Memory and learning with rapid audiovisual sequences

Arielle S. Keller
Volen Center for Complex Systems, Brandeis University, Waltham, MA, USA

Robert Sekuler
Volen Center for Complex Systems, Brandeis University, Waltham, MA, USA

We examined short-term memory for sequences of visual stimuli embedded in varying multisensory contexts. In two experiments, subjects judged the structure of the visual sequences while disregarding concurrent, but task-irrelevant auditory sequences. Stimuli were eight-item sequences in which varying luminances and frequencies were presented concurrently and rapidly (at 8 Hz). Subjects judged whether the final four items in a visual sequence identically replicated the first four items. Luminances and frequencies in each sequence were either perceptually correlated (Congruent) or were unrelated to one another (Incongruent). Experiment 1 showed that, despite encouragement to ignore the auditory stream, subjects' categorization of visual sequences was strongly influenced by the accompanying auditory sequences. Moreover, this influence tracked the similarity between a stimulus's separate audio and visual sequences, demonstrating that task-irrelevant auditory sequences underwent a considerable degree of processing. Using a variant of Hebb's repetition design, Experiment 2 compared musically trained subjects and subjects who had little or no musical training on the same task as used in Experiment 1. Test sequences included some that intermittently and randomly recurred, which produced better performance than sequences that were generated anew for each trial. The auditory component of a recurring audiovisual sequence influenced musically trained subjects more than it did other subjects. This result demonstrates that stimulus-selective, task-irrelevant learning of sequences can occur even when such learning is an incidental by-product of the task being performed.

Introduction

Agus, Thorpe, and Pressnitzer (2010) and Gold, Aizenman, Bond, and Sekuler (2014) took parallel approaches to study auditory and visual short-term memory. In each study, test subjects judged whether the temporal structure of the second half of a stimulus was the same as or different from the temporal structure of the first half. In both studies, the halves of each stimulus were presented back-to-back with no interruption between, and the resulting judgments provided an indirect measure of memory. Agus and colleagues' (2010) auditory stimuli comprised sequences of randomly generated Gaussian auditory noise; Gold and colleagues' (2014) stimuli were sequences of eight randomly generated luminances delivered rapidly to the center of a display. Both groups of researchers noted substantial, though unexplained, differences among subjects' performance levels. Additionally, by intermittently recycling particular stimulus exemplars, each research group found strong evidence of what has been termed "incidental learning" (Kelly, Burton, Kato, & Akamatsu, 2001; McGeoch & Irion, 1952; Seitz & Watanabe, 2009). This task-irrelevant, stimulus-selective improvement in performance was more rapid for auditory than for visual stimuli, although performance for each kind of stimulus was marked by substantial individual differences.

For the experiments presented here, we adapted Gold and colleagues' (2014) task and stimuli, embedding them in a multisensory context of concurrent synchronous sequences made up of varying luminances and tones. We had several reasons for this choice. First, we believed that sequences in each modality would lend themselves to systematic examination of structural similarity, a variable strongly implicated in memory formation and in learning (Dubé, Zhou, Kahana, & Sekuler, 2014; Kahana, 2012). Second, concurrent stimuli would make it possible to manipulate the similarity between temporally correlated auditory and visual sequences. With subjects instructed to attend to and judge only the luminances in a sequence, we expected that intermodal similarity might overwhelm selectivity of attention, allowing the task-irrelevant auditory sequences to impact subjects' judgments.

Citation: Keller, A. S., & Sekuler, R. (2015). Memory and learning with rapid audiovisual sequences. *Journal of Vision, 15*(15):7, 1–18. doi:10.1167/15.15.7.

doi: 10.1167/15.15.7 Received March 22, 2015; published November 17, 2015 ISSN 1534-7362

Third, the robust learning expected with concurrent, temporally correlated audiovisual stimulation (Shams & Seitz, 2008; Thelen, Matusz, & Murray, 2014) might allow us to examine subsequent unlearning and relearning, phenomena that afford unique perspectives on memory and learning. Fourth, the paradigm allowed us to probe the influence of prior musical training, which is known to powerfully affect processing of various kinds of auditory signals (Skoe & Kraus, 2012). Finally, Gold and colleagues' (2014) subjects were instructed to judge whether the last four luminances in a trial's stimulus sequence replicated the sequences first four luminances. These instructions suggest that subjects should separately encode and remember the first four items in a sequence so that each can be compared against the corresponding item in that sequence's second half. However, Gold and colleagues (2014) found that subjects' judgments might have been based on ensemble or summary representations (Alvarez, 2011; Dubé et al., 2014) rather than on the individual items in a sequence. We wanted to verify that result, extend it to the task-irrelevant sequences of tones, and use it to probe the origin of individual differences.

Experiment 1

Experiment 1 measured how task-irrelevant, nominally ignored auditory sequences influence processing of attended, concurrent visual sequences.

Stimuli

On each trial, a subject saw an eight-item sequence of quasi-random luminances, presented at 8 Hz. These luminances were delivered one after another, without interruption, to a $4.1° \times 4.1°$ region centered on a computer display. On approximately half of the trials, the order of the final four luminances in a sequence identically repeated the first four luminances (hereafter, Repeat stimuli). On remaining trials, the final four luminances were generated independently of the first four (hereafter, Non-Repeat stimuli). A tone sequence whose members varied in frequency was concurrent and synchronous with items in each luminance sequence.

To generate the visual stimuli for each trial, eight random samples were drawn from a normal distribution, $N(0.0, 0.2)$. Samples $\geq 2\sigma$ from the mean were discarded and replaced by new samples. Next, the eight samples were translated into luminances using a lookup table that linearized the display. With the display's background luminance fixed at 19.03 cd/m^2, this process generated luminance samples ranging from 2 to 42 cd/m^2. Together with the distribution's relatively small standard deviation, censoring extreme values homogenized items that comprised each stimulus sequence. This would reduce the value of individual highly-distinctive, "oddball" items as the basis for subjects' judgments.

Visual stimuli were presented on a Sony Trinitron UltraScan P780 CRT monitor (Sony Electronics, Tokyo, Japan) whose resolution was set to 1024×768 pixels. The display refresh rate was 75 Hz. Stimuli were generated and presented by an Apple iMac running MATLAB (version 7.7), along with extensions from the Psychophysics Toolbox (Brainard, 1997). A calibration-based lookup table linearized the luminances of visual stimuli. Viewing was binocular through natural pupils at a distance of 57 cm, which was enforced by means of a chin support. The computer's display provided the only source of illumination in the room.

Each auditory sequence comprised a seamless stream of eight, equal-duration pure tones, each approximately 133 ms in duration. Tones were sampled at 44.1 kHz and presented binaurally at 70–72 db(A) through Sennheiser HD280 supraaural earphones (Sennheiser, Wedemark, Germany). Audible transients that would arise from abrupt changes in frequency from one tone to another were eliminated by tapering the leading and trailing edges of each tone with a raised cosine.

In order to generate multisensory sequences whose components would be perceptually correlated (hereafter, Congruent stimuli), we drew on some cross-modal matching results from Marks (1974). In Congruent stimuli, the frequency of each tone was a monotonically increasing function of the luminance that it accompanied. Equation 1 presents the mapping used to generate stimulus frequencies:

$$\text{Hz} = 56.297L^{0.6228} \quad (1)$$

in which L is luminance in cd/m^2.

Although individual subjects' audiovisual matches might not all align perfectly with this particular monotonic function (Marks, 1974), to simplify the intended analyses, this single function generated auditory stimuli for all subjects. For luminances in the stimulus range we used, 2 to 42 cd/m^2, Equation 1 generated auditory stimuli that range from a low of 87 Hz to a high of 577 Hz (or, in musical terms, from F2 to approx. D5).

To generate stimuli whose visual and auditory sequences were not correlated, (hereafter, Incongruent stimuli), a second set of eight luminance samples was drawn from an $N(0.0, 0.2)$ distribution. Then, frequency equivalents to items in this second set were found and substituted for the original set of frequencies. The result was a set of frequencies not correlated with the original set of luminances. Throughout our

	Audiovisual relationship	
	Congruent	Incongruent
Visual repeat?		
Repeat	Rcon	Rincon
Non-Repeat	Ncon	Nincon

Table 1. Names used for stimulus conditions.

experiments, the subjects' task was to judge whether the last four items in a visual sequence replicated the first four (a Repeat stimulus) or did not (a Non-Repeat stimulus), while disregarding the accompanying auditory sequence.

Subjects

Fifteen subjects between 18 and 21 years old were tested. All had Snellen visual acuity of at least 20/40, and clinically normal hearing as defined by pure tone thresholds at 0.25, 0.5, 1, 2, 4, and 8 kHz of at least 25 dB (Mueller & Hall, 1998). Each received $10 (U.S.) for participation. In a version of the task used here, subjects' performance was found to be correlated with their musical training. Specifically, Aizenman, Gold, and Sekuler (2013) found that subjects' ability to detect repetitions of four-item patterns in visual sequences as well as with auditory sequences was related to subjects' musical training. To minimize the intrusion of such effects here, we recruited subjects whose musical training fell between the extremes represented in Aizenman and colleagues' (2013) two groups. Using criteria adapted from Skoe and Kraus (2012), subjects either (a) had played an instrument for 6 years or less but were not currently playing, or (b) had played for more than 6 years, but were not currently playing, or (c) had played for less than 3 years.

