Published in final edited form as:
J Exp Psychol Learn Mem Cogn. 2011 September ; 37(5): 1081–1091. doi:10.1037/a0023700.

Learning Across Senses: Cross-Modal Effects in Multisensory Statistical Learning

Aaron D. Mitchel and **Daniel J. Weiss**
Pennsylvania State University

Abstract

It is currently unknown whether statistical learning is supported by modality-general or modality-specific mechanisms. One issue within this debate concerns the independence of learning in one modality from learning in other modalities. In the present study, the authors examined the extent to which statistical learning across modalities is independent by simultaneously presenting learners with auditory and visual streams. After establishing baseline rates of learning for each stream independently, they systematically varied the amount of audiovisual correspondence across 3 experiments. They found that learners were able to segment both streams successfully only when the boundaries of the audio and visual triplets were in alignment. This pattern of results suggests that learners are able to extract multiple statistical regularities across modalities provided that there is some degree of cross-modal coherence. They discuss the implications of their results in light of recent claims that multisensory statistical learning is guided by modality-independent mechanisms.

Keywords

multimodal statistical learning; multisensory perception; speech segmentation

Statistical learning is the process by which learners rapidly acquire structured information from variable environmental inputs in the absence of explicit reward or feedback (Aslin & Newport, 2009). This process has been demonstrated to operate at numerous levels within the domains of language acquisition and visual processing, leading researchers to question whether statistical learning is supported by a singular, modality-general mechanism or by a set of modality-specific mechanisms. After discovering that statistical learning plays a critical role in speech segmentation (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996), several studies demonstrated comparable learning with nonspeech auditory stimuli (Saffran, Johnson, Aslin, & Newport, 1999), visual shapes (Fiser & Aslin, 2002), and motor tasks (Hunt & Aslin, 2001). Because the levels of performance and types of computations appear to be equivalent across these modalities (e.g., Creel, Newport, & Aslin, 2004), these initial findings were interpreted as empirical support for a modality-general mechanism (Kirkham, Slemmer, & Johnson, 2002). However, this view has been countered

Correspondence concerning this article should be addressed to Aaron D. Mitchel or Daniel J. Weiss, Department of Psychology and Program in Linguistics, Pennsylvania State University, 643 Moore Building, University Park, PA 16802. adm241@psu.edu or djw21@psu.edu.

by evidence demonstrating modality constraints. Using a different type of statistical learning task, Conway and Christiansen (2005) found a quantitative advantage for the audio modality relative to the visual and tactile modalities, as well as qualitative differences in the types of computations performed in each modality. From this pattern of results, the authors have argued for the existence of multiple, parallel statistical learning mechanisms that are directly linked to the sensory modality of the input (see also Conway & Christiansen, 2006, 2009), consistent with theories of embodied cognition that more broadly posit that cognitive representations are rooted in modality-specific sensorimotor systems (e.g., Barsalou, Simmons, Barbey, & Wilson, 2003).

In addressing whether statistical learning is rooted in a central mechanism or multiple modality-specific mechanisms, researchers initially relied on comparisons across individual modalities. However, the sensory environment is seldom limited to a single modality or input source (Stein & Stanford, 2008); thus, it is likely that statistical learning mechanisms encounter multiple statistical regularities across modalities on a regular basis. As a consequence, an alternative approach to examining the modality-specificity of statistical learning is to explore the degree to which multimodal input sources are processed independently. According to a modality-general view, simultaneous multisensory input should be processed by a unitary mechanism, suggesting that statistical learning should not be independent across modalities. Thus, evidence of independence in multisensory statistical learning would provide strong support for modality-specific models of statistical learning.

Seitz, Kim, van Wassenhove, and Shams (2007) investigated this issue by presenting participants with a stream of audiovisual bigrams constructed from distinct two-dimensional shapes and highly contrastive audio stimuli with varied spectrotemporal properties. Familiarization consisted of either unimodal (audio or visual streams in isolation) or multimodal (audiovisual) stimuli. Participants in the multimodal condition were able to identify correctly audio, visual, and audiovisual bigrams that appeared in the familiarization stream when tested against novel bigrams constructed from the same stimuli. This result suggests that statistical learning mechanisms are capable of simultaneously extracting multiple sequential dependencies from multimodal input. Further, Seitz et al. (2007) noted that there was no difference in performance in either modality across both familiarization conditions (unimodal and multimodal), suggesting that simultaneous learning in one modality did not impact learning in the second modality. From this pattern of findings, the authors claim that both streams were processed independently, consistent with a modality-specific account of statistical learning.

The conclusions of Seitz et al. (2007) regarding sensory independence must be tempered, as their stimuli contained perfect cross-modal correspondence between the input streams. Each audio segment always occurred concurrently with a single visual shape. It is possible that the absence of a significant difference in performance between unimodal and multimodal conditions hinged on the perfect correlation across modalities. Prior research investigating multistream visual statistical learning found evidence of distinct learning patterns for perfectly correlated input (Turk-Browne, Isola, Scholl, & Treat, 2008). Further, it is possible that learning could have occurred primarily in one modality and then transferred at test to the second modality. Given these concerns, the goals of the current study are twofold. First,

we endeavor to test systematically whether multistream statistical learning is processed independently for each modality. Second, we explore the types of cross-modal relationships that influence simultaneous learning of multimodal input streams. We accomplish both of these goals by manipulating the statistical correspondence between individual elements across streams as well as cross-modal boundary alignment (described in the following). Boundary information may be particularly important for cross-modal learning, as demonstrated by Cunillera, Càmara, Laine, and Rodríguez-Fornells (2010), who found that audiovisual contiguity between a picture presented to learners and dips in transitional probabilities (a statistical cue to word boundaries) facilitated successful performance in a speech-segmentation task (see also Thiessen, 2010). Here we ask whether such contiguity might also impact segmentation of both an auditory and visual stream in a cross-modal statistical learning paradigm.

In a series of four experiments, we investigate the influence of cross-modal associations on multimodal statistical learning. In Experiment 1, we collect baseline measures of performance for our input streams presented in isolation. The streams are based on the tone stream from Saffran et al. (1999) and the visual shape stream from Fiser and Aslin (2002). In Experiment 2, we present both streams simultaneously, with perfect predictability between audio and visual elements (similar to Seitz et al., 2007). In Experiment 3, this predictability between streams is removed, as each audio element is equally likely to co-occur with each visual element (and vice versa), although the triplet boundaries are aligned across streams. Finally, in Experiment 4, we further disrupt cross-modal relationships by offsetting the streams such that triplet boundaries are not aligned across streams. If statistical learning is achieved independently across modalities, then disrupting cross-modal relationships should not alter segmentation performance. On the other hand, evidence that statistical learning is sensitive to these manipulations would constrain the types of modality-specific theories to be considered.

Experiment 1a: Tone Sequence Alone

Previous statistical learning studies have demonstrated statistical learning with tone streams (Creel et al., 2004; Saffran et al., 1999). Here we test the ability of participants to segment a similar tone stream adapted from the stimuli used in Experiment 3 of Saffran et al. (1999). The goal of Experiment 1a is to replicate this effect and establish a baseline level of performance for the tone stream when presented in isolation.

Method

Participants—Twenty-six naïve undergraduate introductory psychology students (19 female and seven male), participating for course credit, were included in the analysis. All participants were monolingual English speakers. Given the role of attention in statistical learning tasks (Toro, Sinnett, Soto-Faraco, 2005; Turk-Browne, Jungé, & Scholl, 2005; see Weiss, Gerfen, & Mitchel, 2009), we excluded from analysis any participant who gave a self-reported effort level below seven on a 10-point scale (three participants). By excluding participants on the basis of effort, we adopted a conservative approach to ensure that negative results could not be attributed to inattentiveness.

