Crossmodal statistical learning is facilitated by modality predictability

Daniel Duato [2]; Francesco Giannelli [2, 3]; Ruth de Diego-Balaguer [2,3,4,5]; Alexis Pérez-Bellido [1,2,3,4]

* Correspondence concerning this article should be addressed to Dr. Alexis Pérez-Bellido, [1] Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain. [2] Department of Cognition, Development and Educational Psychology, University of Barcelona, Spain. [3] Cognition and Brain Plasticity Unit [Bellvitge Biomedical Research Institute – IDIBELL], 08097 L'Hospitalet de Llobregat, Barcelona, Spain. [4] Institute of Neurosciences, University of Barcelona, Barcelona, Spain. [5] ICREA, Barcelona, Spain. Email: alexisperezbellido@gmail.com

Abstract

Statistical learning (SL) refers to the ability to extract statistical regularities from the environment. Previous research has suggested that regularity extraction is modality-specific, occurring within but not between sensory modalities (Frost et al., 2015). The present study investigates the circumstances under which SL can occur between modalities. In the first experiment, participants were presented with a stream of meaningless visual fractals and synthetic sounds while performing an oddball detection task. Stimuli were grouped into unimodal (AA, VV) or crossmodal (VA, AV) pairs based on higher transitional probability between the elements. Using implicit and explicit measures of SL, we found that participants only learned the unimodal pairs. In a second experiment, we presented the pairs in separate unimodal (VVVV, AAAA) and crossmodal (AVAV, VAVA) blocks, allowing participants to anticipate which modality would be presented next. We found that SL for the crossmodal pairs outperformed that of unimodal pairs. This result suggests that modality predictability facilitates a correct crossmodal attention deployment that is crucial for learning crossmodal transitional probabilities. Finally, a third experiment demonstrated that participants can explicitly learn the statistical regularities between crossmodal pairs even when the upcoming modality is not predictable, as long as the pairs contain semantic information. This finding suggests that SL between crossmodal pairs can occur when sensory-level limitations are bypassed, and when learning can unfold at a supramodal level of representation. This study demonstrates that SL is not a modality-specific mechanism and compels revision of the current neurobiological model of SL in which learning of statistical regularities between low-level stimuli features relies on hard-wired learning computations that take place in their respective sensory cortices.

Keywords: Statistical learning, multisensory, prediction

Introduction

Environmental information is highly structured conforming temporal and spatial patterns. The human brain has developed an exquisite sensitivity to implicitly detect these statistical regularities and exploit them to efficiently anticipate upcoming environmental changes. This ability, termed statistical learning (SL), has been thoroughly investigated in the last decades showing that the human brain is ready to pick-up on statistical regularities at very early ages of development (Saffran et al., 1996). However, whereas most of the studies have successfully shown that humans and other animals can learn statistical regularities in the auditory, visual or tactile modalities (Conway & Christiansen, 2005; Fiser & Aslin, 2001; Saffran et al., 1999), the evidence showing that SL also takes place between sensory modalities is limited.

Most of the current empirical evidence supports the hypothesis that SL is modality-specific (Frost et al., 2015; Walk & Conway, 2016). For instance, prior research has demonstrated that humans can concurrently learn statistical regularities independently in two sensory modalities without interference (Conway & Christiansen, 2006; Mitchel & Weiss, 2011; Seitz et al., 2007). In addition, statistical learning of an artificial grammar does not transfer across modalities (Redington & Chater, 1996; Siegelman & Frost, 2015). However, SL in one modality is not entirely independent of stimulation in other sensory modalities when statistical relations are not consistent between modalities. Multiple studies have shown that asynchronous temporal alignment of boundaries between independent learning units in different sensory modalities can hinder the learning of transitional probabilities in either the visual or auditory modality(Cunillera, Càmara, et al., 2010; Cunillera, Laine, et al., 2010; Glicksohn & Cohen, 2013; Mitchel & Weiss, 2011; Thiessen, 2010). In addition, there is indirect evidence indicating that the learning of transitional probabilities can take place between audio-visually inconsistent but integrated percepts such as the McGurk illusion (Mitchel et al., 2014). Although this learning may be occurring at a supra-modal level it implies indirect evidence supporting SL between modalities. Nevertheless, other than that, to date, there is no empirical evidence showing that humans can detect statistical regularities between inputs presented in different sensory modalities (Walk & Conway, 2016).

As a result, the most accepted framework in the literature considers that while statistical learning relies on a domain-general set of computations, learning of the statistical associations between low-level sensory features is modular and constrained by the specific characteristics of each sensory (Frost et al., 2015). From this framework, crossmodal statistical learning would be exclusively limited to sensory input that encode high-level representations of meaningful stimuli, and consequently would occur at multimodal or amodal brain regions.