Procedure

Repeat and Non-Repeat visual sequences were presented equally often, and in random order. In addition, the auditory accompaniment to a visual sequence was either Congruent with items in the visual sequence (the frequency of a tone was monotonically related to the accompanying visual item's luminance), or Incongruent with items in the visual sequence (tone frequencies were independent, nonrepeating random samples). Sequences were generated anew for each trial. By crossing two types of visual sequences, Repeat and Non-Repeat, with two types of congruence, Congruent and Incongruent, we produced four stimulus types. Table 1 gives the scheme used to designate the stimulus types. When referring to any one of these four types, an upper case R or N designates the type of *visual* sequence in the stimulus. Thus, an upper case R designates a stimulus whose visual sequence is Repeat, while an upper case N designates a stimulus whose visual sequence is Non-Repeat. The congruency or incongruency between a sequence's *auditory* component and its visual component is designated by the suffix "con" or the suffix "incon" appended to R or N. Thus, Rcon, for "Repeat Congruent," designates a stimulus whose visual components include a repeated sequence, *and* whose auditory components are perceptually congruent to the visual. Correspondingly, we refer to each of the other three stimulus types as Repeat Incongruent (Rincon), Non-Repeat Congruent (Ncon), or Non-Repeat Incongruent (Nincon).

In the first and fourth blocks of trials, only those four stimulus types were presented equally often and in random order. In the second and third blocks, a new type of stimulus, which we call Nrep, was added to the mix. As the schematic example in Figure 1 shows, in such an Nrep stimulus, a Non-Repeat visual sequence was accompanied by an independently generated Repeat auditory sequence; that is, an auditory sequence whose final four items duplicated its first four. These Nrep stimuli tested whether a Repeat sequence within the nominally ignored, auditory stream influenced subjects' judgment of whether a Repeat had occurred within the attended, visual stream. We hypothesized that these Nrep stimuli would attract more false positives; that is, erroneous Repeat judgments, than would Nincon stimuli. Table 2 summarizes the names given to various classes of stimuli, and identifies for each, which dimension—visual, auditory, both, or neither—repeats within exemplars of that class.

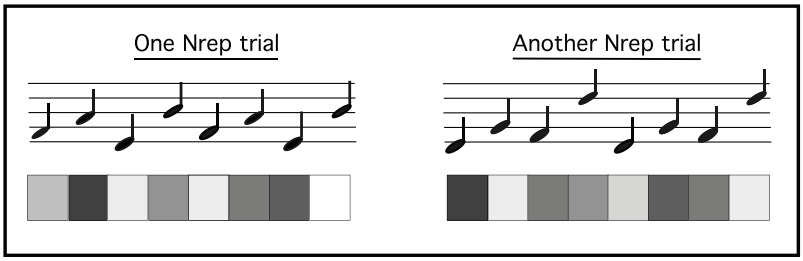

Figure 1. Schematic examples of Nrep stimuli presented on different trials.

Experiment 1	Experiment 2	Repeat modality
Rcon	Rcon	Visual, auditory
Rincon	Rincon	Visual
Ncon	Ncon	–
Nincon	Nincon	–
Nrep	–	Auditory
–	FRcon	Visual, auditory
–	FRincon	Visual

Table 2. Classes of stimuli in Experiments 1 and 2.

Subjects were informed that in some sequences, the last four brightness levels would repeat the first four identically, and that they should categorize such sequences as Repeat. Subjects were also told that, in some sequences, the last four brightnesses would not replicate the first four, and that these sequences should be categorized as Non-Repeat.

To promote their understanding of the task, subjects were shown diagrams containing schematic exemplars of both types of sequences, and were asked to categorize each according to the instructions just described. Again, subjects were shown printed schematic exemplars, this time with musical notes at various heights on the musical staff accompanied by squares of various luminance levels. Additional instructions emphasized that subjects should ignore whatever tones accompanied the visual sequences, judging only the visual sequences. They were then tested for their understanding of this crucial part of the task. Following this test, a subject received 24 practice trials (an equal mix of Repeat/Non-Repeat and Congruent/Incongruent), each followed by immediate feedback about response correctness. During the experiment, subjects signaled their judgments by pressing one of two keys on the keyboard corresponding to "Repeat" or "Non-Repeat" sequences. Immediate feedback after each response conveyed whether the response had been correct or not.

Each subject served in 600 trials, distributed over four blocks. Blocks 1 and 4 comprised 140 trials each, and to accommodate trials needed for Nrep stimuli, Blocks 2 and 3 comprised 160 trials each. To maintain their interest and motivation, during short breaks between blocks, interesting facts about the brain were presented to subjects on the computer display (Anguera et al., 2013).

Results and discussion

Influence of irrelevant auditory stimuli

We asked first whether the task-irrelevant, auditory sequence of the audiovisual stimuli influenced subjects' judgments of items in the accompanying visual sequence. One answer was provided by performance with Nrep stimuli. These stimuli, in which auditory but not visual sequences contained a Repeat sequence, attracted a great many incorrect judgments; that is, the visual sequence was misjudged as Repeat. Specifically, Nrep stimuli elicited correct responses significantly less often than any other stimulus type did (all ps < 0.01, HSD, $0.38 < d < 0.9$). In fact, Nrep stimuli drew incorrect proportion (Repeat) responses on approximately half of all trials ($M = 0.54$, $SD = 0.12$), a value not significantly different from chance, $p = 0.50$; $t(14) = 1.653$, $p = 0.121$, $d = 0.62$.

Having seen that task irrelevant auditory sequences influenced judgments of accompanying visual sequences, we next examined possible differences between Congruent and Incongruent sequences. Such differences would also show the possible impact of task-irrelevant auditory stimuli. We calculated the mean *pr* (Repeat) values for each of the four main stimulus types, converted those values to standard scores, and found the difference between standard scores for Rcon and Ncon stimuli, and also between standard scores for Rincon and Nincon stimuli. Figure 2A shows these differences as d' values for Congruent and Incongruent sequences. Congruent stimuli produced higher mean d' values than did Incongruent stimuli, $M = 1.57$, $SD = 0.57$ and $M = 0.93$, $SD = 0.27$, respectively. A matched samples t test showed that this difference was statistically significant, $t(14) = 5.124$, $p < 0.001$, $d = 1.32$. This outcome, which replicated previous results with similar stimuli (Aizenman et al., 2013), confirms that although auditory sequences were task-irrelevant and nominally ignored, they nevertheless influenced how visual sequences were categorized as Repeat or Non-Repeat.

Potential strategies for judgments

A one-way ANOVA compared the *pr* (Repeat) responses associated with each of four stimulus types, Rcon, Rincon, Ncon, and Nincon. The ANOVA showed that overall differences among stimulus types were statistically significant, $F(3, 59) = 10.47$, $p < 0.001$, $\eta^2 = 0.36$. Post hoc comparisons revealed that part of this effect reflected the high *pr* (Repeat) responses made with Rcon stimuli, $M = 0.82$, $SD = 0.13$; all ps < 0.01, by Tukey's HSD test, $1.37 < d < 1.88$. The reliable difference between performance with Rcon and Rincon stimuli, $t(14) = 4.51$, $p < 0.001$, $d = 1.16$, is important as those stimulus types differ only in congruence between the sequence's auditory and visual components. As a result, the difference seems to suggest that audiovisual congruence influences the accuracy with which the sequence is categorized. This result is unsurprising; after all, the importance of spatial or temporal congruence has been demonstrated previously in many tasks (Spence, 2011).

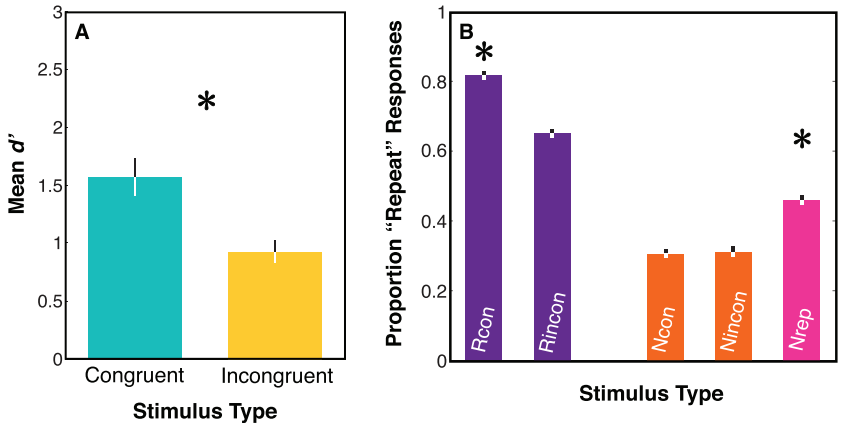

Figure 2. (A) Values of d' produced with Congruent and Incongruent stimuli. The symbol * signifies that performance with Congruent stimuli significantly exceeds performance with Incongruent stimuli. (B) Mean proportion of Repeat responses for each stimulus type. Note that for Rcon and Rincon stimuli, a Repeat response was correct, while for the other three stimulus types a Repeat response was incorrect. The leftmost * symbol indicates that performance is significantly higher with Rcon stimuli than with Rincon stimuli; the rightmost * symbol signifies that performance with Nrep stimuli is significantly worse than all other stimulus types with the exception of Rincon. Error bars represent between-subjects standard errors.