Stimuli—The stimuli in Experiment 1a were modeled after previous studies examining the ability of learners to track statistical dependencies across tone sequences (Creel et al., 2004; Saffran et al., 1999). We chose tone stimuli rather than speech sounds because it was easier to precisely manipulate duration and produce a more homogenous set. In addition, Seitz et al. (2007) used tone stimuli, and we wanted to ensure that our results could be compared directly. The tone sequences in this experiment were created from all 12 pure tones within the one-line Western chromatic octave between C_4, or "middle C," and C_5. Each tone was created using a sine-wave generator in Praat (Boersma & Weenink, 2008) based on pitch frequencies set by the Acoustical Society of America, keeping length constant at 1 s per tone. The tones were arranged into four groups of three (FGD, G#C#B, CF#D#, and EAA#), forming triplets that avoid any standard musical frame (major/minor chords).

The triplets were concatenated in Praat to form a loop consisting of 24 triplets in a pseudorandom order, with each triplet occurring an equal number of times and no triplet ever following itself. There were no silences between tones, nor were there any other acoustic markers of the triplet boundaries. This loop was repeated four times and concatenated into a 96-triplet block, lasting 4 min and 48 s. The familiarization stream was concatenated in Praat and encoded in .WAV format with a sampling rate of 44.1 kHz.

The familiarization stream followed an identical structure to statistical learning speech-segmentation studies (e.g., Saffran, Aslin, & Newport, 1996), such that individual tones were analogous to syllables and the triplets were analogous to statistically defined words. Because each tone appears in only one triplet, the transitional probability (the probability of two sounds co-occurring relative to the sounds' overall frequency of occurrence; see Saffran, 2003) within triplets was 1.00, whereas the transitional probability between triplets dipped to .33 (because each triplet never followed itself, it could only be followed by one of three other tones). As in the aforementioned segmentation studies, the dips in transitional probabilities provided the only reliable cue to triplet boundaries.

Procedure—Participants were instructed to listen to an audio stream followed by a test that would assess the information learned from the stream. There were no explicit instructions given about the nature of the audio stream, nor were participants informed that it was composed of sequences of triplets. The stream was repeated three times using Apple iTunes software with a 1-min silence between each block for a total of 16 min.

Following familiarization, participants were given a 16-item, two-alternative, forced-choice test, equivalent to the test used in previous segmentation studies (e.g., Saffran, Aslin, & Newport, 1996; Weiss, et al., 2009). In each trial, a statistically defined triplet (hereafter referred to as a tone-word) was paired with a statistically incompatible triplet (consisting of the last tone from one triplet concatenated with the first two tones of a different triplet; hereafter referred to as a tone-partword). Between test items, there was a 1 s pause, and intertrial intervals lasted 4 s, during which participants indicated which of the two sequences was most consistent with the audio stream by circling either "1" or "2" on the answer sheet. The test was composed of the four tone-words, along with four tone-partword foils. Each tone-word was paired with two different tone-partwords twice (counterbalancing for order), yielding 16 total test trials. After completing the test, participants filled out a questionnaire

on language background (how many languages spoken, number of years studied, and whether they would label themselves as being bilingual). The questionnaire also included a self-report effort rating (how hard the participants tried in the task). The self-reported effort level is a particularly important metric in subsequent visual and audiovisual conditions, because any visual exposure requires participants to attend to the display.

After completing the questionnaire, participants were tested on a version of the Simon task (Simon & Small, 1969). This measure of general executive functioning requires learners to respond to the color of a presented stimulus while ignoring positional information that is either congruent or incongruent with the response location (see Weiss, Gerfen, & Mitchel, 2010, for a detailed description of the task). The size of the Simon effect (a subtraction of reaction times between incongruent and congruent trials) varies across individuals, and performance is thought to reflect selective attention, inhibitory control, or task-switching abilities (see Lu & Proctor, 1995, for review).

Results and Discussion

The overall results are presented in Figure 1. The mean test score in Experiment 1a was 11.23 out of 16 (70%), with a standard deviation of 1.95. A one-sample *t* test (all tests were two-tailed) revealed that performance for the tone sequence was significantly above chance (in all experimental conditions, chance is defined as 50%, or eight out of 16), $t(25) = 8.47$, $p < .001$, $d = 3.39$. These results represent a successful replication of the findings reported by Saffran et al. (1999), indicating that statistically defined tone sequences can be segmented into their constituent triplets. Our findings provide a baseline learning rate for comparison in subsequent experiments.

Experiment 1b: Visual Sequence Alone

In Experiment 1b we familiarize adults with a visual sequence of arbitrary shapes to replicate the findings of Fiser and Aslin (2002) and provide a baseline level of performance for the visual stream in isolation.

Method

Participants—Twenty-four (17 female and seven male) naïve undergraduate introductory psychology students, participating for course credit, were included in the analysis. All participants were monolingual English speakers. We excluded from analysis any participant who gave a self-reported effort level below seven on a 10-point scale (five participants).

Materials—The stimuli in Experiment 1b were modeled on previous research investigating segmentation of visual-shape sequences (Fiser & Aslin, 2002). We created a movie (using Macromedia Director MX 2004) consisting of a sequence of single shapes (the same 12 simple black shapes used in Fiser & Aslin, 2002; see Figure 2a). A 6-cm (5.19°) wide × 10-cm (8.62°) long static black vertical bar was positioned in the center of a 17-in. (43.18-cm) LCD flat panel monitor at 1024 × 768 pixel resolution. The movie consisted of a single shape moving smoothly, at a constant rate, from the starting position (behind the occluder) out toward the edge of the window (for a distance of 4 cm, 3.47°) in a straight, horizontal path and then returning along the same path, ending behind the occluder. After one shape

disappeared behind the occluder, the subsequent shape would emerge from the opposite side and proceed along the same horizontal plane out toward the opposite side of the screen (see Figure 2b). Each complete movement, from the occluder to the edge and back, lasted exactly 1 s.

The shapes were grouped into the same four base triplets used in Fiser and Aslin (2002; see Figure 2a). Similar to the tone-words, the base triplets were sequences of three consecutive shapes. The four triplets were concatenated into a continuous movie (using Macromedia Director MX 2004) of 24 triplets in a pseudorandom order, with each triplet appearing the same number of times and no triplet following itself. This movie loop was then repeated four times and combined into a movie lasting 4 min 48 s, consisting of 96 triplets. The movie was then exported as an 800 × 600-pixel Quicktime movie (.MOV) with Sorenson 3 video compression, at a frame rate of 30 frames per second.

The statistical structure of the visual sequence was identical to that of the tone sequence in Experiment 1a. The transitional probabilities within triplets were 1.00, whereas the transitional probabilities between triplets dipped to 0.33. There were no other cues to word boundary other than these transitional probabilities.

Procedure—Participants were instructed to watch a movie followed by a test to assess the information they learned from the movie. Participants then watched the movie clip described above, repeated three times with a 1-min pause (during which the screen would turn white) between each block for a total of 16 min. The movie clip was presented using Apple iTunes software. Following the clip, participants were given a 16-item, two-alternative, forcedchoice test, structurally identical to Experiment 1a except that the test items were shape sequences rather than tone sequences. All other aspects of the procedure were identical to Experiment 1a.