In this study, we aim to scrutinize this framework in three experiments while testing two complementary hypotheses that might explain why finding statistical learning between low-level sensory inputs is challenging. Firstly, previous studies on crossmodal statistical learning have only examined whether participants could explicitly recognize crossmodal associated items. However, there is extensive evidence that statistical learning can occur implicitly during the learning phase, facilitating the detection of predicted targets or the discrimination of unexpected oddballs (López-Barroso et al., 2016; Orpella et al., 2021; Richter et al., 2018; Richter & de Lange, 2019). Thus, we hypothesize that if crossmodal SL is more implicit than unimodal SL, we must quantify SL using an online and implicit SL measure (Siegelman et al., 2017). Second, crossmodal SL might be impeded compared to unimodal SL processing multimodal stimuli sequences due to a modality switching cost. This is, in randomly intermixed sequences of multimodal trials, the reaction times are slower when the target stimulus is preceded by a stimulus of a different than by from the same modality (Gondan et al., 2004; Otto & Mamassian, 2012). This phenomenon is associated with an attentional bias to expect information in the last processed modality (Spence et al., 2001). In previous studies investigating crossmodal SL (Seitz et al., 2007; Walk & Conway, 2016), multimodal stimuli were randomly intermixed during the learning phase. Hence, the existence of modality switching costs might have induced biases to perceptually group together inputs delivered within the same sensory modality, limiting the capacity to learn crossmodal transitions.

Therefore, to test our first hypothesis we used an online measure of statistical learning by registering participants' performance and reaction times in detecting predictable and unpredictable oddball trials during exposure to the TPs in an implicit learning task. Importantly, we modeled whether participants' detection of predictable compared with unpredictable oddballs improved with progressive exposure during this implicit phase. In the second experiment, we tested our second hypothesis employing a similar experimental design as in the first experiment, but we structured the stimuli modality combinations in

different blocks, allowing participants to anticipate modality alternations. This manipulation aimed to avoid biases to perceptually group together items from the same sensory modality. Finally, in a third experiment, we tested the hypothesis, proposed by Frost et al., (2015), that crossmodal statistical learning can occur when the stimuli from each modality have a meaningful higher level representation.

Foreshadowing our results, we found that when unimodal and crossmodal stimuli transitions are randomly intermixed in the learning phase, participants only reliably learned the associations between the unimodal pairs (Seitz et al., 2007; Walk & Conway, 2016). However, when crossmodal transitions were predictable, participants could implicitly learn the association between crossmodal pairs. Finally, we demonstrated that participants can learn crossmodal transitional probabilities even when the upcoming modality is unpredictable if the sensory inputs can be semantically represented in the brain suggesting that SL also occurs between higher amodal representations.

Experiment 1

In the first experiment we used both an implicit online and an explicit offline measure of SL to assess whether participants acquired crossmodal statistical regularities between low-level sensory information. In order to be sure that the SL relations could only be learned at the sensory level and not at an abstract level of representation, participants were exposed to a continuous stream of non-meaningful synthetic stimuli (i.e. fractals in the visual modality and synthetic sounds in the auditory modality). The stream contained hidden adjacent dependencies between stimuli pairs such that each stimuli in the pair was deterministically associated to another one with 100% probability. After the exposure, participants were tested about their knowledge of such dependencies. We included unimodal statistical dependencies to validate our experimental paradigm. Therefore, we used four different modality combinations (Fig. 1B): auditory to auditory (AA), visual to visual (VV), auditory to visual (AV) and visual to auditory (VA). All four conditions were pseudo-randomly sorted conforming one single stream (Fig. 1D).

Importantly, during the implicit learning phase, participants were not informed about the presence of associated pairs and their task was simply to pay attention to both modalities in order to detect noisy oddball stimuli (both visual and auditory). The task was orthogonal to the hidden structure of pairs, that randomly appeared during the stream (an example at the upper right, Fig. 1A). The inclusion of this orthogonal task allowed us to use the reaction time (RT) to oddballs as an implicit measure of learning: We anticipated faster detection of the

oddballs when stimuli could be expected than when they were unexpected. Moreover, it ensured that participants equally attended the auditory and visual stimuli without explicitly intending to extract the hidden statistical regularities. After the implicit learning phase, we assessed whether participants could explicitly recognize the pairs with a two-interval forced choice task (2 AIC) (Fig. 1E). Given two pairs of stimuli, participants had to identify which one was a pair they were previously exposed to. In addition, we aimed to determine the participants' confidence level in their decisions. To measure this, participants were asked to report their level of confidence about their answer after each test trial.