It is interesting to note that not every comparison between pairs of conditions was consistent with the idea that audiovisual congruence was important to performing our task. Consider the Ncon and Nincon conditions. These conditions differed only by whether the audio and visual sequences were congruent or not. Yet, they produced levels of performance that were essentially indistinguishable from one another, $t(14) = 0.71$, $p = 0.49$, $d = 0.18$. This result made us consider the possibility that subjects might have been influenced by some variable other than audiovisual congruence when attempting to tell Repeat from Non-Repeat sequences. One candidate for such a variable is the number of Repeat sequences contained in the stimuli. To appreciate this possibility, note that Rcon and Rincon stimuli differ in more than just one way. They differ of course according to whether tones in the auditory sequence are perceptually correlated with the luminances in the visual sequence or not, but they differ also in the total number of Repeat sequences contained in each type of stimulus, as Table 3 shows. Thus, while an Rcon stimulus comprises two different Repeat sequences (one visual and one auditory), an Rincon stimulus comprises only one Repeat sequence (visual). We asked whether Rcon's superiority over Rincon might have arisen from the presence of two distinct sources of Repeat information in Rcon, rather than from the perceptual congruence in that condition. The hypothesis that the number of Repeat information sources is more influential than audio-visual congruence in this task would also explain the similarity in performance with Ncon and Nincon stimuli, as these stimuli have equal numbers of Non-Repeat information sources.

To test this possibility we focused on Rincon stimuli, evaluating each one according to how closely its random sequence of Non-Repeat auditory tones approximated a Repeat sequence. For the purposes of our evaluation, we computed a variable that we call "repeatedness." For the auditory sequence in an Rincon stimulus, repeatedness is defined as the summed absolute differences in frequency between tone n and tone $n+4$, for $n = 1{:}4$. The resulting set of repeatedness values computed for all Rincon stimuli were sorted from smallest to largest, and divided into 10 equally populous bins. For the purposes of plotting, these repeatedness values were normalized such that, if a sequence had actually repeated, the value of its repeatedness variable would have been one. Finally, for

Stimulus type	Visual repeat?	Auditory repeat?	Repeat sequences
Rcon	Yes	Yes	2
Rincon	Yes	No	1
Ncon	No	No	0
Nincon	No	No	0
Nrep	No	Yes	1

Table 3. Repeat sequences contained in audiovisual stimuli in Experiment 1.

Model	Estimate	Standard error	*z*-value	*pr*(>|z|)
(Intercept)	0.25	0.32	0.79	0.43
Visual	−0.19	0.07	−2.59	0.0097**
Auditory	−0.03	0.02	−1.46	0.14
Visual:Auditory	0.002	0.005	0.47	0.64

Table 4. Logistic regression of summary statistic for Experiment 1. *Note*: $^{**}p < 0.01$.

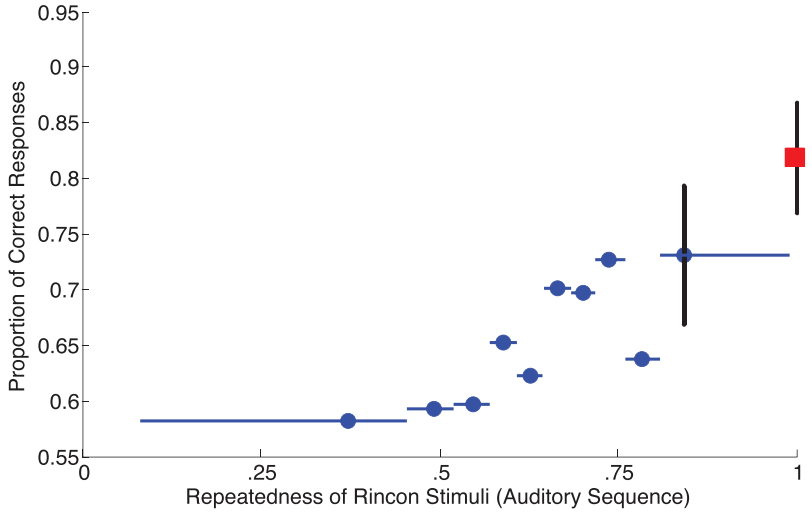

Figure 3. Proportion of correct responses in Experiment 1 with Rincon stimuli of varying degrees of auditory "repeatedness" in 10 equally populous bins. The degree of repetition in the auditory component of a Rincon stimulus is calculated by taking the mean difference between items n and $n + 4$ for $n = 1{:}4$. The resulting value was normalized so that a repeatedness value of 1 would correspond to a perfect repeat. Vertical bars represent 95% confidence intervals, while horizontal bars show the range of values in a bin. Also plotted (■) is the mean proportion of correct responses with Rcon stimuli. The overlap in confidence intervals between the Rcon stimuli and the bin of Rincon stimuli closest to perfectly repeated provides evidence for the claim that the effect of congruency can be accounted for by the additional Repeat information in the auditory component.

each bin, we found the proportion of Repeat responses made to stimuli in that bin.

Although the auditory sequence in an Rincon stimulus was never actually congruent with its accompanying visual sequence, subjects' Repeat judgments tracked the repeatedness of the task-irrelevant auditory sequences. This relationship is shown by the data points in Figure 3. A logistic regression on individual Rincon trials confirmed that the *pr* (Repeat) judgments was related to the mean repeatedness of a stimulus' auditory sequence ($p < 0.0001$). The red square in the figure represents the mean performance with Rcon stimuli; that is, stimuli in which both auditory and visual sequences repeated. Note that the 10th bin includes Rincon trials whose auditory sequences most closely approximated a genuine Repeat. The 95% confidence intervals around the result with Rcon stimuli (red square) overlapped the confidence intervals around the mean for trials in the 10th Rincon bin, which signifies that mean performance with the 10th bins' Rincon stimuli did not differ significantly from mean performance with Rcon stimuli (red square). The similarity between these two proportions of Repeat judgments suggests that the difference observed between Rcon and Rincon stimuli could have resulted from repeatedness of the auditory sequence. This suggests that a process akin to aggregation of Repeat information across both modalities could explain the difference between Congruent and Incongruent stimuli. This outcome suggests that congruency, defined by a cross-modal relationship between luminance levels and auditory frequencies, by itself did not matter for performance with this task. However, as this cross-modal relationship was held constant across subjects rather than defined by individual subjects' matching behavior, we cannot rule out the possibility that cross-modal relationships tailored to individual subjects might have affected performance even more strongly. However, it is clear that in our task the presence of Repeat items within an auditory sequence does lure subjects into categorizing a sequence's visual component as Repeat.

Ours is certainly not the first study to demonstrate that a task-irrelevant auditory stimulus alters responses to a concurrent or near-concurrent visual stimulus (e.g., Rosenthal, Shimojo, & Shams, 2009; Sekuler, Sekuler, & Lau, 1997; Shams, Kamitani, Thompson, & Shimojo, 2001; Zhou, Wong, & Sekuler, 2007). In such studies, the mere presence of the task-irrelevant stimulus was sufficient for its effect to be seen. By contrast, the auditory stimuli in Experiment 1 apparently had to undergo cognitively complex processing in order for them to impact subjects' judgments. Specifically, Figures 2 and 3 suggest that the influence of an auditory sequence requires that it be processed sufficiently to extract the sequence's repeatedness.

Subjects' instructions implicitly encouraged them to make item-by-item comparisons between corresponding items in a sequence's two halves (i.e., to compare item n to item $n + 4$ for items $n = 1{:}4$). Of course, no matter what strategy is implied by an experimenter's instructions, subjects may not follow it. This could be

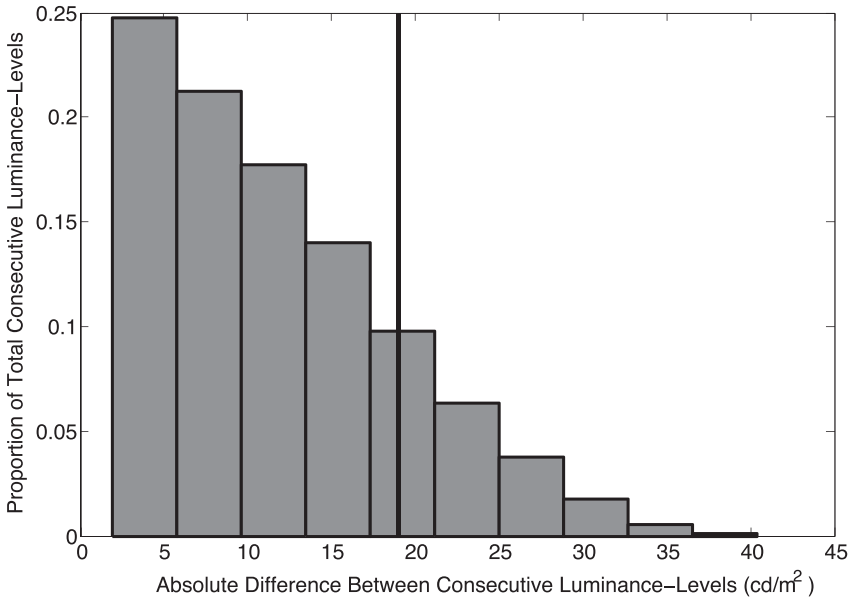

Figure 4. Distribution of absolute differences between the luminances of successive items for Nincon trials in Experiment 1, divided into 10 equally spaced bins. The vertical line represents the background luminance, 19.03 cd/m^2.

because too many stimuli are presented, because stimuli are presented too rapidly, or because subjects hit upon some alternative strategy. In our paradigm, the rate at which relatively homogeneous items succeeded one another made item-by-item comparisons difficult. Additionally, making such comparisons would have challenged working memory by requiring that subjects perform a complex variant of the *n*-back task (Jaeggi, Buschkuehl, Perrig, & Meier, 2010). Moreover, item-by-item comparisons might not be required for some measure of success in our task. For example, Warren and Bashford (1993) showed that subjects could judge the order of items in auditory sequences of rapidly presented phonemes or tones without actually segmenting each sequence into an ordered series of items. Instead, subjects seemed to form and base judgments on perceptual compounds extracted from the sequences. This led us to conjecture that rather than make a series of item-by-item comparisons, subjects might have adopted some alternative, shortcut strategy.