Results and Discussion

The mean test score in Experiment 1b was 10.71 out of 16 (67%), with a standard deviation of 2.90 (see Figure 1). A one-sample *t* test indicated that performance for the visual shape sequence was significantly above chance, $t(23) = 4.56$, $p < .001$, $d = 1.86$. An independent samples *t* test between performance on the visual test (Experiment 1a) and auditory test (Experiment 1b) revealed no significant difference, $t(48) = 0.75$, $p = .455$, $d = 0.22$. These findings represent a successful replication of the findings reported in Fiser and Aslin (2002). When the visual stream was presented in isolation, the underlying constituent triplets were learnable through their statistical properties. The results provide a baseline comparison for performance on the visual sequence in subsequent experiments in which the shape sequences are presented with accompanying tone sequences. In Experiment 2, we begin to explore whether such audiovisual sequences are learnable and whether this learning is dependent on the correspondence between the audio and visual stimuli.

Experiment 2: Correlated, Synchronized Streams

The goal of Experiment 2 is to determine whether learners can track the statistical structure of two input streams simultaneously in two modalities. Recent work has demonstrated that

learners are capable of tracking multiple sets of sequential statistics within the auditory modality (Gebhart, Aslin, & Newport, 2009; Mitchel & Weiss, 2010; Weiss et al., 2009). Here we test whether this ability extends to sequential statistics presented simultaneously in different modalities, exposing participants to an audiovisual familiarization stream composed of the audio stream from Experiment 1a and the visual stream from Experiment 1b.

Method

Participants—Fifty (28 female and 22 male) naïve undergraduate introductory psychology students, participating for course credit, were included in the analysis. All participants were monolingual English speakers. We excluded from analysis participants who failed to follow instructions (three) or gave a self-reported effort level below seven on a 10-point scale (10). Failure to follow instructions included falling asleep or closing eyes for extended periods of time, repeatedly removing headphones, stopping or skipping ahead in the familiarization stream, or failing to complete the test.

Materials—The familiarization stream in Experiment 2 was created by combining the tone stream in Experiment 1a with the visual stream in Experiment 1b. Using Adobe Premiere, we synched the audio and visual streams, aligning the onset and offset of each shape and tone (each individual tone and shape presentation lasted 1 s). The combined, audiovisual stream was then exported as an 800 × 600-pixel Quicktime movie encoded with Sorenson 3 compression, with a total duration of 4 min 48 s. The statistical structure of each stream was identical to those presented in Experiment 1. The audio and visual streams contained a one-to-one correspondence between each individual tone and a shape element such that tone triplet ABC always coincided with the visual triplet ABC, tone triplet DEF with visual triplet DEF, and so on (see Figure 3a). A pseudorandom ordering was used for presentation, such that no element followed itself and all elements occurred equally often.

Procedure—All aspects of the procedure were identical to Experiment 1b, except that half of the participants were given the audio test from Experiment 1a, and the other half received the visual test from Experiment 1b. We elected to give participants only one test, rather than both, to avoid any potential transfer, interference, or shifts in strategy from one test to the other. Unlike Seitz et al. (2007), we did not use an audiovisual test, because they can be solved using multiple strategies (e.g., audio, visual, or audiovisual) and are thus somewhat difficult to interpret. Further, because we manipulate cross-modal relationships in subsequent experimental conditions, it would be difficult to implement an audiovisual test that would be comparable across experiments.

Results and Discussion

The mean test score for the audio test in Experiment 2 was 10.72 out of 16 (67%), with a standard deviation of 2.62 (see Figure 1). A one-sample *t* test indicated that performance on the audio test was significantly above chance, $t(24) = 5.19$, $p < .001$, $d = 2.12$. This was not significantly different from performance in Experiment 1a, when the audio stream was presented in isolation, $t(49) = 0.79$, $p = .432$, $d = 0.23$. The mean test score for the visual test in Experiment 2 was 10.12 out of 16 (63%), with a standard deviation of 2.62 (see Figure 1).

A one-sample *t* test indicated that performance on the visual test was significantly above chance, $t(24) = 3.89$, $p = .001$, $d = 1.59$. This was not significantly different from performance in Experiment 1b, when the visual stream was presented in isolation, $t(47) = 0.73$, $p = .468$, $d = 0.21$. An independent samples *t* test between performance on the visual and audio tests revealed no significant difference, $t(48) = 0.79$, $p = .432$, $d = 0.23$.

The results of Experiment 2 indicate that learners are able to track two sets of sequential statistics simultaneously, extending previous work demonstrating this sequentially within a single modality (Gebhart et al., 2009; Mitchel & Weiss, 2010; Weiss et al., 2009). Using a paradigm that tests trigrams structures and is more comparable with earlier statistical learning methods (e.g., Fiser & Aslin, 2002; Saffran et al., 1999), Experiment 2 replicates the effect observed by Seitz et al. (2007). Our results provide further evidence that statistical structures in two modalities can be learned simultaneously. As noted previously, it is possible that participants only segmented one of the two streams and then transferred knowledge to the other stream during test, a confounding factor that was also present in the Seitz et al. study. The design of the streams in each of these experiments correlated every audio and visual item with a token in the other modality (e.g., each audio tone was presented simultaneously with the same visual shape throughout familiarization and vice versa). Learners could have transferred knowledge regarding stimulus-specific associations between individual elements across modalities (i.e., tones and shapes) or positional information (Endress & Bonatti, 2007; Endress & Mehler, 2009). For example, learners could have segmented the auditory stream and then noticed that a particular shape always occurred at the onset of an auditory triplet, thereby cueing boundary information. In Experiments 3 and 4 we explore the possibility that multimodal statistical learning is supported by transfer of element-to-element association or positional information, respectively.

Experiment 3: Uncorrelated, Synchronized Streams

In Experiment 3, the triplets within each stream are ordered such that each element is equally likely to occur with one of four possible elements in the other stream. If the bimodal learning in Experiment 2 was a product of element-to-element transfer of learning, then we would predict that only one stream in Experiment 3 should be learned. However, if both streams were learned concurrently during familiarization in Experiment 2, then here we predict no decrement in performance for either stream.

Method

Participants—Forty-nine (27 female and 22 male) naïve un-dergraduate introductory psychology students, participating for course credit, were included in the analysis. All participants were monolingual English speakers. We excluded from analysis additional participants who failed to follow instructions (four) or gave a self-reported effort level below seven on a 10-point scale (14), as well as instances of technical failure during the experiment (two).

Materials and procedure—In Experiment 3 we presented participants with an audiovisual familiarization stream composed of the audio input stream from Experiment 1a and visual input stream from Experiment 1b. The streams were reordered such that each

triplet had an equal probability of co-occurrence with all triplets in the other modality (see Figure 3b). For example, tone triplet ABC was presented an equal number of times with shape triplets ABC, DEF, GHI, and JKL. Thus, unlike Experiment 2, there was no reliable correspondence between particular shape and tone elements. We created a loop of 48 tone triplets with a pseudorandom ordering in Praat and a loop of 48 shape triplets with a distinct pseudorandom ordering in Macromedia Director MX 2004. These loops were then combined using Adobe Premiere. The resulting clip had a duration of 4 min 48 s and consisted of 96 triplets. The movie was then exported as an 800 × 600-pixel Quicktime movie with Sorenson 3 compression.

The procedure was identical to Experiment 2. Twenty-five participants completed the audio test, and 24 participants completed the visual test.