Methods

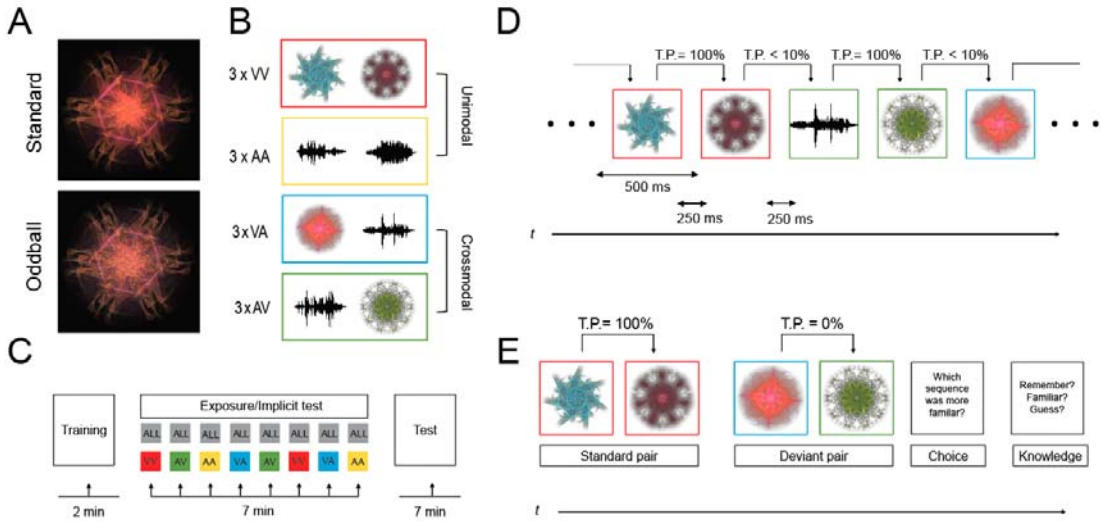

Figure 1.**A** Example of a standard visual stimulus and its oddball version in the experiments 1 and 2 **B** Twelve visual and 12 auditory meaningless synthetic (experiment 1 and 2) or meaningful natural (experiment 3; not displayed here) stimuli were combined conforming 12 standard pairs (3 x each modality combination). **C** Each experiment consisted in three parts: First, a training phase in which the participants were familiarized with the task and the different stimuli. Second, an implicit learning phase. In experiment 1 and 3, during the implicit learning phase, all the modality combinations were randomly intermixed in each block. In experiment 2 only one modality combination was presented in each block. Third, after the implicit learning phase, participants performed an explicit recognition test. **D** Example of a stimuli sequence during the implicit learning phase in the first experiment (for illustrative purposes, color frames denote standard pairs as represented in A). **E** An example trial of the explicit 2AIC test. The “old” pair is one of the previously presented pairs and a “new” pair is a new stimuli combination that was never presented during the implicit learning phase. The color frame is added here for illustrative purposes: red for VV, yellow for AA pairs, green for AV and blue for VA pairs.

Participants

Forty-five university students participated in experiment 1 (40 females, 5 males, mean age 20.36, range 18-28). None of the participants reported history of auditory problems, and all had normal or corrected-to-normal vision. All participants signed informed consent before the experiment following the guidelines of the Ethics committee of the University of Barcelona. Participants were paid with course credit for their participation. The sample size for this and subsequent experiments was set to a minimum of forty-five participants considering a medium effect size (Cohen's d > 0.5) with a power of 95%.

Stimuli

Visual stimuli consisted of twelve artificially generated images of symmetric round fractals. Images were created in Apophysis 7x, a fractal flame renderization software that uses iterated function systems (Spotworks & Berthoud, 2008). The luminance saliency of the selected stimuli was normalized using histogram equalization applied to the value channel of their HSV (hue, saturation, value) representation. Every image had a diameter of 800 px. and was presented in the middle of a CRT monitor, on a black background.

Auditory stimuli were created using random parametrization of multiple synthesizers from the digital audio workstation FL Studio. Twelve different, unequivocally distinguishable abstract sounds were generated. Stimuli were then edited using the software Audacity: their amplitude was normalized, and their length reduced to 500 ms by cutting and time-stretching. To remove any clicking in the sounds an envelope of 100 ms of attack and 100 ms of release was applied. Auditory stimuli were presented through headphones to reduce any possible environmental noise.

Oddball versions of all stimuli were created to have a measure of implicit learning during the implicit learning phase. The oddball versions were designed so that they could be easily discriminated from the standard stimuli while still sharing enough characteristics with their corresponding standard stimulus. In the visual modality, a layer of upscaled RGB white Gaussian noise with the same shape as the stimuli was added to each stimulus. The noise was upscaled to get an additional pixelated effect (Fig. 1A). In the auditory modality, white Gaussian noise with the same amplitude as the stimuli was added to the sounds. Each stimulus appeared in its oddball version 10 times, out of a total of 110. Oddball appearances were randomly scattered across the implicit learning phase and they appeared half of the times on the first item of the pair (leading stimulus) or on the second item of the pair (trailing stimulus).

For each participant, visual and auditory stimuli (twenty-four in total) were pseudo-randomly grouped into twelve standard pairs: three pairs for each one of the four pair conditions.

Transitional probabilities (TPs) between the stimuli within a standard pair always was set at 100%, while the TPs between stimuli of different pairs was set at 8%. Each pair was presented 110 times during exposure leading to a total of 1320 pair presentations at the end of the implicit learning phase.

After being exposed to the stimuli, the participants underwent an explicit test to assess their learning. The test consisted of three standard pairs previously exposed in the learning phase for each condition (e.g., V1-V4, V2-V5, V3-V6), which were presented alongside six new pairs. The new pairs were created by combining the three leading stimuli of the standard pair modality with their trailing stimuli, but randomly associating them (e.g., V1-V5, V2-V6, V3-V4), and by taking the leading and trailing stimuli of the same modality but from a crossmodal trial (e.g., A1-V7, A2-V8, A3-V9 and V10-A7, V11-A8, V12-A9). These deviant pairs consisted of pairs that had not been presented to the participants before and were matched in frequency of occurrence for each stimulus, preventing participants from relying on differences in the probability of appearance during the explicit test phase to solve the task. We generated six two-interval forced choice (2IFC) trials for each of the four modality combinations, resulting in a total of 36 test trials. Each trial consisted in a standard pair and a new pair presented sequentially, and participants had to identify which of the two was the standard pair.

Procedure

The participants seated in a dimly light, sound attenuated room, facing a computer screen. The viewing distance from the participants to the monitor was approximately 50 cm. Before the experiment, participants were informed they would watch and listen to a series of images and sounds that could either be standard or noisy (oddball). To ensure that the participants understood the difference between standard stimuli and noisy stimuli, they were presented with an example of a stimulus per modality and their corresponding oddball before starting. Instructions were given to the participants both orally by the experimenter and in written form printed on the screen.

