371.

Conway, C.M., & Christiansen, M.H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 24-39.

Conway, C.M. & Christiansen, M.H. (2006). Statistical learning within and between modalities: Pitting abstract against stimulus-specific representations. *Psychological Science, 17,* 905-912.

Conway, C.M., Pisoni, D.B., Anaya, E.M., Karpicke, J., & Henning, S.C. (2011). Implicit sequence learning in deaf children with cochlear implants. *Developmental Science, 14,* 69-82.

Cree, G.S. & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning *chipmunk, cherry, chisel, cheese,* and *cello* (And many other such concrete nouns). *Journal of Experimental Psychology: General, 132* (2), 163-201.

Emberson, L.L., Conway, C.M., & Christiansen, M.H. (in press). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *Quarterly Journal of Experimental Psychology*.

Gómez, R.L. (2002). Variability and detection of invariant structure. *Psychological Science, 13,* 431-436.

Gómez, R.L. (1997). Transfer and complexity in artificial grammar learning. *Cognitive Psychology, 33,* 154-207.

Kim, R.S., Seitz, A.R. & Shams, L. (2008). Benefits of stimulus congruency for multisensory facilitation of visual learning. *PLoS One, 3*(1), e1532.

Leclercq, A. & Majerus, S. (2010). Serial order short term memory predicts vocabulary development: Evidence from a longitudinal study. *Developmental Psychology, 46* (2), 417-427.

Lewkowicz, D.J. (2004). Serial order processing in human infants and the role of multisensory redundancy. *Cognitive Processes, 5,* 113-122.

Manza, L & Reber, A.S. (1997). Representing artificial grammars: Transfer across stimulus forms and modalities. In D.C. Berry (Ed.), *How implicit is implicit learning?* (pp.73-106). New York, NY: Oxford University Press.

McGurk, H. & MacDonald, J. W. (1976). Hearing lips and seeing voices. Nature, 2 64 , 746-748.

McNorgan, C., Reid, J. & McRae, K. (2011). Integrating conceptual knowledge within and across representational modalities. *Cognition, 118,* 211-233.

Perruchet, P. & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences, 10*(5), 233-238.

Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6,* 855-863.

Robinson, C.W. & Sloutsky, V.M. (2007). Visual statistical learning: Getting some help from the auditory modality. Paper session presented at the meeting of Cognitive Science Society, Nashville, TN.

Rosenblum, L.D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science, 17* (6), 405-409.

Seitz, A.R., Kim, R., Wassenhoven, V., & Shams, L. (2007). Simultaneous and independent acquisition of multisensory and unisensory associations. *Perception, 36*(10), 1445-1453.

Seger, C.A. (1994). Implicit learning. *Psychological Bulletin, 115 (2),* 163-196.

Zacks, J.M. & Swallow, K.M. (2007). Event segmentation. *Current Directions in Psychological Science, 16* (2), 80-84.