Results and Discussion

The mean test score for the audio test in Experiment 3 was 10.28 out of 16 (64%), with a standard deviation of 1.79 (see Figure 1). A one-sample *t* test indicated that performance on the audio test was significantly above chance, $t(24) = 6.36$, $p < .001$, $d = 2.60$. Performance on the audio test in Experiment 3 was lower than performance in Experiment 1a, when the audio stream was presented in isolation (*M*s = 10.28 and 11.28, respectively), although this difference did not reach conventional standards of significance testing, $t(49) = 1.81$, $p = .076$, $d = 0.52$. The mean test score for the visual test in Experiment 3 was 11.33 out of 16 (71%), with a standard deviation of 2.55 (see Figure 1). A one-sample *t* test indicated that performance on the visual test was significantly above chance, $t(23) = 6.41$, $p < .001$, $d = 2.67$. This was not significantly different from performance in Experiment 1b, when the visual stream was presented in isolation, $t(46) = -0.79$, $p = .431$, $d = -0.23$. An independent samples *t* test between performance on the visual and audio tests revealed no significant difference, $t(47) = -1.68$, $p = .100$, $d = -0.49$. Further independent samples *t* tests revealed no difference in performance in Experiment 2 and Experiment 3 on the audio test, $t(48) = 0.69$, $p = .492$, $d = 0.19$, or visual test, $t(47) = -1.61$, $p = .115$, $d = -0.47$. A 2 (Experiment 1 vs. Experiment 3) × 2 (Visual Test vs. Auditory Test) analysis of variance (ANOVA) revealed no significant main effect of experiment, $F(1, 95) = .121$, $p = .728$, or test type, $F(1, 95) = .322$, $p = .572$. The interaction term, although approaching significance, was not statistically significant at conventional levels, $F(1, 95) = 2.84$, $p = .095$.

The results of Experiment 3 demonstrate that the findings reported in Experiment 2 were not likely due to element-to-element transfer from one modality to the other. Rather, these results suggest that learners are able to extract multiple statistical regularities simultaneously from audiovisual input. However, because triplet boundaries were aligned across streams, participants who successfully segmented one modality could have used positional information to segment the other stream (see Endress & Bonatti, 2007). In Experiment 4, we offset boundary alignment across streams to remove this source of positional information.

Another explanation for the learning observed in Experiment 3 is that audio and visual pairings may have been represented as unified, bound objects (i.e., AV1→AV2→AV3) as opposed to individual elements (i.e., A1→A2→A3 and V1→V2→V3). If learners perceived the stream in an object-based manner, then the transitional probabilities of this AV stream

would have still provided consistent word boundary cues (1.0→1.0→0.33). Although this account may seem less plausible (because of the infrequent occurrence of each AV triplet), recent evidence suggests that multistream statistical learning may be object-based when the streams are aligned (Turk-Browne et al., 2008). When the streams are partially decoupled, statistical learning switches to a feature-based parsing strategy. Thus, in Experiment 4 we examine whether multimodal statistical learning occurs when the triplet boundaries are temporally decoupled across modalities.

Experiment 4: Uncorrelated, Desynchronized Streams

In this experiment, we decouple the streams by misaligning word boundaries across the audio and visual streams. Consequently, participants should be unable to transfer positional knowledge of one stream to segment the other. Further, the transitional probabilities of the AV, object-based stream (1.0→0.33→0.33) no longer provide a consistent boundary cue, which should preclude statistical learning on coupled AV stimuli. If learners segment each input stream independently, then we would not expect any change in learning relative to previous conditions. On the other hand, if learning in the two modalities is not entirely independent, then in Experiment 4 we predict that this manipulation should disrupt successful segmentation.

Method

Participants—Fifty-one (25 female and 26 male) naïve undergraduate introductory psychology students, participating for course credit, were included in the analysis. All participants were monolingual English speakers. We excluded from analysis participants who failed to follow instructions (13) or gave a self-reported effort level below seven on a 10-point scale (17), as well as instances of technical failure during the experiment (one).

Materials and procedure—In Experiment 4 we modified the audiovisual familiarization stream from Experiment 3, such that the cross-modal coherence was disrupted by offsetting triplet boundaries across streams. This was achieved in Adobe Premiere by moving the initial segment of the visual stream to the end of the stream. This effectively shifted each visual segment forward in ordinal position relative to the audio stream, which remained the same as in Experiment 3. By adjusting the visual stream while keeping the audio stream constant, the triplet boundaries became misaligned across streams. For example, the beginning of the familiarization stream consisted of the visual elements BCD (from the triplets ABC and DEF), and the audio elements ABC. Thus, the triplet boundaries were offset, disrupting the boundary alignment between the audio and visual streams (see Figure 3c). The correspondence between individual elements in the audio and visual streams was identical to that in Experiment 3 (0.25). These two streams were then combined in Adobe Premiere and exported as an 800 × 600-pixel Quicktime movie with Sorenson 3 compression.

All other aspects of the materials and procedure were identical to those in Experiment 3. Twenty-six participants completed the audio test, and 25 participants completed the visual test.

Results and Discussion

The mean test score for the audio test in Experiment 4 was 8.96 out of 16 (56%), with a standard deviation of 3.14 (see Figure 1). A one-sample *t* test indicated that performance on the audio test was not significantly above chance, $t(25) = 1.56$, $p = .131$, $d = 0.62$. This was significantly lower than performance in Experiment 1a, when the audio stream was presented in isolation, $t(50) = -3.13$, $p = .003$, $d = 0.89$. The mean test score for the visual test in Experiment 4 was 8.92 out of 16 (56%), with a standard deviation of 2.48 (see Figure 1). A one-sample *t* test indicated that performance on the visual test was not significantly above chance according to conventional standards, $t(24) = 1.85$, $p = .076$, $d = 0.76$. It is notable that this was significantly different from performance in Experiment 1b, when the visual stream was presented in isolation, $t(47) = 2.32$, $p = .025$, $d = 0.68$. An independent samples *t* test between performance in Experiment 4 on the visual and audio tests revealed no significant difference, $t(49) = .052$, $p = .959$, $d = 0.01$.

It is possible that the at-chance performance was a consequence of averaging the scores of subgroups of participants that learned one of the two streams (e.g., half the participants in the auditory test condition successfully learned the auditory stream, and the other half learned the visual stream). To test whether participants in Experiment 4 were learning one or neither of the two streams, we examined the distribution of scores in each test. Separate one-sample Kolmogorov-Smirnov analyses for the auditory and visual tests revealed that each was normally distributed around the group mean (auditory $K\text{-}S = .64$, $p = .801$; visual $K\text{-}S = .78$, $p = .582$). This suggests that participants failed to successfully segment either of the two streams at above chance levels.

A one-way ANOVA across all experiments revealed a significant difference in audio test score, $F(3, 98) = 4.12$, $p = .009$. A second one-way ANOVA revealed a significant difference in visual test scores across all experiments $F(3, 94) = 3.65$, $p = .015$. Planned contrasts[1] confirmed that, for each test type, performance in Experiment 4 was significantly lower than performance in Experiments 2 and 3, audio test $t(98) = -2.61$, $p = .011$, $d = -0.53$; visual test $t(94) = -2.76$, $p = .007$, $d = -0.57$.[2]

In sum, the results of Experiment 4 demonstrate that when the audio and visual streams were offset, the learning observed in Experiments 2 and 3 was attenuated. Learners require alignment of boundaries to successfully segment both streams. These results indicate that learning is not independent across modalities.