The experiment consisted of three phases (Fig. 1C): a *training phase*, in which participants got familiar with the stimuli and the task; an *implicit learning phase*, in which they were exposed to statistical regularities; and an explicit *test phase*, in which participants were tested regarding their knowledge acquired from such regularities. Each phase was separated by a short break of 1 minute.

During the training and implicit learning phases, the participants' task consisted of detecting oddball stimuli by pressing the spacebar key whenever they saw or heard one. They were

instructed to direct their gaze at a white fixation cross at the center of the monitor. Each visual or auditory stimulus was presented for a duration of 500 ms and they were separated by a 250 ms interval.

In the *training phase*, stimuli were presented to the participant in random order, without any paired structure. Oddballs appeared at a higher rate than during the implicit learning phase (35% of the stimuli), because the objective of this phase was to teach participants to discriminate between standard and oddball stimuli. The training phase lasted for one minute, but it was repeated until the participant was able to detect at least the 80% of the oddball stimuli correctly. This way we ensured that the participant would perform the oddball detection task correctly.

The *implicit learning phase* had the objective of training the participants to learn the TPs and measure the implicit learning of the pairs. This part of the experiment was divided into eight blocks, each of them lasting for approximately six and a half minutes. Between blocks, participants had the opportunity to rest and then resume the experiment by pressing a key whenever they wanted. In order to maintain them engaged with the task, after each block, participants received feedback about the percentage of oddball stimuli detected. To constrain the temporal window between oddballs and keypresses and calculate the proportion of correct oddball detection, we considered an oddball stimulus as detected whenever a keypress occurred no later than 2 seconds after the stimulus' onset. After the implicit learning phase, we revealed for the first time to the participants that they had to undertake an additional task. In the *explicit test phase*, we informed the participants that the visual and auditory stimuli previously seen during the implicit learning phase had been presented conforming pairs. To assess whether the participants had explicit knowledge about the pairs presented, we asked them to perform a 2IFC task. New instructions were given to the participants both printed on the screen and orally by the experimenter. Each trial consisted of two stimuli pairs: a standard one, corresponding to a pair presented during implicit learning phase; and a new one, that violated the statistical regularities from the implicit learning phase.

After being shown both pairs in sequential order, participants were asked to select, using the left and right arrow keys, which one was more familiar to them. Left corresponded to the first pair presented and right for the second pair. After choosing a pair, participants were asked to report the confidence of their decision. They had to select one of the following options by pressing either the 'Z', 'X' or 'C' key: "I remembered the sequence", "The sequence was more familiar to me" or "I guessed". These answers correspond to the confidence conditions Remember, Know and Guess, respectively. We considered the Remember and Know conditions as indicators of explicit knowledge being accessed either through recall or

familiarity, respectively. The test was self-paced: participants had to press a key before each pair appeared and had no time limit to respond.

Data analysis

To simplify the interpretability of the results, we took a two levels of analysis approach. Firstly, we conducted a repeated measures ANOVA on the proportion of detected oddballs and the corresponding reaction times to investigate the potential impact of the "modality" and "modality transition" variables on participant performance. Secondly, we aimed to assess implicit statistical learning by modelling the proportion of detected oddballs and reaction times as a function of trial number during the implicit test task, while also including stimulus position within the pair (leading vs. trailing) as an additional regressor in the model. Our hypothesis was that as participants became more familiar with the stimuli associations during the implicit learning phase, they would be able to use the leading stimulus to anticipate the trailing stimulus in each pair, leading to improved accuracy and reaction times when detecting trailing oddballs. To assess the impact of oddball position within each pair on the probability of detection, we implemented a generalized linear mixed-effect model (GLMM). Given the binomial nature of the responses, we modeled trial-level hits and misses using a logistic regression in lme4 package (Baayen et al., 2008; Bates et al., 2014) and R (version 3.6.2, R Foundation for Statistical Computing). GLMMs were optimized using the *bobyca* algorithm. On the other hand, as the reaction times conform a continuous dependent variable, we modeled them at the trial-level using a linear mixed model (LMM). We analyzed only those reaction times larger than 200 ms and shorter than 2 s relative to oddball onset. To reduce the positive skewness that characterize reaction times distributions, we normalized them by calculating the reciprocal of the latencies (Whelan, 2008). GLMMs and LMMs have more statistical power than classical ANOVA analyses given that variability in the proportion of hits and reaction times due to the individual differences between the participants or the stimuli can be accounted as random errors.

To assess implicit learning of the transition probabilities over the course of the experiment, we used the trial number of each stimulus presentation as a continuous regressors as a function of the modality transitions (i.e. unisensory transition if the oddball was preceded by a stimulus of the same modality or crossmodal if it was preceded by a stimulus of a different sensory modality). We rescaled trial-numbers to z-scores to improve the convergence of the mixed model. By default, we also included several random error effects in the model, allowing the intercept to vary independently for each participant and stimuli. To account for

individual variability in performance along the experiment, we also modelled the effect of trial number at the participant-level as another random effect.

To investigate potential differences in implicit learning between unimodal and crossmodal transition conditions, regardless of the modality of the oddball stimulus, we combined the auditory and visual unisensory conditions together, apart from the audiovisual and visuo-auditory conditions that were also analyzed together (see Table 1; Eq. 1). Our models included 'oddball position' (the difference in performance between detecting trailing versus leading oddball stimuli, with positive values indicating better performance in detecting trailing oddballs), as well as its interaction with modality transition. We also included and interaction between oddball position and trial number to measure changes in the 'learning rate'. To account for performance variability associated with the modality of the oddball stimulus (visual or auditory), we included an additional random-error term. We repeated the same analysis approach for the visual and auditory oddballs independently to obtain a modality-specific description of changes in performance (see Eq. 2).