To test this conjecture, we examined two potential strategies that subjects might have adopted. First, we asked whether subjects might have based their judgments on some especially salient feature within a given sequence, detecting that feature in the first part of a sequence and then assessing whether that feature repeated in the latter part of the sequence. Specifically, we asked if performance might be related to the presence of some unusually large difference between consecutive items in a sequence (Pollack, 1956). Figure 4 shows the distribution of differences between successive items' luminances for Nincon sequences. Although the generating algorithm produced a narrow range of individual luminances (2 to 42 cd/m^2), some differences between successive items would have been several times the expected discrimination threshold (Graham & Kemp, 1938). As such differences might have been particularly salient, we hypothesized that they might be useful as short-term memory cues. To test this possibility, a logistic regression analysis compared performance on each Nincon sequence to the largest absolute difference between consecutive items in a sequence. Disconfirming the hypothesis, the logistic analysis showed nonsignificant effects for both visual ($p = 0.48$) and auditory sequences ($p = 0.24$).

Having seen that performance was not tied to the size of a sequence's largest successive difference, we turned to another possible cue that subjects might have relied upon, namely, some summary or ensemble representation of items in a sequence (Alvarez, 2011; Dubé & Sekuler, 2015; Rosenholtz, Huang, Raj, Balas, & Ilie, 2012). Subjects seem to exploit various summary statistics when a variable such as rapidity of presentation makes it difficult to process individual components of a stimulus. Such statistics capture the gist of a stimulus, while sacrificing its details. We examined the possibility that subjects summed a sequence's first four items, summed its last four, and based a response on the absolute difference between the two sums. Adapting Sorkin's (1962) differencing model for same–different judgments, we assume that the difference between sums from two halves of a stimulus is compared to a criterion. If the difference between the two aggregates exceeds that criterion, the stimulus will be categorized as Non-Repeat, otherwise, as Repeat. Although the combination of this summary statistic and the differencing operation should apply for all our stimulus types, the structure of Nincon stimuli makes them the

most amenable for evaluating predictions from this strategy. Making or not making a false positive ("Repeat" judgment) to any Nincon stimulus is a binary variable. Therefore, we applied a logistic regression to the proposed model, using aggregations computed solely in the visual domain, solely in the auditory domain, or in both domains simultaneously. The results of this logistic regression, shown in Table 4, suggest that subjects' judgments are consistent with a differencing model applied to a summary statistic, but only for visual sequences ($p < 0.01$). It is interesting to note that neither auditory sequences nor the combination of visual and auditory sequences support such behavior ($p = 0.14$ and $p = 0.64$, respectively). This difference between vision and audition might reflect the fact that our task instructions prioritized visual information. Alternatively, it might reflect the fact that audition's superior temporal acuity (Julesz & Hirsch, 1972) made it unnecessary for subjects to use a summary statistic to represent auditory sequences.

Experiment 2

The subjects in Experiment 1 performed a task that can be described as a form of short-term memory. The task required subjects to compare their memory of each sequence's first four visual items against some representation of the trial's last four visual items. Throughout the experiment, all stimuli were generated afresh on each trial. Because subjects never saw precisely the same visual (or auditory) sequence, the results were mute about the durability of information that subjects acquired and used on any trial. A previous experiment with unisensory visual stimuli and a task like Experiment 1's (Gold et al., 2014) showed that the short-term memory from one trial could cumulate over trials. That experiment's design was modeled on one devised by Hebb (1961) to examine the robustness and vulnerability of short-term memory. Hebb assessed subjects' short-term memory for quasi-random sequences of nine digits, and discovered that performance improved steadily when precisely the same digit sequence recurred periodically. Remarkably, his subjects learned the single repeated sequence despite its having been embedded with many other different sequences. Results with Hebb's repetition design (Gold et al., 2014; Horton, Hay, & Smyth, 2008; Page & Norris, 2009; Stadler, 1993) encouraged us to ask, in Experiment 2, whether similar repetition-based learning would be seen with multisensory, audiovisual stimulus sequences. In addition, Experiment 2 was designed to contrast audiovisual integration and learning in musically trained and nontrained subjects. The theoretical value of this contrast rested on two lines of previous research. First, musical training has been shown to sharpen temporal processing, which should impact subjects' ability to parse any rapidly presented audiovisual sequences (Lee & Noppeney, 2011; Lu, Paraskevopoulos, Herholz, Kuchenbuch, & Pantev, 2014) and second, musical training has been shown to strengthen some forms of audiovisual interaction (Paraskevopoulos, Kuchenbuch, Herholz, & Pantev, 2012, 2014; Strait, Parbery-Clark, Hittner, & Kraus, 2012). Finding that musical training affects performance on our task would be a valuable outcome given multisensory learning's importance in everyday life (Shams & Seitz, 2008) and the pervasive benefits associated with musical training (Skoe & Kraus, 2012).

Working with unisensory, visual sequences identical to the visual components of our multisensory sequences, Gold et al. (2014) demonstrated stimulus-selective incidental learning. As particular visual sequences recurred intermittently, interspersed among many other nonrecurring stimuli, subjects grew more adept at judging whether the visual items within that sequence repeated or not. Note that the subjects' task was defined exclusively in terms of within-trial comparisons between sets of items comprising an individual visual sequence; they did not judge whether a sequence had been seen before or not. So, Gold and colleagues' (2014) subjects had "no specific motive or a formal instruction" to treat the intermittently presented visual sequences as anything other than one among dozens of other similar visual sequences (McGeoch & Irion, 1952). However, multiple exposures to particular visual sequences allowed subjects to learn to categorize those stimuli more accurately, a form of unsupervised statistical learning (Fiser & Aslin, 2001).

We wanted to know whether subjects who had been instructed to detect repetition within the visual component of an audiovisual sequence do better as a result of multiple encounters with the same, recurring exemplar. In order to capture stimulus-selective learning, we introduced two new classes of stimuli, FRcon and FRincon. Both included Repeat visual sequences, and were made to recur intermittently and randomly. We describe these sequences as "frozen," after that term's use in auditory psychophysics to refer to reproducible noise stimuli (Guttman & Julesz, 1963; Pfafflin, 1968; Pfafflin & Matthews, 1966).

In Experiment 2, frozen stimuli were either particular, randomly generated Rcon stimuli (which we call FRcon stimuli), or particular, randomly generated Rincon stimuli (which we call FRincon stimuli). Pairs of yoked subjects shared exactly one FRcon and one FRincon exemplar; these exemplars varied between yoked subject pairs. Each of these exemplars was presented on 100 trials, interspersed among randomly generated, nonfrozen stimuli. Figure 5 shows schematic examples of the visual sequences on Non-Repeat,

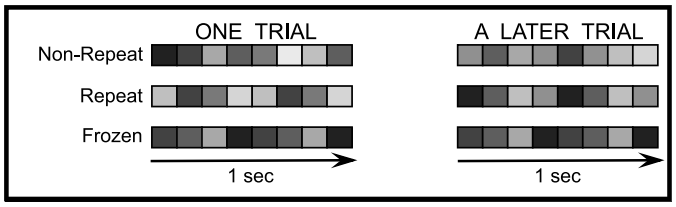

Figure 5. Examples of visual components of Non-Repeat, Repeat, and Frozen stimuli. See text for details.

Repeat, and Frozen trials. On average, 5.75 trials intervened between successive occurrences of FRcon stimuli, and 5.82 trials between occurrences of FRincon stimuli. Additionally FRcon stimuli and FRincon stimuli appeared on successive trials only 0.17% of the time.

Experiment 1 showed that, although nominally ignored, auditory sequences can influence judgments of concurrent visual sequences. Therefore, we hypothesized that judgments of recurring, frozen visual sequences would be influenced by concurrent auditory sequences. To test this hypothesis, we altered the auditory component of a FRincon sequence partway through the experiment but left the visual component intact. This maneuver, which was applied to half the subjects, would reveal whether what subjects had learned before the change included some representation of the auditory sequence. In addition, after these subjects had been exposed to these changed conditions, we reinstated their original stimuli. This reinstatement, or switchback, gauged how well subjects' learning had been preserved in the face of intervening exposure to altered conditions (Cohen, Ivry, & Keele, 1990, experiments 3 and 4).

Subjects

Twenty-four subjects, ages 18 to 23 years, took part. All had normal vision and hearing, measured as in Experiment 1. Each was paid $10 for participation. According to criteria formulated by Skoe and Kraus (2012), 14 of the subjects had considerable musical training, while 10 had little or none. Without intending any commitment to subjects' actual musical prowess, we will use a terminological shorthand for the two groups, calling one "Musicians" and the other "Non-Musicians."

Half of each group's subjects were tested in what we call the Crossover-Switchback condition. In that condition, the task-irrelevant auditory component of the FRincon sequence was changed partway through the experiment and then changed back. The remaining subjects were tested in what we call the Constant condition (Piantadosi, 2005); for them, the same FRincon stimulus was maintained throughout the

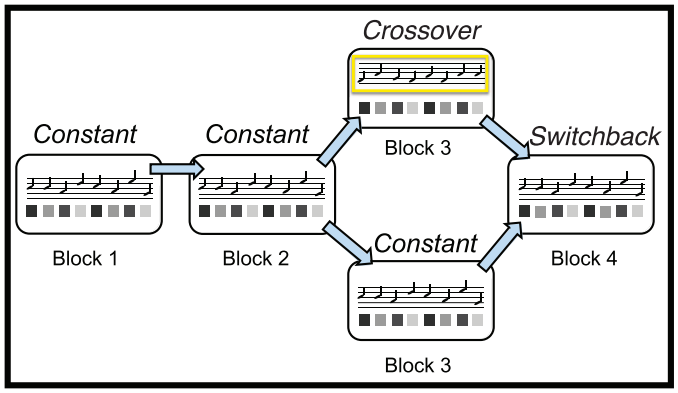

Figure 6. Diagram showing exemplars of Frozen Incongruent sequences. Each subject's Frozen Incongruent stimulus (FRincon) recurred identically and intermittently throughout the experiment. A subject's FRincon stimulus was changed only in Block 3 and only for subjects in the Crossover-Switchback condition. Each FRincon stimulus comprised a Repeat visual sequence accompanied by an incongruent auditory sequence. For subjects assigned to the Crossover-Switchback condition, the sequence's auditory component was switched in Block 3 to a different, incongruent auditory sequence. In contrast, subjects assigned to the Constant condition continued to receive the same FRincon sequence throughout.

experiment. This change was introduced in the third block of trials, and then removed (switched back) for the fourth and final block, as shown in Figure 6.