General Discussion

The primary goals of this research were to test systematically whether multistream statistical learning is processed independently for each modality and to explore the types of cross-

[1]The weights for the contrast analysis were [0, −1, −1, 2] for Experiments 1, 2, 3, and 4, respectively.

[2]Because of the unusually high number of excluded participants in this experiment, we conducted a similar set of analyses that included participants removed on the basis of self-reported effort. This analysis was not substantially different from the filtered analyses. A one-way ANOVA revealed a significant difference in performance between experiments, audio $F(3, 124) = 3.89$, $p = .011$; visual $F(3, 116) = 3.06$, $p = .031$. Contrast analyses with identical weights each revealed that performance in Experiment 4 was significantly lower than in Experiments 2 and 3, audio test $t(124) = -2.10$, $p = .038$, $d = -0.53$; visual test $t(116) = -2.35$, $p = .032$, $d = -0.57$.

modal relationships that influence simultaneous learning of multimodal input streams. In Experiment 1, we provided baseline levels of performance for segmentation of both a tone and shape input stream presented in isolation. In Experiment 2, we presented learners with an audiovisual stream that consisted of the simultaneous presentation of the audio and visual streams from Experiment 1 and maintained a one-to-one correspondence between elements across modalities. We found that learners were able to segment both the visual and auditory input streams successfully. In Experiment 3, we removed the correspondence between individual elements in the audio and visual streams, a manipulation that did not result in a decrement in performance relative to Experiment 2, as learners were equally successful in learning both streams. In Experiment 4, we further disrupted the relationship between the audio and visual streams by offsetting the triplet boundaries. In this condition, learners were unable to segment either the audio or visual streams successfully at above chance levels.

These findings inform theoretical accounts of cross-modal statistical learning, particularly regarding the extent to which learning in one modality is achieved independently of learning in a second modality. Previous studies have claimed that during multimodal statistical learning, each stream is segmented independently. According to this view, statistical learning within a particular modality should be unaffected by cross-modal relationships, and consequently learners should be capable of simultaneously segmenting two streams in different modalities without suffering any decline in performance because of cross-modal associations (Seitz et al., 2007). However, the results presented here in Experiment 4 fail to confirm this theory; learners were unable to segment either stream when presented with multimodal input in which the boundaries were not aligned. Thus, we conclude that multisensory statistical learning appears to be contingent on some degree of cross-modal coherence and therefore cannot be described as independent (*sensu* Seitz et al., 2007).

We note that the critical comparisons in this series of studies are made across experiments, and such comparisons must be interpreted with caution. Ideally, participants for these experiments would have been randomly selected from our student population at the same point in time, as systematic disparities among subject populations or experimental conditions could have resulted in different levels of segmentation performance. However, the four experiments here were largely run sequentially, as the results from the early experiments dictated the design of the subsequent experiments (e.g., if there had been no learning in Experiment 3, it would not have made sense to conduct Experiment 4). Consequently, there was a significant difference in the time of the semester at which each experiment was conducted, $F(3, 196) = 26.89$, $p < .001$, with Bonferroni post hoc analyses indicating that Experiment 4 was completed significantly later in the semester than Experiments 1–3 (ps <.05). That said, each experiment (including Experiment 4) was conducted over a range of times within the semester, and a subset of participants in Experiments 2 and 3 were tested after the completion of Experiment 4. Given this variability, we examined whether the time in the semester would correlate with segmentation performance. Correlation analyses across all four Experiments revealed that performance on neither the audio test, $r(100) = -.051$, $p = .608$, nor the visual test, $r(96) = -.141$, $p = .167$, significantly correlated with the time of the semester during which the experiments were conducted. Further, performance on the audio or visual tests did not

significantly correlate with day of semester within each experiment (*p*s >.05). There was also no significant difference in the size of the Simon effect across experiments, $F(3, 195) = 0.87$, $p = .458$. Thus, although we acknowledge the need for caution when making comparisons between experiments, we do not have any reason to suspect that there were significant between-experiment differences in the subject population. The time of the semester did not appear to influence performance, and we did not observe a significant difference in performance on the Simon task. Future research using random assignment can verify this claim.

The second goal of this research was to identify the types of cross-modal relationships that affect multisensory statistical learning. As described above, we systematically varied cross-modal relationships across three experimental conditions, reducing the correspondence between individual segments (Experiment 3) and misaligning boundary information (Experiment 4) across streams. We found no evidence that removing the correspondence between elements impacted learning, yet misaligning the boundaries disrupted learning. This suggests that boundary information is critical for multimodal statistical learning. Below, we speculate about why we observed this pattern of learning across Experiments 3 and 4.

A possible explanation for the observed pattern of results is that the successful multisensory statistical learning in Experiments 2 and 3 was achieved by learning a single stream and then transferring knowledge from that stream to segment the other. Had participants attempted to transfer word-boundary knowledge across streams in Experiment 4, then this strategy would have resulted in an incorrect mapping of word boundaries to the secondary stream and a potential decrement in performance. However, to adopt a transfer strategy, we might have expected learners to segment one of the two streams successfully, effecting an asymmetrical pattern of learning. Our results do not support this assertion. Neither the aggregated results nor additional Kolmogorov-Smirnov analyses on the distribution of scores in Experiment 4 provide evidence of an asymmetry in learning. Instead, our findings indicate that learners in Experiment 4 segmented neither the visual nor the audio stream. These results cast doubt on the notion that participants were using a transfer strategy to achieve multisensory statistical learning in Experiments 2 and 3.

It is also possible that the results of Experiments 3 and 4 reflect learners' use of an object-based parsing strategy (see Turk-Browne et al., 2008). The transitional probabilities between audiovisual objects in Experiment 3 provided robust cues to triplet boundaries (1.0→1.0→.33), whereas they did not in Experiment 4 (1.0→.33→.33). Therefore, if segmentation was based on transitional probabilities calculated over integrated AV objects, then the absence of transitional probability cues in Experiment 4 could account for the decrement in learning. Although we do not rule out this possibility entirely, the large number of possible audiovisual triplets (16) in Experiment 3 and the relative infrequency of each audiovisual triplet (18 total instances throughout familiarization) make this account appear less tenable. There is recent evidence that increasing the number of words and decreasing the frequency of exposure to individual words in a segmentation task both result in a decrement in learning (Frank, Goldwater, Griffiths, & Tenenbaum, 2010). These findings are consistent with Bayesian models of statistical learning that assign weight to items according to the model's degree of confidence in that word (e.g., Goldwater, Griffiths, & Johnson, 2009; Orbán,

Fiser, Aslin, & Lengyel, 2008). By extension, if participants in the present study were segmenting integrated AV objects, then we might have expected to observe a decline in performance in Experiment 3 (which contained 16 AV words each occurring 18 times) relative to Experiment 2 (four AV words each occurring 72 times) because of an increase in the number of words and a decrease in word frequency. Because learning was equivalent in Experiments 2 and 3, our results are not entirely consistent with an audiovisual object-based parsing strategy.