We used post-hoc likelihood-ratio ($\chi 2$) model comparisons to assess the predictive power and significance of all main effects and interactions between the regressors included in the mixed models. We compared a reduced model (i.e., a model that only included trial number and modality transition as fixed effects) with other models of increasing complexity (i.e. a model that includes oddball position as regressors). After multiple model comparisons, we determined a model that better explained the data, and reported the obtained $\chi 2$ against the immediately less complex model. We then assessed the significance of each regressor estimate with respect to the others by calculating the Wald (z) statistic for each fixed effect estimate in the full model. Additionally, we used the regressor coefficients and associated standard errors from the most complex model that included all the main effects and interaction (Table 1; Eq. 1 and Eq. 2) to compare the effects across different conditions and calculate the confidence interval for each regressor estimate. It should be noted that the lme4 output estimates represent relative contrasts between regressors. To facilitate visual comparisons across conditions, the parameter estimates displayed in the figures were recalculated as absolute values. For instance, the crossmodal intercept was calculated by adding the unimodal intercept for leading oddballs to the estimated crossmodal intercept for leading oddballs. Similarly, the crossmodal oddball estimate was obtained by summing the unimodal leading oddball estimate and the crossmodal leading oddball estimate, and so on.

Finally, to test whether participants could explicitly recognize the associations between the leading and trailing stimuli, we modelled the explicit learning test responses using GLMM (family binomial logit). As this task consisted of a 2IFC, chance level is equal to a 50% of

correct responses (logistic regression intercept = 0). We modelled each "modality" and "modality combination" (AA, VV, AV and VA) separately. If the intercept of the model is significantly larger than 0, it means that participants can discriminate the standard pairs above chance level (>50%). Moreover, to determine whether the participants guessed or explicitly recognized the standard pairs, we also modelled their reported level of confidence in the decision (Table 1; Eq.3). We only report the results and statistics of the post hoc model comparisons in the main results section. Again, the reported $\chi 2$ values reflect the contrast of the best model against the immediately simpler contrasted model. Bayes-factor (BF) estimates derived from model comparisons showing a null effect are reported as an additional statistic, whose validation is essential to support or reject a relevant hypothesis in this study. The stimuli and code used to run the experiments, and reproduce the analyses and figures is available at https://osf.io/tcxe9/.

Table 1: Full Linear mixed models fitted to the data to quantify implicit statistical learning effects. Trial regressor modulation represents a change in performance as a function of the trial number. Oddball position (leading trailing) regressor modulation indicates that there is some learning, as performance changes as a function of the oddball position in the trial. The interaction between trial number and oddball position regressor represents a dynamic change in learning during the experiment (learning rate). We compared our base model including a trial regressor and its interaction with the modality pair regressor against other extended versions of the model including a main effect [+] or an interaction [*] with the oddball position regressor. Random effects of the GLMM models can be: Random $Error_A$ = (modality|participant) + (1|stimulus) + (1+ trial|participant); Random $Error_B$ = (1|stimulus) + (1+ trial|participant)

	Dependent variable	**Regressors**	**Modelling…**	**Equation ID**
Modalities together	P(Hits) or 1/RT	~ trial * modality pair [+ or *] oddball position (learning) + Random $Error_A$	…A and V modality oddballs together	Eq 1
Implicit statistical learning effects	P(Hits) or 1/RT	~ trial * modality pair [+ or *] oddball position (learning) + Random $Error_A$	…A and V modality oddballs separately	Eq 2
Explicit statistical learning effects	P(Correct)	~ confidence level + (1|stimulus)	…VV, AA, AV and VA separately	Eq 4

Results

Implicit learning: oddball detection task

Accuracy analyses: During the implicit learning phase, participants were highly accurate in detecting both the visual and auditory oddball stimuli (group average: $P(Hits)_V = 0.99$ and $P(Hits)_A = 0.85$). In order to explore how participants oddball detection depended on modality and modality-transition, we ran a repeated measures ANOVA (Fig. 2A). Our results showed that participants detected more oddballs in the visual compared to the auditory conditions ($F(1, 44) = 25.5$, $P < 0.01$), but their detection performance did not change as a function of the modality transition ($F(1, 44) = 0.16$, $P = 0.6$).

We used GLMM and post hoc likelihood-ratio ($\chi 2$) model comparisons to explore participants' ability to detect oddball stimuli (hits) during the implicit learning phase and to determine whether they implicitly learned the stimuli pairs (Fig. 2B). In our first analysis we combined the visual and auditory modalities. Following a comparison of several models with varying levels of complexity, we identified a model that included trial number and modality transition as regressors to best explain hit performance ($\chi 2 = 11.82$, $df = 1$, $P < 0.001$). Our analyses (Fig. 2C) demonstrated that in consistency with a practice effect, participants' performance improved with time, particularly in the unimodal conditions ($z = 4.0$, $P < 0.001$). However, we did not observe any evidence that supported the notion that participants implicitly learned the unimodal or crossmodal pairs ($\chi 2 = 0.88$, $df = 1$, $P = 34$, $BF = 0.01$). We conducted the same analysis for each modality condition separately and replicated the performance improvement with number of trials, primarily in the auditory unimodal and crossmodal oddballs (Fig. 2C), but again we did not identify any statistical evidence of learning effects in either modality (i.e. visual $BF = 0.66$ or auditory $BF < 0.01$). We only found that detection of trailing stimuli was significantly larger in unimodal compared to the crossmodal modality pairs for the visual modality ($z = 1.4$, $P = 0.03$), however the regressor estimate did not significantly differed from 0, limiting our capacity to conclude that participants detected more trailing than leading oddballs in the unimodal visual condition.