Pairs of subjects in each condition were yoked according to their group status, Musicians or Non-Musicians. One subject in each yoked pair was assigned to the Constant condition, while the other was assigned to the Crossover-Switchback condition. This yielded seven Musicians and five Non-Musicians in each condition. In all other respects, subjects in each yoked pair received the same stimulus sequences presented in the same order. Those sequences and orders differed between yoked pairs. For all subjects, a block of trials comprised 150 trials. Throughout, all subjects' had the same task as in Experiment 1: to judge whether the sequence of eight varying luminances was composed of two repeated sets of varying luminances, while ignoring the accompanying auditory stimuli.

Stimuli and procedure

Within a block of 150 trials, six different stimulus types were presented, each on 25 trials. The order in which stimuli were presented was randomized anew for every block and pair of yoked subjects. Between blocks of trials, subjects were allowed a short break, and were shown several "fun facts" about the brain, as in Experiment 1.

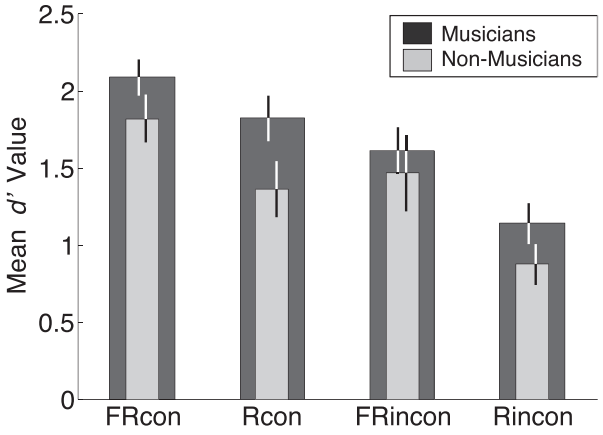

Figure 7. Performance (expressed as mean d' values) with Frozen and Repeat stimuli, in Congruent and Incongruent conditions. As in Experiment 1, performance with Congruent stimuli exceeds performance with Incongruent stimuli. Moreover, performance with Frozen exemplars exceeds performance with Repeat stimuli, and Musicians outperform Non-Musicians on average for all stimulus types. Error bars represent between-subjects standard errors.

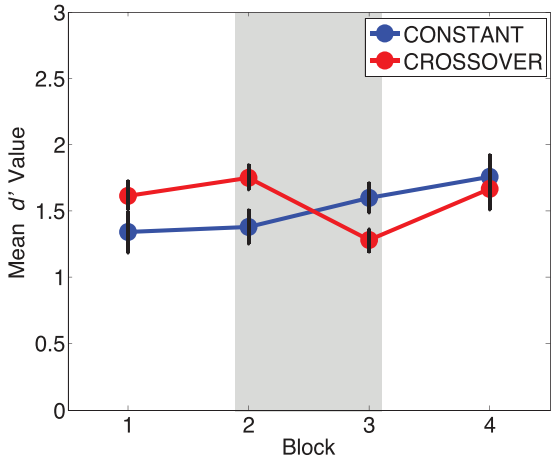

Figure 8. Performance with FRincon stimuli by subjects in the Constant and Crossover-Switchback groups. While subjects in the Constant condition show a small, statistically nonsignificant change in d' over the experiment's four blocks, subjects in the Crossover-Switchback group show a sharp decrease in Block 3. Error bars represent between-subject standard errors.

As mentioned above, half the subjects in each yoked pair were assigned to the Constant condition, while the other half were assigned to the Crossover-Switchback condition. This arrangement yoked subjects (one in each condition) who shared the same musicianship status (either Musician or Non-Musician). Figure 6 shows schematically how FRincon stimuli were manipulated in those conditions. In the Constant condition, on 25 of the 150 trials in every block, a subject was presented with exactly the same exemplar Frozen Incongruent stimulus (FRincon). Subjects in the Crossover-Switchback condition and subjects in the Constant condition were treated identically except for the third block of trials. In that block, subjects in the Crossover-Switchback condition received a FRincon stimulus comprising the very same visual sequence as in previous blocks, but now that visual sequence was accompanied by a *different* auditory sequence on each of the 25 presentations in that third block.

Results

Changes to frozen auditory sequences are disruptive

Figure 7 shows that, overall, Musicians outperformed Non-Musicians, and did so for every stimulus type. Additionally, for both Congruent and Incongruent sequences, performance with Frozen Repeat stimuli was higher than that with nonfrozen Repeat stimuli. With data averaged over Blocks 1 and 2 (to avoid contamination from Block 3's change in conditions), a $2 \times 2 \times 2$ mixed ANOVA compared the influence of audiovisual congruency for Frozen and nonfrozen Repeat stimuli, for Musicians and Non-Musicians. A main effect of Musicianship confirmed that in general Musicians significantly outperform Non-Musicians, $F(1, 22) = 241.025$, $p < 0.001$, $\eta_p^2 = 0.174$). A main effect of Frozen versus Repeat showed that performance with Frozen stimuli exceeded that with Repeat stimuli, $F(1, 22) = 22.708$, $p < 0.001$, $\eta_p^2 = 0.958$. A main effect of Congruency showed that Congruent stimuli yielded significantly better performance than Incongruent stimuli, replicating a finding from Experiment 1, $F(1, 22) = 12.183$, $p = 0.002$, $\eta_p^2 = 0.356$). There were no significant pairwise interactions between any of the three independent variables: Musicianship × Frozen/Repeat, $F(1, 22) = 0.505$, $p = 0.485$, $\eta_p^2 = 0.023$; Musicianship × Congruency, $F(1, 22) = 0.886$, $p = 0.357$, $\eta_p^2 = 0.039$; Frozen/Repeat × Congruency, $F(1, 22) = 0.305$, $p = 0.587$, $\eta_p^2 = 0.014$; Musicianship × Frozen/Repeat × Congruency, $F(1, 22) < 0.001$, $p = 0.994$, $\eta_p^2 < 0.001$). Consistent with Aizenman and colleagues' (2013) finding with unisensory visual sequences, over multiple encounters with their own assigned Frozen Repeat sequence, subjects demonstrate learning, becoming more successful at categorizing the sequence as Repeat.

In Block 3, when the auditory portion of a FRincon stimulus was changed to a different random auditory sequence, subjects' performance with their assigned FRincon stimuli fell. Figure 8 shows Constant and Crossover-Switchback groups' performance with FRincon stimuli over successive blocks of trials. A 2×2 mixed ANOVA compared performance in Block 2, the block preceding the change in FRincon stimuli, and performance in Block 3, the block in which the change was made. Although neither the difference between

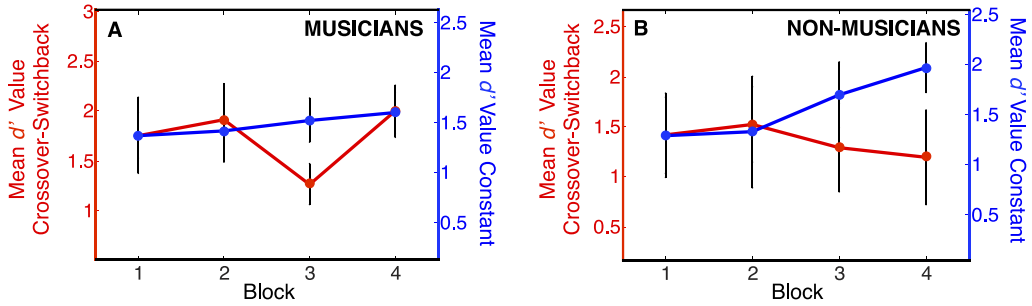

Figure 9. (A) Musicians' mean performance (expressed as d') with FRincon stimuli in the Crossover-Switchback and Constant conditions. (B) Non-Musician's mean performance with FRincon stimuli in the Crossover-Switchback and Constant conditions. Note the use of two separate *y*-axes in each panel, with each axis adjusted such that performance in Block 1 lies at the same point in the plot for both groups of subjects. Error bars in both panels represent between-subjects standard errors.

Constant and Crossover-Switchback groups nor the effect of block was significant, $F(1, 22) = 0.007$, $p = 0.934$, $\eta_p^2 = 0.0003$ and $F(1, 22) = 1.061$, $p = 0.314$, $\eta_p^2 = 0.01$, respectively, as predicted, the interaction between group and block was statistically significant, $F(1, 22) = 7.699$, $p = 0.011$, $\eta_p^2 = 0.26$.