Another alternative explanation of our results is that the decrement in performance in Experiment 4 may have stemmed from increased task demands or attentional constraints relative to Experiments 2 and 3. There is evidence that attention is necessary for successful statistical learning (e.g., Turk-Browne et al., 2005). If the manipulation in Experiment 4 served to increase the demands on attention, then this may have impaired participants' ability to segment each stream. For example, if learners were switching attention between streams during learning, then perhaps the costs associated with shifting attention were greater when triplet boundaries were misaligned across streams in Experiment 4. Consequently, the decline in learning observed in Experiment 4 might not disconfirm modality independence. Although we cannot entirely rule out this possibility, in our view it is unlikely to account for our findings. It is important to note that there is a great deal of individual variation in selective attention and task-switching capacities (see Lu & Proctor, 1995; Simon & Small, 1969)—some people are better able to ignore an unattended stimulus or task and focus their attention on the target stimulus/task. If the absence of learning in Experiment 4 was due to increased difficulty selectively attending to each stream or switching back and forth between the streams, then we might have expected participants' segmentation ability to be correlated with their selective attention capacity as indexed by performance on the Simon task (see Weiss et al., 2010, for an example of how the Simon task may correlate with performance on a statistical learning task). However, there was no significant correlation between performance on the segmentation and Simon tasks in Experiment 4—audio test $r(24) = .113$, $p = .584$; visual test $r(23) = -.344$, $p = .092$—although the correlation between the visual task and the Simon task approached significance. This casts doubt on the notion that increased demands on attentional resources can fully explain the decrement in performance observed in Experiment 4.

Related to constraints on attention, the decrement in performance in Experiment 4 could reflect increased task demands. As noted earlier, a greater number of participants in Experiment 4 were excluded from analysis for failure to follow instructions (e.g., falling asleep) or for low self-reported effort levels. This higher level of attrition may indicate that the task in Experiment 4 was more difficult than the task in the previous experiments. However, the only difference in task demands between Experiment 3 and Experiment 4 was the relative position of boundary information across streams. According to a modality-independent account of statistical learning, this manipulation should not introduce a difference in task demands across experiments. Thus, even if participants failed to segment the audio and visual streams due to an increase in the task demands of Experiment 4, then this suggests that statistical learning is not modality-independent (*sensu* Seitz et al., 2007).

Finally, it is possible that the observed decrement in performance in Experiment 4 may have arisen because of the importance of boundary information for statistical learning. Word offsets are known to be highly salient events for the process of speech segmentation (Cunillera, Gomilla, & Rodriguez-Fornells, 2008; see also, Echols & Newport, 1992); thus, it is possible that the misaligned boundaries amplified the difficulty of tracking multiple streams simultaneously. This possibility is supported by a recent study of speech segmentation in which learners were presented with a speech stream paired with static visual images (Cunillera et al., 2010). Performance declined when the onset and offset of the static pictures were misaligned with the onset and offset of words in the speech stream. Like the results presented here, these findings suggest that cross-modal boundary alignment is an important component for multimodal statistical learning. The importance of boundary alignment is further demonstrated by a recent proposal that learners may extract positional information in conjunction with (or in addition to) statistical information (see Endress & Mehler, 2009). According to this theory, learners encode edge-based positional information and elements that occur at an edge undergo specialized processing (Endress & Mehler, 2010). The present study does not address whether learners possess specialized mechanisms for detecting edges, but the results of Experiment 4, in which incongruent positional information disrupts learning, are consistent with the idea that alignment of edges or positional codes may exert a strong influence on participants' abilities to successfully segment a multimodal stream.

Irrespective of the underlying source of these findings, our results confirm that adult learners are capable of multisensory statistical learning. This conclusion is consistent with recent demonstrations that statistical learning is sensitive to multimodal input. Although early studies of statistical learning typically focused on learning within a single modality (e.g., Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996), more recent studies have confirmed that visual input can impact auditory statistical learning (e.g., Cunillera et al., 2010; Hollich, Newman, & Jusczyk, 2005; Mitchel & Weiss, 2010; Sell & Kaschak, 2009). To the best of our knowledge, only one other study (Seitz et al., 2007) has explored multisensory statistical learning in which statistics are simultaneously tracked in separate modalities, although there is evidence that adults can learn two interleaved artificial grammars presented in separate modalities (Conway & Christiansen, 2006).

If, as our results suggest, multimodal statistical learning is not achieved via independent mechanisms, what are the consequences for statistical learning? One possibility is that statistical learning may benefit from the availability of multisensory input. It has recently been proposed that perceptual learning mechanisms most likely evolved to operate optimally over multimodal input, reflecting the multimodal nature of both the learning environment (Shams & Seitz, 2008) and perceptual systems (Stein & Stanford, 2008). In support of this claim, there is evidence that rule-learning mechanisms benefit from multisensory input. Frank et al. (2010) found that infants were able to acquire rule structures (e.g., ABA or AAB) from multisensory input (speech and tones) at an earlier age than they were able to acquire similar structures from unisensory input. This suggests that unimodal assessments of language-learning capacities may underestimate infants' ability to acquire structure from the environment and studies of early perceptual learning may benefit from using multimodal

displays (e.g., Teinonen, Aslin, Alku, Csibra, 2008). Thus, future work should investigate the impact of multisensory input during development.

Our results may also provide insight into whether statistical learning has its basis in one central mechanism or in several sensory-specific mechanisms. Had learners been successful in all conditions, regardless of our manipulations of cross-modal relationships, then it would have provided compelling evidence for a modality-specific view of statistical learning (e.g., Conway & Christiansen, 2005, 2006). However, lack of evidence for modality independence suggests that statistical learning includes at least some modality-general component (i.e., the strictest modality-specific account, as implied by Seitz et al., 2007, does not seem viable). Nevertheless, we do not rule out the possibility that statistical learning might also include a modality-specific component. Indeed, the equivalent level of learning in Experiment 3 and Experiment 2 is consistent with modality-specific accounts of statistical learning. In Experiment 3, we removed the correlation between the individual elements of the strings, yet this did not affect participants' ability to segment each stream when compared with performance in Experiment 2, in which the streams were perfectly correlated. Although this is compatible with modality-specific learning mechanisms, a modality-general account might have predicted that removing the correlation between streams should have disrupted learning. Thus, our results lend support for a comprehensive model of statistical learning that is composed of both modality-specific and modality-general components. Such models emphasizing the mixture of modality-specific and modality-general influences have been posited in other domains, including attention (e.g., Driver & Spence, 1998) and early sensory perception (Besle, Fort, & Giard, 2005). It may be possible to adjudicate among potential models of multisensory statistical learning by examining the time course of cross-modal effects to determine the point at which information is integrated across senses. Using the temporal precision of EEGs, planned studies will measure responses to multimodal stimuli in early sensory perception (see Besle et al., 2005) to detect whether information is integrated immediately or further downstream.

In summary, statistical learning operates within a multidimensional, multimodal sensory environment; thus, we asked whether learners can parse multiple input streams across modality at the same time and, if so, whether these streams are learned independently. Our experiments demonstrate that learners are capable of simultaneously segmenting two input streams presented in separate modalities. However, this ability demands a certain level of cross-modal coherence, as disrupting boundary information across modalities attenuates learning, providing evidence against modality independence during multisensory statistical learning.

Acknowledgments

We would like to thank Beth Buerger, Molly Jamison, and Troy Gury for conducting experiments. We would also like to thank Morten Christiansen for helpful comments during the preparation of this article. This research was supported by National Institutes of Health Grant R03 HD048996-01 to Daniel J. Weiss.

References

Aslin, RN.; Newport, EL. What statistical learning can and can't tell us about language acquisition. In: Colombo, J.; McCardle, P.; Freund, L., editors. Infant pathways to language: Methods, models, and research disorders. Erlbaum; New York, NY: 2009. p. 15-29.