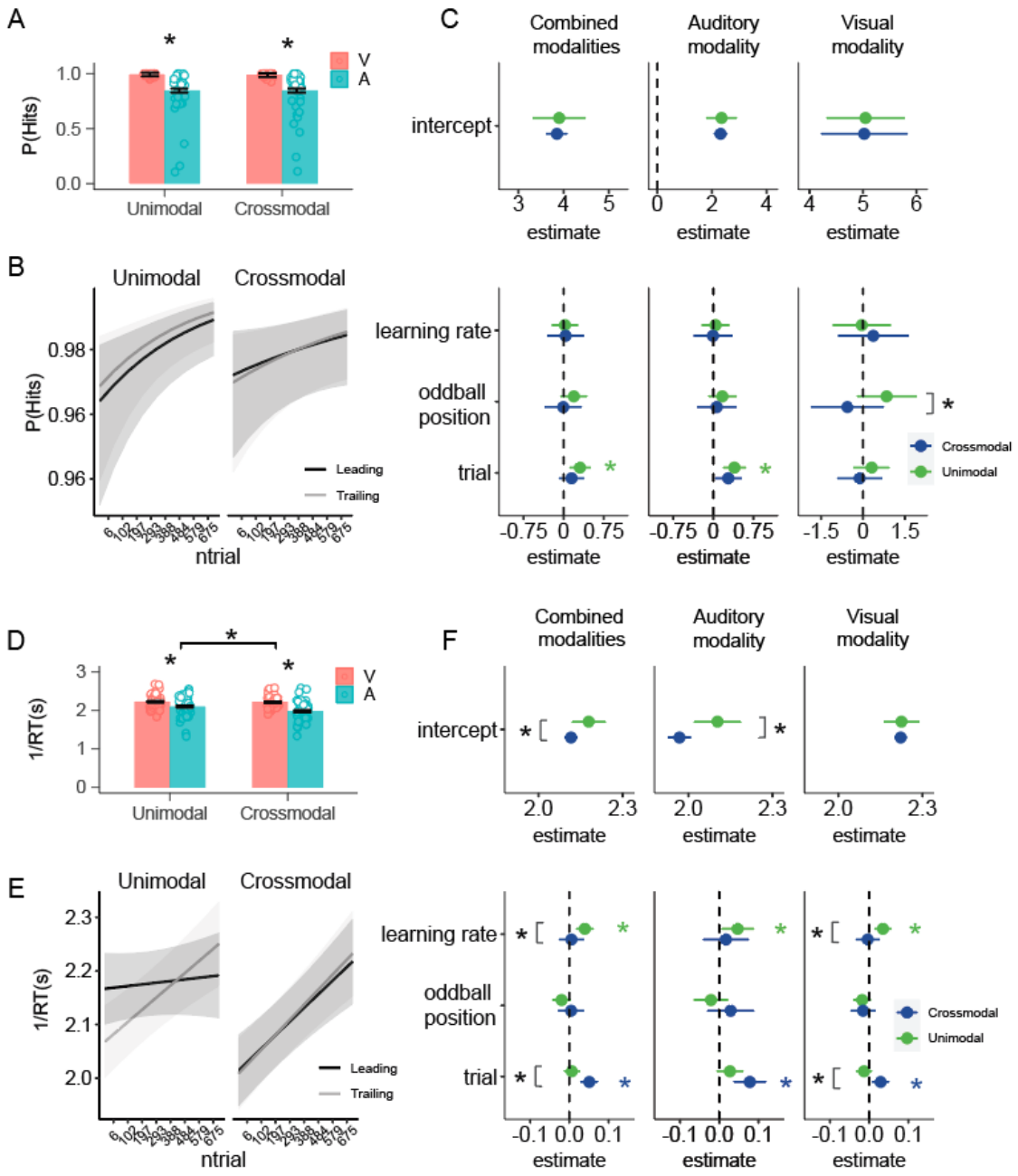

Figure 2. Performance during the implicit learning phase in experiment 1: A. Group-average proportion of hits for the unimodal and crossmodal pairs, separated by stimulus modality (V, A) and modality transition. B. Full model predictions for the unimodal and crossmodal pairs, modeling the visual and auditory modalities together. C. Parameter estimates of the full model applied to the combined auditory and visual modalities. The intercept of the model represents the average number of hits per modality (in log odds scale; 0 is equal to 0.5 proportion of hits). The trial regressor estimate represents the modulation in hits as a function of the number of trials. The learning estimate represents the average difference in hits between trailing and leading oddballs. Finally, the learning rate estimate represents linear modulations in the difference in detection between leading and trailing oddballs as a function of the number of trials. D, E, and F represent the same information as in the previous panels for reaction times. Confidence intervals (error bars) are calculated as the parameter estimate standard error multiplied by 1.96. Generally, an estimate whose confidence interval overlaps with a dashed vertical line indicates that the regressor modulation is not significantly different from 0. Based on model parameter contrasts, black asterisks indicate significant differences between

conditions estimates, while colored asterisks indicate significant modulations of a given parameter with respect to 0. The characters “V” and “A” represent visual and auditory oddballs, respectively.