For a clearer idea of how the Crossover-Switchback manipulation affected performance, we compared results from subjects in each yoked pair, one from the Crossover-Switchback condition and the other from the Constant condition. Overall, the Constant group's performance did not differ reliably from that of the Crossover-Switchback group in Block 1, Bootstrapped 95% CIs (0.85, 1.85) and (1.21, 2.02), respectively, and Block 2, (0.83, 1.87) and (1.21, 2.18). To facilitate the main comparison of interest, we normalized each subject's d' values relative to that subject's performance with FRincon stimuli in Block 1. A paired-samples t test compared yoked subjects' d' values in Blocks 1 and 2 to their values in Block 3. The result confirms what can be seen in Figure 8, namely that the performance of the two groups diverges significantly from Blocks 1 and 2 to Block 3, $t(11) = -3.030$, $p = 0.011$, $d = 0.88$.

Effects of musical training

Previous research suggests an association between musical training and ability to process rapidly presented stimulus sequences (Lu et al., 2014; Paraskevopoulos et al., 2014). That result led us to hypothesize that in our experiment, Musicians and Non-Musicians would be differentially influenced by to-be-ignored auditory components of Frozen visual sequences.

Figure 9 shows that the crossover introduced in Block 3 affects performance differentially for Musicians and Non-Musicians. In particular, the performance of Musician subjects dips sharply from Block 2 to Block 3. In contrast, Non-Musicians in the Crossover-Switchback condition show no such dip, instead showing a small, gradual, nonsignificant decline over blocks. A 2 × 2 mixed ANOVA with only Musicians' data from Block 2 and Block 3, compared subjects who received the Crossover-Switchback treatment to ones who had not. The analysis revealed no main effect of group, $F(1, 12) = 0.108$, $p = 0.748$, $d = 0.01$; no main effect of block, $F(1, 12) = 3.746$, $p = 0.077$, $d = 0.24$); but, importantly, a significant interaction between condition and block, $F(1, 12) = 7.246$, $p = 0.020$, $d = 0.38$. The analogous ANOVA on Non-Musicians' data revealed no difference between Constant and Crossover-Switchback conditions, $F(1, 8) = 0.037$, $p = 0.852$, $d < 0.005$; no effect of Block, $F(1, 8) = 0.096$, $p = 0.765$, $d = 0.01$; and no significant interaction between Condition and Block, $F(1, 8) = 1.746$, $p = 0.223$, $d = 0.18$. So, although Musicians were impacted significantly by the crossover treatment, Non-musicians were not. It should be kept in mind that these results may have been impacted by the different numbers of Musicians and Non-Musicians (14 vs. 10).

Following the change in FRincon stimuli for Block 3, reinstatement of the original FRincon stimuli restored Musicians' performance in Block 4 to its precrossover level. In contrast, Non-Musicians showed no evidence of such a recovery in performance.

Potential strategies for judgments

As in Experiment 1, we were interested in the strategy that subjects might have adopted in performing the task they were given. Specifically, we wanted to know whether the subjects in Experiment 2 exploited the combined summary statistic and differencing operation for which Experiment 1 provided evidence. Therefore, for each trial, we computed the difference between the summed luminances or frequencies in the first and second halves of the stimulus. Then, a logistic regression evaluated the relationship between the difference between the two halves of every stimulus and the responses made to those stimuli. This process produced a result very much like the one we saw for Experiment 1. The summary statistic accounted for a

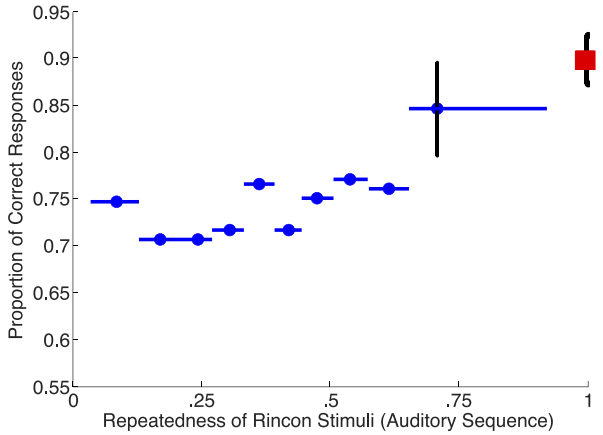

Figure 10. Proportion of correct responses in Experiment 2 with auditory components of Rincon stimuli of varying "repeatedness" in 10 equally populous bins. The degree of repetition in the auditory component of a particular Rincon stimulus is calculated by taking the mean difference between items *n* and *n* + 4 for *n* = 1:4. The mean repeatedness within each bin were transformed such that a repeatedness value of 1 corresponds to a perfect repeat. Vertical bars denote 95% confidence intervals, while horizontal bars denote the range of each bin. Also plotted (■) is the mean proportion of correct responses with Rcon stimuli.

significant fraction of the variance for visual components, ($p < 0.01$) but not for auditory components ($p = 0.14$); the interaction between summary statistics for visual and auditory domains was also nonsignificant ($p = 0.30$). It appears, therefore, that subjects' responses could have made some use of the summary statistic extracted from visual sequences.

To understand why performance in Experiment 2 differed for Rcon and Rincon stimuli, we took the same approach as in the previous experiment. Specifically, we asked whether that difference reflected the repeatedness of auditory sequences in Rincon stimuli rather than audiovisual congruency itself. Again, we calculated the repeatedness of the auditory component in every Rincon stimulus. A logistic regression on individual Rincon trials confirmed that the *pr* (Repeat) judgments made to Rincon stimuli did indeed track the repeatedness of the auditory components of the stimuli ($p < 0.0001$). This relationship is represented by the data points of Figure 10. For that figure, trials have been sorted into 10 equally populous bins according the repeatedness of stimuli in the bin, and the mean proportion of correct (Repeat) responses is shown for each bin. Note that, as in results from the previous experiment, the proportion of Repeat responses in the 10th bin, in which auditory sequences most closely approximate Repeat sequences, does not differ significantly from the proportion of Repeat responses elicited by Rcon stimuli (shown by the red square). This similarity suggests that repeated information within an auditory sequence contributes to what seems to be an effect of audiovisual congruency.

Discussion

Multiple intermittent presentations of identical audiovisual sequences improved subjects' ability to categorize a sequence's visual component as Repeat or Non-Repeat. However, performance was significantly reduced for FRincon stimuli when their visual sequences were maintained, but the concurrent auditory sequence was changed to a set of frequencies that *differed* from those that the subject had experienced previously in FRincon stimuli. This change in the auditory component of FRincon stimuli affected both groups of subjects, but it had particular impact on musically trained subjects. This result suggests that the Musician subjects had been more reliant on the auditory sequences prior to the change, even though the auditory sequences were task-irrelevant and nominally ignored. This outcome is consistent with prior research that compared Musicians and Non-Musicians on tasks quite different from ours. In that research, Musicians more readily integrated auditory and visual information (Paraskevopoulos et al., 2012). Additionally, Musicians' performance recovered completely in Block 4, when the auditory portion of the stimulus was restored to the state it had originally, in Blocks 1 and 2. That outcome is consistent with Paraskevopoulos and colleagues' (2014) demonstration that musical training enhances multisensory processing. The complete recovery shown by Crossover-Switchback Musicians in Block 4 suggests that despite intervening exposure to changed conditions in Block 3, Musicians had some relatively robust memory for the stimuli they encountered in Blocks 1 and 2.

Subjects' sole task was to judge whether the visual component of a trial's sequence did or did not contain visual items that repeated from the first half stimulus to the last. This task entailed no need to encode or store any particular sequence exemplar, although our results make it clear that some information was stored and cumulated over trials. This improvement in performance produced by multiple exposures to a particular exemplar qualifies as what McGeoch and Irion (1952) defined as incidental learning, which Gold et al. (2014) showed for unisensory visual sequences like ours. Although subjects were not asked directly whether they had seen any stimulus multiple times, during post-experiment debriefing a few subjects reported a suspicion that they had encountered the same stimulus more than once. That most subjects offered no such suspicion suggests, but certainly does not prove, that sequence-selective learning might occur with minimal awareness (Chong, Husain, & Rosenthal, 2014). This

interesting possibility, which is outside the focus of the present research, may deserve further study.

General discussion

Previous studies have demonstrated the various benefits of multisensory interactions (Chen & Spence, 2010; Mendonça, Santos, & López-Moliner, 2011; Parise, Spence, & Ernst, 2012; Seitz, Kim, & Shams, 2006; Shams & Seitz, 2008; van Atteveldt, Murray, Thut, & Schroeder, 2014), but only a few studies have explored the disruptive influences of such interactions. Exceptions include studies of spatial ventriloquism (McGurk & MacDonald, 1976) and other visual influences on auditory perception (Teramoto, Kobayashi, & Hidaka, 2013), as well as various demonstrations that auditory signals alter visual perception (Hidaka et al., 2009; Sekuler et al., 1997; Shams, Kamitani, & Shimojo, 2000; Shipley, 1964; Teramoto, Hidaka, & Sugita, 2010; Zhou et al., 2007). However, those previous studies focused on relatively low-level audiovisual tasks; none targeted a higher-order cognitive function such as memory. In order to understand how task-irrelevant auditory signals affected short-term memory, our first experiment investigated how nominally ignored auditory sequences influenced categorization of concurrent visual sequences. Our second experiment used Hebb's (1961) repetition effect to examine the buildup of this influence (incidental learning) over many trials and the transfer of information from short-term memory restricted to the stimuli within a single trial, to longer-term memory that spanned multiple trials. As Sperling and Dosher (1986) have noted, in higher-order tasks, subjects' strategies enjoy an expanded influence on performance. Because our study's task recruited higher order processes such as selective attention and memory, it was able to reveal some features of the strategies that subjects call upon. For example, Experiment 2 showed that prior experience (namely, musical training), affected subjects' reliance on task-irrelevant information. Such results confirm the usefulness of studying audiovisual interactions in a task that draws upon top-down as well as bottom-up processing.