Barsalou LW, Simmons WK, Barbey AK, Wilson CD. Grounding conceptual knowledge in modality-specific systems. Trends in Cognitive Sciences. 2003; 7:84–91. doi:10.1016/S1364-6613(02)00029-3. [PubMed: 12584027]

Besle J, Fort A, Giard M-H. Is the auditory sensory memory sensitive to visual information? Experimental Brain Research. 2005; 166:337–344. doi:10.1007/s00221-005-2375-x. [PubMed: 16041497]

Boersma, P.; Weenink, D. Praat: Doing phonetics by computer (Version 5.0.35) [Computer program]. 2008. Retrieved from http://www.praat.org/

Conway CM, Christiansen MH. Modality-constrained statistical learning of tactile, visual, and auditory sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005; 31:24–39. doi:10.1037/0278-7393.31.1.24.

Conway CM, Christiansen MH. Statistical learning within and between modalities: Pitting abstract against stimulus specific representations. Psychological Science. 2006; 17:905–912. doi:10.1111/j.1467-9280.2006.01801.x. [PubMed: 17100792]

Conway CM, Christiansen MH. Seeing and hearing in space and time: Effects of modality and presentation rate on implicit statistical learning. European Journal of Cognitive Psychology. 2009; 21:561–580.

Creel SC, Newport EL, Aslin RN. Distant melodies: Statistical learning of non-adjacent dependencies in tone sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2004; 30:1119–1130. doi:10.1037/0278-7393.30.5.1119.

Cunillera T, Càmara E, Laine M, Rodríguez-Fornells A. Speech segmentation is facilitated by visual cues. The Quarterly Journal of Experimental Psychology. 2010; 63:260–274. doi:10.1080/17470210902888809. [PubMed: 19526435]

Cunillera T, Gomila A, Rodríguez-Fornells A. Beneficial effects of word final stress in segmenting a new language: Evidence from ERPs. BMC Neuroscience. 2008; 9:23. [PubMed: 18282274]

Driver J, Spence C. Crossmodal attention. Current Opinion in Neurobiology. 1998; 8:245–253. doi:10.1016/S0959-4388(98)80147-5. [PubMed: 9635209]

Echols CH, Newport EL. The role of stress and position in determining first words. Language Acquisition. 1992; 2:189–220. doi: 10.1207/s15327817la0203_1.

Endress AD, Bonatti LL. Rapid learning of syllable classes from a perceptually continuous speech stream. Cognition. 2007; 105:247–299. doi:10.1016/j.cognition.2006.09.010. [PubMed: 17083927]

Endress AD, Mehler J. The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. Journal of Memory and Language. 2009; 60:351–367. doi:10.1016/j.jml.2008.10.003.

Endress AD, Mehler J. Perceptual constraints in phonotactic learning. Journal of Experimental Psychology: Human Perception and Performance. 2010; 36:235–250. doi:10.1037/a0017164. [PubMed: 20121307]

Fiser J, Aslin RN. Statistical learning of higher order temporal structure from visual shape-sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2002; 28:458–467. doi:10.1037/0278-7393.28.3.458.

Frank MC, Goldwater S, Griffiths T, Tenenbaum JB. Modeling human performance in statistical word segmentation. Cognition. 2010; 117:107–125. doi:10.1016/j.cognition.2010.07.005. [PubMed: 20832060]

Gebhart AL, Aslin RN, Newport EL. Changing structures in mid-stream: Learning along the statistical garden path. Cognitive Science. 2009; 33:1087–1116. doi:10.1111/j.1551-6709.2009.01041.x. [PubMed: 20574548]

Goldwater S, Griffiths T, Johnson M. A Bayesian framework for word segmentation: Exploring the effects of context. Cognition. 2009; 112:21–54. doi:10.1016/j.cognition.2009.03.008. [PubMed: 19409539]

Hollich G, Newman RS, Jusczyk PW. Infants' use of synchronized visual information to separate streams of speech. Child Development. 2005; 76:598–613. doi:10.1111/j.1467-8624.2005.00866.x. [PubMed: 15892781]

Hunt RH, Aslin RN. Statistical learning in a serial reaction time task: Simultaneous extraction of multiple statistics. Journal of Experimental Psychology: General. 2001; 130:658–680. doi:10.1037/0096-3445.130.4.658. [PubMed: 11757874]

Kirkham NZ, Slemmer JA, Johnson SP. Visual statistical learning in infancy: Evidence for a domain-general learning mechanism. Cognition. 2002; 83:B35–B42. doi:10.1016/S0010-0277(02)00004-5. [PubMed: 11869728]

Lu C-H, Proctor RW. The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. Psychonomic Bulletin & Review. 1995; 2:174–207. doi:10.3758/BF03210959. [PubMed: 24203654]

Mitchel AD, Weiss DJ. What's in a face? Visual contributions to speech segmentation. Language and Cognitive Processes. 2010; 25:456–482. doi:10.1080/01690960903209888.

Orbán G, Fiser J, Aslin RN, Lengyel M. Bayesian learning of visual chunks by human observers. Proceedings of the National Academy of Sciences, USA. 2008; 105:2745–2750.

Saffran JR. Statistical language learning: Mechanisms and constraints. Current Directions in Psychological Science. 2003; 12:110–114. doi:10.1111/1467-8721.01243.

Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. Science. 1996; 274:1926–1928. doi:10.1126/science.274.5294.1926. [PubMed: 8943209]

Saffran JR, Johnson EK, Aslin RN, Newport EL. Statistical learning of tone sequences by human infants and adults. Cognition. 1999; 70:27–52. doi:10.1016/S0010-0277(98)00075-4. [PubMed: 10193055]

Saffran JR, Newport EL, Aslin RN. Word segmentation: The role of distributional cues. Journal of Memory and Language. 1996; 35:606–621. doi:10.1006/jmla.1996.0032.

Seitz AR, Kim R, van Wassenhove V, Shams L. Simultaneous and independent acquisition of multisensory and unisensory associations. Perception. 2007; 36:1445–1453. doi:10.1068/p5843. [PubMed: 18265827]

Sell AJ, Kaschak MP. Does visual speech information affect word segmentation? Memory & Cognition. 2009; 37:889–894. doi:10.3758/MC.37.6.889. [PubMed: 19679867]

Shams L, Seitz AR. Benefits of multisensory learning. Trends in Cognitive Sciences. 2008; 12:411–417. [PubMed: 18805039]

Simon JR, Small AM. Processing auditory information: Interference from an irrelevant cue. Journal of Applied Psychology. 1969; 53:433–435. doi:10.1037/h0028034. [PubMed: 5366316]

Stein BE, Stanford TR. Multisensory integration: Current issues from the perspective of the single neuron. Nature Reviews Neuroscience. 2008; 9:255–266. doi:10.1038/nrn2331.

Teinonen T, Aslin RN, Alku P, Csibra G. Visual speech contributes to phonetic learning in 6-month-old infants. Cognition. 2008; 108:850–855. doi:10.1016/j.cognition.2008.05.009. [PubMed: 18590910]

Thiessen ED. Effects of visual information on adults' and infants' auditory statistical learning. Cognitive Science. 2010; 34:1093–1106. doi: 10.1111/j.1551-6709.2010.01118.x. [PubMed: 21564244]

Toro JM, Sinnett S, Soto-Faraco S. Speech segmentation by statistical learning depends on attention. Cognition. 2005; 97:B25–B34. doi: 10.1016/j.cognition.2005.01.006. [PubMed: 16226557]

Turk-Browne NB, Isola PJ, Scholl BJ, Treat TA. Multidimensional visual statistical learning. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2008; 34:399–407. doi:10.1037/0278-7393.34.2.399.