Reaction time analyses: We ran a repeated-measures ANOVA to analyze the reaction times of participants in detecting visual and auditory oddballs (Fig. 2D). The results showed that participants were faster at detecting visual oddballs than auditory ones ($F(1,44) = 29.3$, $P < 0.01$). Furthermore, slower reaction times were observed for crossmodal compared to unimodal transitions ($F(1,44) = 15.4$, $P < 0.01$). This effect was more pronounced for auditory oddballs in the crossmodal transitions, as indicated by a significant interaction between the modality and modality transition factors ($F(1,44) = 15.2$, $P < 0.01$). These findings suggest that detecting an oddball that is preceded by a stimulus delivered in the complementary sensory modality results in a detection cost.

We used linear mixed-effects models (LMMs) and conducted post hoc likelihood-ratio ($\chi2$) model comparisons to investigate the impact of trial number and oddball position (as illustrated in Fig. 2E) on participants' response times for detecting oddballs. Our analysis showed that the full model, which included an interaction between trial number, modality transition, and oddball position, provided a more robust explanation for the data ($\chi2 = 14$, $df = 3$, $P < 0.01$). Specifically, we observed that participants' reaction times improved as the number of trials increased, particularly in the crossmodal conditions ($z = 3.8$, $P < 0.01$). Furthermore, participants exhibited a learning effect for the unimodal conditions, as they became faster in detecting unimodal trailing oddballs compared to leading ones over time ($z = 3.46$, $P < 0.01$). However, we did not observe a similar learning effect for crossmodal oddballs ($z = 0.29$, $P < 0.76$; $\chi2 = 0.035$, $df = 2$, $P = 0.98$, $BF < 0.01$). We conducted the same analysis separately for each modality (Fig. 2F) and found that the full model also provided a good fit for both the auditory and visual modalities ($\chi2 = 14$, $df = 3$, $P < 0.01$ for both modalities). Again, our results revealed significant positive learning rates for both the auditory ($z = 2.26$, $P = 0.02$) and visual modality ($z = 3.2$, $P < 0.01$) conditions, suggesting the participants became progressively faster in detecting trailing compared to leading oddballs.

Taken together, our findings suggest that participants gradually acquired and utilized the stimulus associations to improve oddball discrimination only for the unimodal pairs.

Explicit learning of auditory transitional probabilities

The GLMM analyses on the proportion of correct responses in the explicit learning test phase revealed that participants were able to recognize the AA standard pairs above chance level ($z = 3.61$, $P < 0.001$). However, no evidence of standard pair recognition was found in the VV ($z = 0.7$, $P = 0.46$, BF = 0.064), AV ($z = -0.64$, $P = 0.51$, BF = 0.061), or VA ($z = -0.049$, $P = 0.96$, BF = 0.05) conditions (Fig. 3A). Additionally, the confidence level regressor only improved the model's fit in the AA condition ($\chi 2 = 7.4$, $df = 2$, $P < 0.05$). Specifically, participants reported a higher proportion of correct responses in the "know" ($z = 2.65$, $P < 0.005$) and "remember" ($z = 2.18$, $P < 0.05$) conditions compared to the "guess" level of confidence (Fig. 3B). These results suggest that participants were only able to explicitly recognize the unimodal auditory standard pairs.

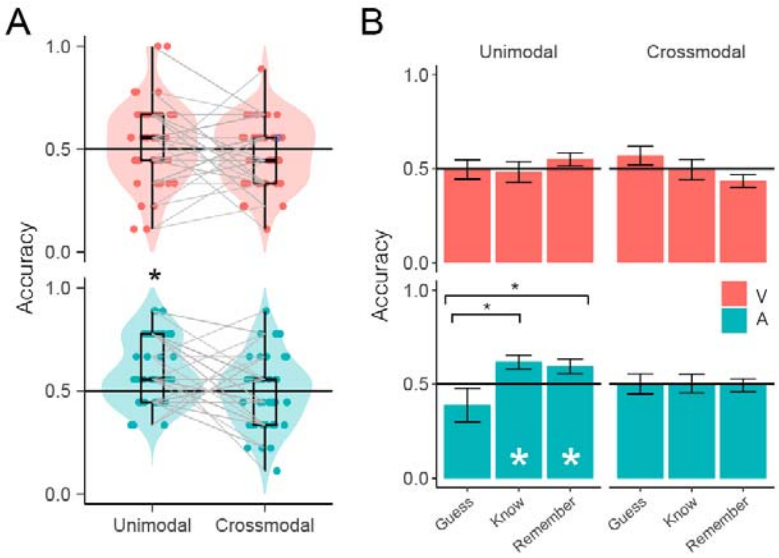

Figure 3. Explicit test performance in experiment 1: A. The accuracy in discriminating standard from deviant pairs is significantly above chance level for the auditory unimodal pairs. B. The analysis of participants accuracy as a function of the confidence in their responses demonstrates that participants could explicitly recognize the associations between the auditory stimuli (right). "V" and "A" characters stand for visual and auditory oddballs. Black and white asterisks indicate significant differences between conditions and significant difference against chance level respectively.