Auditory influences

Experiment 1 showed that a Repeat *auditory* sequence increases the likelihood that a concurrent *visual* sequence will be judged as Repeat, regardless of its actual state. In other words, a Repeat auditory pattern can subvert judgments of a Non-Repeat visual sequence, degrading performance. In the extreme, performance is degraded all the way to chance level, as we saw with Nrep stimuli in Experiment 1. Our results suggest also that when a stimulus' task-relevant visual component was a Repeat, the presence of a perceptually correlated Congruent auditory component facilitates detection of the visual Repeat. This result is consistent with the demonstration that nominally ignored auditory sequences influence performance with concurrent visual sequences.

However, the advantage that Congruent audiovisual sequences have over Incongruent ones seems not to be explained entirely by audiovisual correspondence per se. Instead, it seems that *pr* Repeat judgments for audiovisual stimuli reflects the number of independent Repeat sequences, either visual, auditory, or both, that are present in an audiovisual stimulus. As a consequence, Congruent audiovisual sequences, which contain more Repeat sequences than any other type of stimulus that we tested, produces the highest *pr* Repeat judgments. Within the cognitive framework of Blurton, Greenlee, and Gondan (2014), the audiovisual combinations demonstrated in our experiments more likely reflected a superposition of modality-dependent processing than some form of audiovisual coactivation.

This perspective throws an interesting light on the audiovisual congruence effect demonstrated in both of our experiments. Unlike some cross-modal effects previously reported (e.g., Kayser, Logothetis, & Panzeri, 2010; Molholm et al., 2002), the audiovisual congruence effects produced in our task are unlikely to entail modulation of sensory activity in early sensory cortices. In fact, the complex short-term visual memory component of a subject's task meant that the task recruited cortical areas beyond previously identified multisensory regions.

Experiment 2 demonstrated that task-irrelevant auditory sequences have a particular influence on the performance of subjects with musical training. When the nominally ignored tones in a Frozen sequence were switched to a different, random sequence of tones, Musician subjects' performance dropped. This suggests that Musicians rely more heavily on the auditory component of audiovisual stimuli. Because of their reliance on the auditory component, the unannounced crossover manipulation undermined Musicians' performance, giving Non-Musicians an advantage over Musicians. However, when it came to the switchback manipulation in the fourth block of trials, Musicians again showed superior performance to Non-Musicians: a result that is also explained by Musicians' reliance on the auditory component before the crossover took place. Note that we cannot be sure that musical training per se is responsible for these results. In particular, our results cannot rule out the possibility that, when they experience audiovisual stimuli, indi-

Model	Estimate	Standard error	*z*-value	*pr*(>|*z*|)
(Intercept)	0.87	0.41	2.12	0.03
Visual	−0.30	0.10	−3.19	0.0014**
Auditory	−0.04	0.03	−1.46	0.14
Visual:Auditory	0.01	0.01	1.04	0.30

Table 5. Logistic regression of summary statistic for Experiment 2. *Note*: $^{**}p < 0.01$.

viduals who would give extra weight to auditory input would also be more likely to start and persist in musical training.

Ensemble statistics, false positives, and individual differences

One of our aims was to understand the origins of individual differences that Gold et al. (2014) saw in their study of short-term memory for visual sequences like the ones we used. Although we do not have a complete explanation, part of an explanation might lie in differences in individual subjects' reliance upon a summary statistical representation of stimulus information. The summary statistic we chose to examine was the absolute difference between the summed luminances within each half of a stimulus sequence. The hypothesis was that larger differences between summed luminances in each half sequence would promote "Non-Repeat" responses. Supporting this idea, logistic regressions (Tables 3 and 5) revealed highly significant statistical relationships between subjects' responses on one hand, and the difference between summed luminances within each half of a visual sequence on the other.

To examine individual subjects' reliance on this summary statistic, we generated logistic regression models for responses on Nincon trials, first for all subjects as an aggregate, and then, as our focus was on differences among subjects, for each subject individually. The logistic models gave predicted *pr* Repeat and *pr* Non-Repeat for various values of difference between the integrated half-sequences of an Nincon stimulus. Then, to evaluate the agreement between the predictions and the actual data, pROC, (an R package; Robin et al., 2011), cumulated the predicted values to generate a ROC (receiver operating characteristic), the area under the ROC, and stratified 95% bootstrap confidence intervals on that area. This operation treated the difference in summed luminances as the "strength" variable on the *x*-axis of a signal detection model, determining for Repeat trials the rates of response at each point along that *x*-axis, repeating the process for Non-Repeat trials, and then using the resulting pairs to define the ROC.

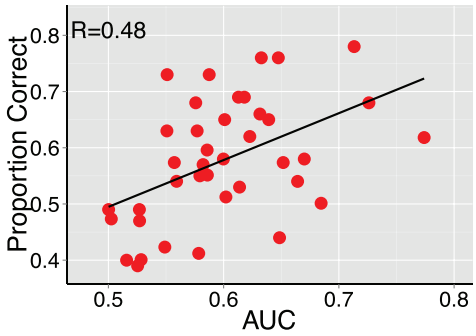

Figure 11. Proportion correct versus AUC. Each data point represents one subject. Also shown is the best fitting regression line and the 95% confidence region around that line. Note that because the stimuli were Nincon, a response of "Non-Repeat" was correct.

The area under the ROC Area Under the Curve (AUC) is a measure of how well the differenced summary statistic alone predicts which response will be made (Zhou, McClish, & Obuchowski, 2011). With AUC = 0.5 the differenced summary statistic would have no predictive value whatever; that is, predictions based on the differenced summary statistic would be at chance level; with AUC = 1.0, the statistic would have perfect predictive value, with the differenced ensemble statistic completely determining response selection. Averaged over all subjects in Experiment 1, AUC was 0.581, 95% CIs (0.556, 0.606); over all subjects in Experiment 2, AUC was similar, 0.600, 95% CIs (0.577, 0.621). So, for aggregated subjects in each experiment, AUC was significantly above the chance level of 0.5, as expected from Tables 3 and 5. To assess individual differences, we computed ROCs and then values of AUC for individual subjects in both experiments. The resulting values of AUC varied considerably over subjects, ranging from a low of 0.50 to a high of 0.77 ($\bar{x} = 0.60$, $SD = 0.06$). This considerable range suggests that subjects vary in their reliance on the summary statistic described above. To examine how this summary statistic impacted subjects' performance, we regressed individual subjects' proportion of correct responses against their AUC values. In Figure 11, individual subjects' AUC values are plotted on the horizontal axis; the subjects' proportion of correct responses is plotted on the vertical axis (e.g., responding "Non-Repeat" to the Nincon stimulus sequence, which was actually Non-Repeat). The figure shows that the two measures were in fact reliably correlated, $r = 0.48$, $t(28) = 2.895$, $p < 0.01$, two-tailed test, but that individual subjects differed considerably in this association.

Differences in subjects' reliance on the summary statistic does not completely explain the individual differences in performance seen in our experiments and in those of Gold et al. (2014) and Agus et al. (2010), but it may be a significant component of a complete explanation. Of course, subjects might also rely on other summary statistics in order to perform our task. For example, they might rely on the mean differences between consecutive items in the first and second halves of the sequence, or might give extra attention to relatively salient items. These additional possibilities notwithstanding, it does appear subjects can extract and exploit macro information (the "gist"), not only from auditory streams (Warren & Bashford, 1993) or from complex visual scenes (Greene & Oliva, 2009) as others have shown, but also from rapidly presented, random sequences of luminances. Individual differences in the degree to which subjects utilize the summary statistic confirms the advantage that comes from studying audiovisual interactions with a task in which top-down processes play a major role (Sperling & Dosher, 1986).

Theoretical implications

In Hebb's (1961) original application of the repetition design, subjects heard 24 series of nine-digit sequences. After each sequence, they tried to repeat those digits in the order that they heard. On every third trial, the same sequence was repeated, and performance slowly, but steadily improved on that recurring sequence. In fact, by the fourth repetition, performance with the recurring sequence exceeded that for interspersed random, nonrecurring sequences. To Hebb (1961), this cumulation suggested that hearing one sequence set up in memory some neural trace, which was not "wiped clean" or completely overwritten by the subsequent unrelated sequence. Rather, the trace was maintained long enough to cumulate with the next occurrence of that same sequence. Over the years following that original report, partly in response to Hebb's (1961) portrayal of the mechanism responsible for the repetition effect, researchers have assumed that this effect could be produced only when individual items in a sequence were distinctive. For example, that assumption received support from Horton and colleagues' (2008) study of serial recall for sequences of upright and inverted unfamiliar faces. In that study, a repetition effect was seen only for sequences of faces when the faces were upright, which was taken to mean that the ability to encode list items distinctively was crucial to the repetition effect. However, that assumption does not align with one result from our study: a repetition effect generated by sequences comprising items that are not perfectly distinctive. As perceptual similarity is key in so many other forms of memory and learning (e.g., Dubé et al., 2014; Kahana, 2012; Wickelgren, 1965), it would be useful to determine how the cumulation rate in Hebb's (1961) repetition effect depended upon the distinctiveness of stimulus items that comprise the sequences.

Taken together, our experiments show that even when time-varying auditory information is not perceptually correlated with the visual information it accompanies, and even when that auditory information is supposedly unattended, it retains some power to influence short- and long-term memory of concurrent visual sequences.

Keywords: audiovisual interaction, incidental learning, Hebb repetition effect, musical training, ensemble statistics

Acknowledgments

This study was supported by CELEST, an NSF Science of Learning Center grant (SBE-0354378). A. S. K. was also supported by an NIH grant for Undergraduate and Graduate Training in Computational Neuroscience, 1T90DA032435-01. We are indebted to Trevor Agus, Aaron Seitz, Barbara Shinn-Cunningham, and Chad Dubé for helpful suggestions. We are especially indebted to Avigael M. Aizenman for developing much of the code used for our experiments.