Turk-Browne NB, Jungé JA, Scholl BJ. The automaticity of visual statistical learning. Journal of Experimental Psychology: General. 2005; 134:552–564. doi:10.1037/0096-3445.134.4.552. [PubMed: 16316291]

Weiss DJ, Gerfen C, Mitchel AD. Speech segmentation in a simulated bilingual environment: A challenge for statistical learning? Language Learning and Development. 2009; 5:30–49. doi: 10.1080/15475440802340101. [PubMed: 24729760]

Weiss DJ, Gerfen C, Mitchel AD. Colliding cues in word segmentation: The role of cue strength and general cognitive processes. Language and Cognitive Processes. 2010; 25:402–422. doi: 10.1080/01690960903212254.

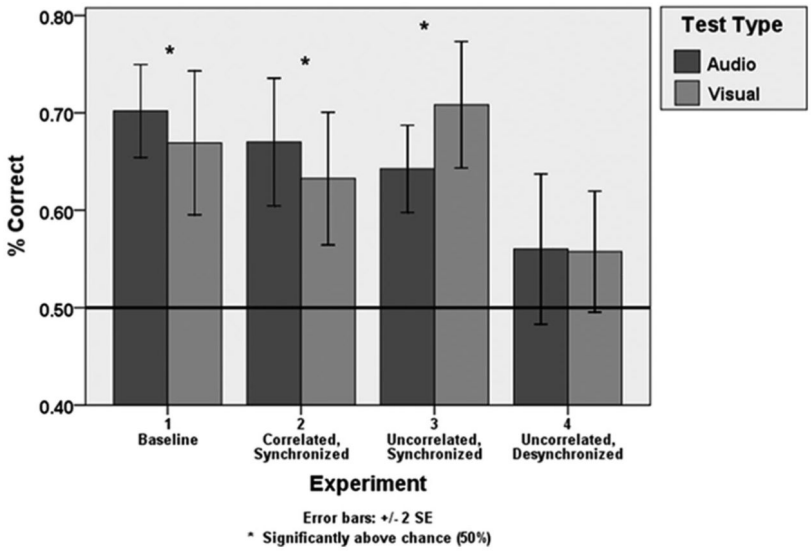

Figure 1.

Percentage correct identification in a two-alternative forced-choice task across four experiments. The bar represents chance (50%). Error bars represent ±2 SE. Asterisk indicates significantly above chance.

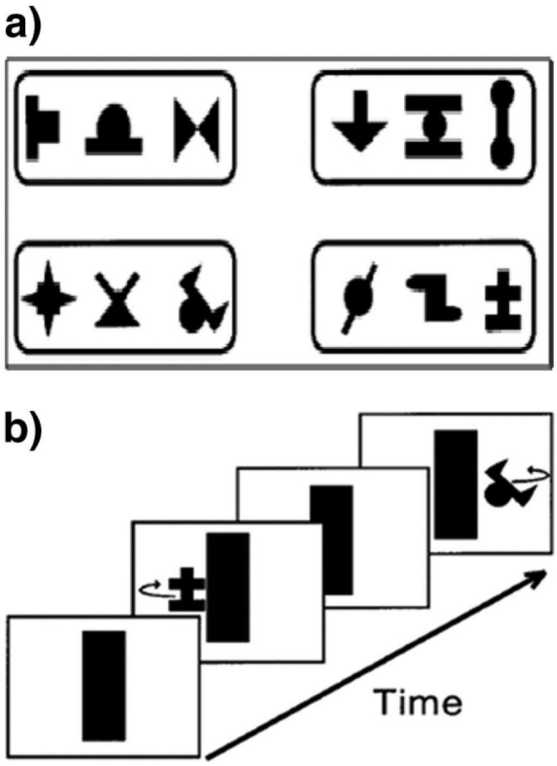

Figure 2.

a. Tokens used to construct the visual input stream. Tokens are grouped in the statistically defined triplets. b. Example temporal sequence of the visual stream. A shape moves from behind a central occluder to the edge of the window and back, at which point the next shape in the sequence emerges from the occluder in the opposite direction. From “Statistical Learning of Higher Order Temporal Structure From Visual Shape-Sequences,” by J. Fiser and R. N. Aslin, 2002, Journal of Experimental Psychology: Learning, Memory, and Cognition, 28, p. 460. Copyright 2002 by the American Psychological Association. Reprinted with permission.

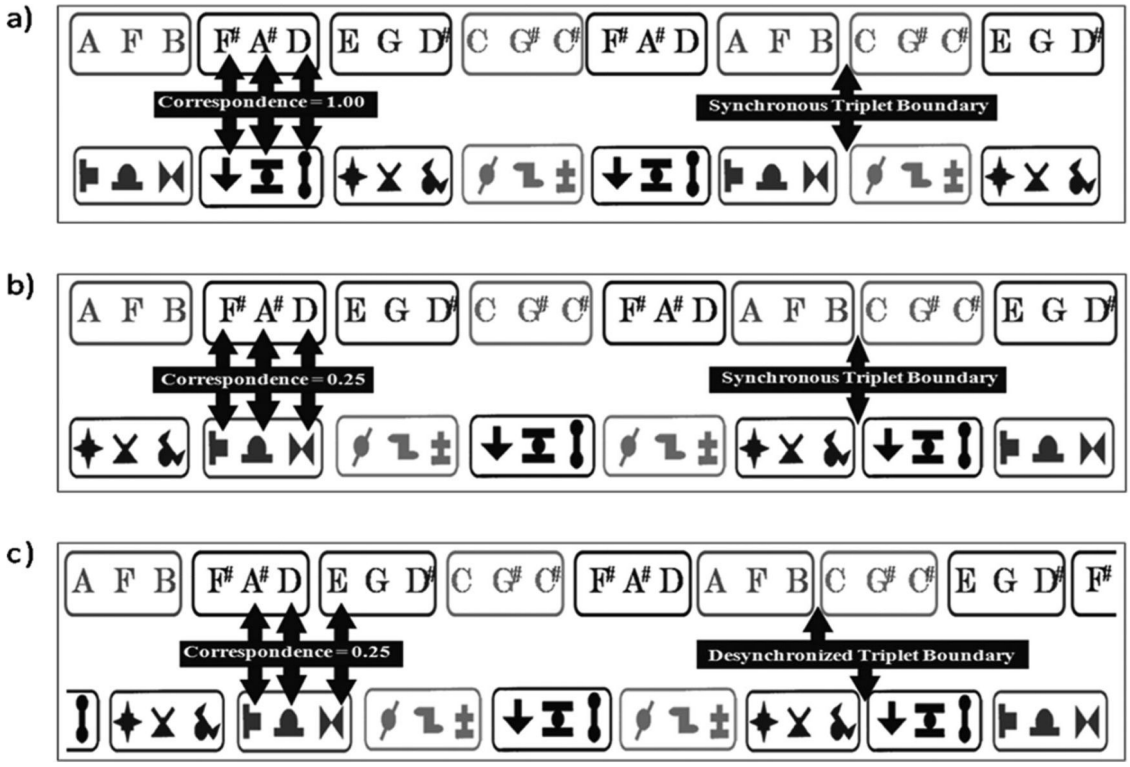

Figure 3.
This figure illustrates the correspondence and boundary alignment across modalities in Experiment 2 (a), Experiment 3 (b), and Experiment 4 (c).