Commercial relationships: none.
Corresponding author: Robert Sekuler.
Email: vision@brandeis.edu.
Address: Volen Center for Complex Systems, Brandeis University, Waltham, MA, USA.

References

Agus, T. R., Thorpe, S. J., & Pressnitzer, D. (2010). Rapid formation of robust auditory memories: insights from noise. *Neuron*, *66*, 610–618.

Aizenman, A. M., Gold, J. M., & Sekuler, R. (2013, May). *Multisensory integration in visual pattern recognition: Music training matters*. Paper presented at Annual Meeting of the Visual Sciences Society, Naples, FL.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*, 122–131.

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-

Hashimi, O., Faraji, F., Janowich, J., & Gazzaley, A. (2013). Video game training enhances cognitive control in older adults. *Nature*, *501*, 97–101.

Blurton, S. P., Greenlee, M. W., & Gondan, M. (2014). Multisensory processing of redundant information in go/no-go and choice responses. *Attention, Perception, & Psychophysics*, *76*, 1212–1233.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Chen, Y. C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition*, *114*, 389–404.

Chong, T. T.-J., Husain, M., & Rosenthal, C. R. (2014). Recognizing the unconscious. *Current Biology*, *24*, R1033–R1035.

Cohen, A., Ivry, R., & Keele, S. W. (1990). Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 17–30.

Dubé, C., & Sekuler, R. (2015). Obligatory and adaptive averaging in visual short-term memory. *Journal of Vision, 15*(4):13, 1–13, doi:10.1167/15.4.13. [PubMed] [Article]

Dubé, C., Zhou, F., Kahana, M. J., & Sekuler, R. (2014). Similarity-based distortion of visual short-term memory is due to perceptual averaging. *Vision Research*, *96*, 8–16.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*, 499–504.

Gold, J. M., Aizenman, A., Bond, S. M., & Sekuler, R. (2014). Memory and incidental learning for visual frozen noise sequences. *Vision Research*, *99*, 19–36.

Graham, C. H., & Kemp, E. H. (1938). Brightness discrimination as a function of the duration of the increment in intensity. *Journal of General Physiology*, *21*, 635–650.

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*, 137–176.

Guttman, N., & Julesz, B. (1963). Lower limits of auditory analysis. *Journal of the Acoustical Society of America*, *35*, 610.

Hebb, D. O. (1961). Distinctive features of learning in the higher animal: A symposium. In A. Fessard, R. Gerard, & J. Konorski (Eds.), *Brain mechanisms and learning* (pp. 37–46). Oxford: Blackwell Scientific Publications.

Hidaka, S., Manaka, Y., Teramoto, W., Sugita, Y., Miyauchi, R., Gyoba, J., & Iwaya, Y. (2009). Alternation of sound location induces visual motion perception of a static object. *PLoS One*, *4*, e8188.

Horton, N., Hay, D. C., & Smyth, M. M. (2008). Hebb repetition effects in visual memory: The roles of verbal rehearsal and distinctiveness. *Quarterly Journal of Experimental Psychology*, *61*, 1769–1777.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*, 394–412.

Julesz, B., & Hirsch, I. J. (1972). Visual and auditory perception –an essay of comparison. In E. E. David, Jr., & P. B. Denes (Eds.), *Human communication: A unified view*. New York: McGraw-Hill.

Kahana, M. J. (2012). *Foundations of human memory*. New York: Oxford University Press.

Kayser, C., Logothetis, N. K., & Panzeri, S. (2010). Visual enhancement of the information representation in auditory cortex. *Current Biology*, *20*, 19–24.

Kelly, S. W., Burton, A. M., Kato, T., & Akamatsu, S. (2001). Incidental learning of real-world regularities. *Psychological Science*, *12*, 86–89.

Lee, H., & Noppeney, U. (2011). Long-term music training tunes how the brain temporally binds signals from multiple senses. *Proceedings of the National Academy of Sciences, USA*, *108*, E1441–E1450.

Lu, Y., Paraskevopoulos, E., Herholz, S. C., Kuchenbuch, A., & Pantev, C. (2014). Temporal processing of audiovisual stimuli is enhanced in musicians: Evidence from magnetoencephalography (MEG). *PLoS One*, *9*, e90686.

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American Journal of Psychology*, *87*, 173–188.

McGeoch, J. A., & Irion, A. L. (1952). *The psychology of human learning* (2nd ed.). New York: Longmans, Green.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *265*, 746–748.

Mendonça, C., Santos, J. A., & López-Moliner, J. (2011). The benefit of multisensory integration with biological motion signals. *Experimental Brain Research*, *213*, 185–192.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: A high-density electrical

mapping study. *Cognitive Brain Research*, *14*, 115–128.

Mueller, G., & Hall, J. W. (1998). *Audiologist's desk reference: Audiolologic management, rehabilitation and terminology* (Vol. II). San Diego, CA: Singular Publishing Group.

Page, M. P. A., & Norris, D. (2009). A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word forms. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *364*, 3737–3753.

Paraskevopoulos, E., Kuchenbuch, A., Herholz, S. C., & Pantev, C. (2012). Musical expertise induces audiovisual integration of abstract congruency rules. *Journal of Neuroscience*, *32*, 18196–18203.

Paraskevopoulos, E., Kuchenbuch, A., Herholz, S. C., & Pantev, C. (2014). Multisensory integration during short-term music reading training enhances both uni- and multisensory cortical processing. *Journal of Cognitive Neuroscience*, *26*, 2224–2238.

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Current Biology*, *22*, 46–49.

Pfafflin, S. M. (1968). Detection of auditory signal in restricted sets of reproducible noise. *Journal of the Acoustical Society of America*, *43*, 487–490.

Pfafflin, S. M., & Matthews, M. V. (1966). Stimulus features in signal detection. *Journal of the Acoustical Society of America*, *39*, 340–345.

Piantadosi, S. (2005). Crossover designs. In S. Piantadosi (Ed.), *Clinical trials: A methodologic perspective* (2nd ed., chap, 15). Hoboken, NJ: Wiley.

Pollack, I. (1956). Identification and discrimination of components of elementary auditory displays. *Journal of the Acoustical Society of America*, *28*, 906–909.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., & Sanchez, J.-C. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Informatics*, *12*, 77, doi:10.1186/1471-2105-12-77.

Rosenholtz, R., Huang, J., Raj, A., Balas, B. J., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision*, *12*(4):14, 1–17, doi:10.1167/12.4.14. [PubMed] [Article]

Rosenthal, O., Shimojo, S., & Shams, L. (2009). Sound-induced flash illusion is resistant to feedback training. *Brain Topography*, *21*, 185–192.

Seitz, A. R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Current Biology*, *16*, 1422–1427.

Seitz, A. R., & Watanabe, T. (2009). The phenomenon of task-irrelevant perceptual learning. *Vision Research*, *49*, 2604–2610.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature*, *385*, 308.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature*, *408*, 788.

Shams, L., Kamitani, Y., Thompson, S., & Shimojo, S. (2001). Sound alters visual evoked potentials in humans. *Neuroreport*, *12*, 3849–3852.

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*, 411–417.

Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science*, *145*, 1328–1330.

Skoe, E., & Kraus, N. (2012). A little goes a long way: How the adult brain is shaped by musical training in childhood. *The Journal of Neuroscience*, *34*, 11507–11510.

Sorkin, R. D. (1962). Extensions of the theory of signal detectability to matching procedures in psychoacoustics. *Journal of the Acoustical Society of America*, *34*, 1745–1751.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*, 971–995.

Sperling, G., & Dosher, B. A. (1986). Strategy and optimization in human information processing. In K. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of human perception and performance* (Vol. I, pp. 2.1–2.65). Hoboken, NJ: Wiley.

Stadler, M. A. (1993). Implicit serial learning: Questions inspired by Hebb (1961). *Memory & Cognition*, *21*, 819–827.

Strait, D. L., Parbery-Clark, A., Hittner, E., & Kraus, N. (2012). Musical training during early childhood enhances the neural encoding of speech in noise. *Brain and Language*, *123*, 191–201.

Teramoto, W., Hidaka, S., & Sugita, Y. (2010). Sounds move a static visual object. *PLoS One*, *5,* e12255.

Teramoto, W., Kobayashi, M., & Hidaka, S. (2013). Vision contingent auditory pitch aftereffects. *Experimental Brain Research*, *229*, 97–102.

Thelen, A., Matusz, P. J., & Murray, M. M. (2014). Multisensory context portends object memory. *Current Biology*, *24*, R734–R735.

van Atteveldt, N., Murray, M. M., Thut, G., & Schroeder, C. E. (2014). Multisensory integration:

Flexible use of general operations. *Neuron*, *81*, 1240–1253.

Warren, R. M., & Bashford, J. A., Jr. (1993). When acoustic sequences are not perceptual sequences: The global perception of auditory patterns. *Perception & Psychophysics*, *54*, 121–126.

Wickelgren, W. A. (1965). Distinctive features and errors in short-term memory for English vowels. *Journal of the Acoustical Society of America*, *38*, 583–588.

Zhou, F., Wong, V., & Sekuler, R. (2007). Multisensory integration of spatio-temporal segmentation cues: One plus one does not always equal two. *Experimental Brain Research*, *180*, 641–654.

Zhou, X.-H., McClish, D. K., & Obuchowski, N. A. (2011). *Statistical methods in diagnostic medicine* (2nd ed.). Hoboken, NJ: Wiley.