In summary, this experiment findings are partially consistent with a previous version of the same experiment (see experiment S1 in supplementary materials), where due to a technical error reaction times were not acquired and the analysis was limited to only hits during the implicit phase and accuracy during the explicit phase. The previous experiment S1 showed that participants implicitly learned the auditory unimodal and crossmodal oddballs, while they also recognized the unimodal visual and auditory pairs in the explicit test (Fig. S2). Based on experiment 1 and experiment S1, there is convergent evidence of unimodal implicit and explicit statistical learning. However, the results related to crossmodal statistical learning are

quite aligned with a modality specific limitation that does not allow the extraction of regularities at the sensory level between modalities, although with some caveats since experiment 1S did show implicit crossmodal learning, at least for visual predicting auditory stimuli.

Experiment 2

In our first experiment, we observed that participants were able to learn and use unimodal transitional probabilities (TPs) effectively, while crossmodal TPs were not learned as reliably. These findings support the notion that statistical learning (SL) is a modality-specific process that operates within each sensory modality (Frost et al., 2015). However, experiment S1 did show implicit crossmodal learning for auditory preceded by visual stimuli, indicating that crossmodal learning at the sensory level can occur. Morevover, it is important to note that in experiments 1 and S1, stimuli were presented in a continuous stream, where unimodal and crossmodal transitions were randomly intermixed. This random presentation may have hindered the learning of crossmodal pairs due to the perceptual costs associated with switching between modalities, as previously demonstrated in multisensory research (Gondan et al., 2004; Otto & Mamassian, 2010; Spence et al., 2001). These switching costs reveal an implicit bias to attend to the modality that delivered the last sensory input, making it challenging to disengage attention from one modality to the other. Consequently, when stimuli from different modalities are presented in random order, this bias may lead to grouping stimuli belonging to the same modality and impede the ability to chunk inputs from different modalities together, constraining the learning of crossmodal statistical regularities. In fact, the lower reaction times that we observed in the crossmodal oddballs compared to the unimodal oddballs are consistent with the existence of a modality switching cost. To reduce the impact of this attentional bias on SL, we designed our second experiment to include eight different experimental blocks, each containing a specific combination of visual-visual (VV), auditory-auditory (AA), audio-visual (AV), and visual-audio (VA) stimuli (Fig. 1C). By doing so, participants could anticipate the modality of the incoming stimuli in the crossmodal blocks (e.g., AVAVA), thus eliminating the possibility of a bias to group consecutive visual or auditory stimuli together.

Methods

Participants

Sixty-three university students participated in experiment 2 (59 females, 4 males, mean age 20.36, range 18-28). Participant's recruitment criteria and the stimuli were identical to the previous experiment.

Procedure

The experimental procedure was almost identical to experiment 1. However, here there was only one modality combination in each experimental block, being either VV, AV, AA, or VA (Fig. 1C). The order of each modality combinations block was randomized for each participant, and two blocks of the same modality could not come one after the other. We kept the TPs between the stimuli within a standard pair at 100%. However, because we presented a smaller number of possible pairs in each block, lowering the number of pair combinations per block to only 3 (6 stimuli), we increased the TPs between non paired stimuli to 33% ($\times$4 higher than in Experiment 1) and potentially rendering the SL of the associated pairs more challenging compared to the previous experiment.

Results

Implicit learning: oddball detection analysis

Accuracy: Participants detected most of the visual and auditory oddball stimuli ($P(Hits)_V$= 0.99 and $P(Hits)_A$= 0.88) during the implicit learning phase (Fig. 4A). A repeated measures ANOVA showed that participants detected more oddballs in the visual compared to the auditory conditions ($F(1, 63) = 52.2$, $P < 0.01$), but their detection performance did not change as a function of modality transition ($F(1, 63) = 0.063$, $P = 0.8$).

To investigate whether participants learned the item associations, we used a GLMM to model hits performance as a function of trial number on the visual and auditory modalities together (Fig. 4B). The best-fitting model included trial number, modality transition, and oddball position as fixed effects ($\chi2 = 12.15$, $df = 3$, $P < 0.01$). Importantly, we found that on average, participants detected more trailing than leading oddballs in the crossmodal conditions ($z = 3.56$, $P < 0.01$), providing evidence of rapid and sustained crossmodal statistical learning ($\chi2 = 6.83$, $df = 2$, $P = 0.032$). In addition, whereas the detection slope increased with the number of trials for the leading crossmodal oddballs, the detection performance for trailing crossmodal oddballs did not change as function of the number of trials. This finding suggests that learning of crossmodal statistical regularities reaches its maximum quickly, and does not continue to improve with additional exposure, probably due to a ceiling effect. On the other hand, we find no evidence to support that participants were

able to learn the transition probabilities in the unimodal conditions ($\chi 2$ = 0.85, df = 1, P = 0.35, BF = 0.01).

After repeating the GLMM analysis for the auditory and visual modalities separately (Fig. 4C), we only replicated the same pattern of results in the auditory modality ($\chi 2$ = 24.75, df = 4, P < 0.01; That is, on average participants detected more trailing than leading auditory oddballs in the crossmodal compared to the unimodal auditory condition (z = 3.93, P < 0.01). However, we did not obtain statistical evidence that participants learned the unimodal or crossmodal pairs containing a visual oddball, as the best fitting model did not include oddball position as a regressor ($\chi 2$ = 0.004, df = 1, P = 0.94, BF = 0.001).

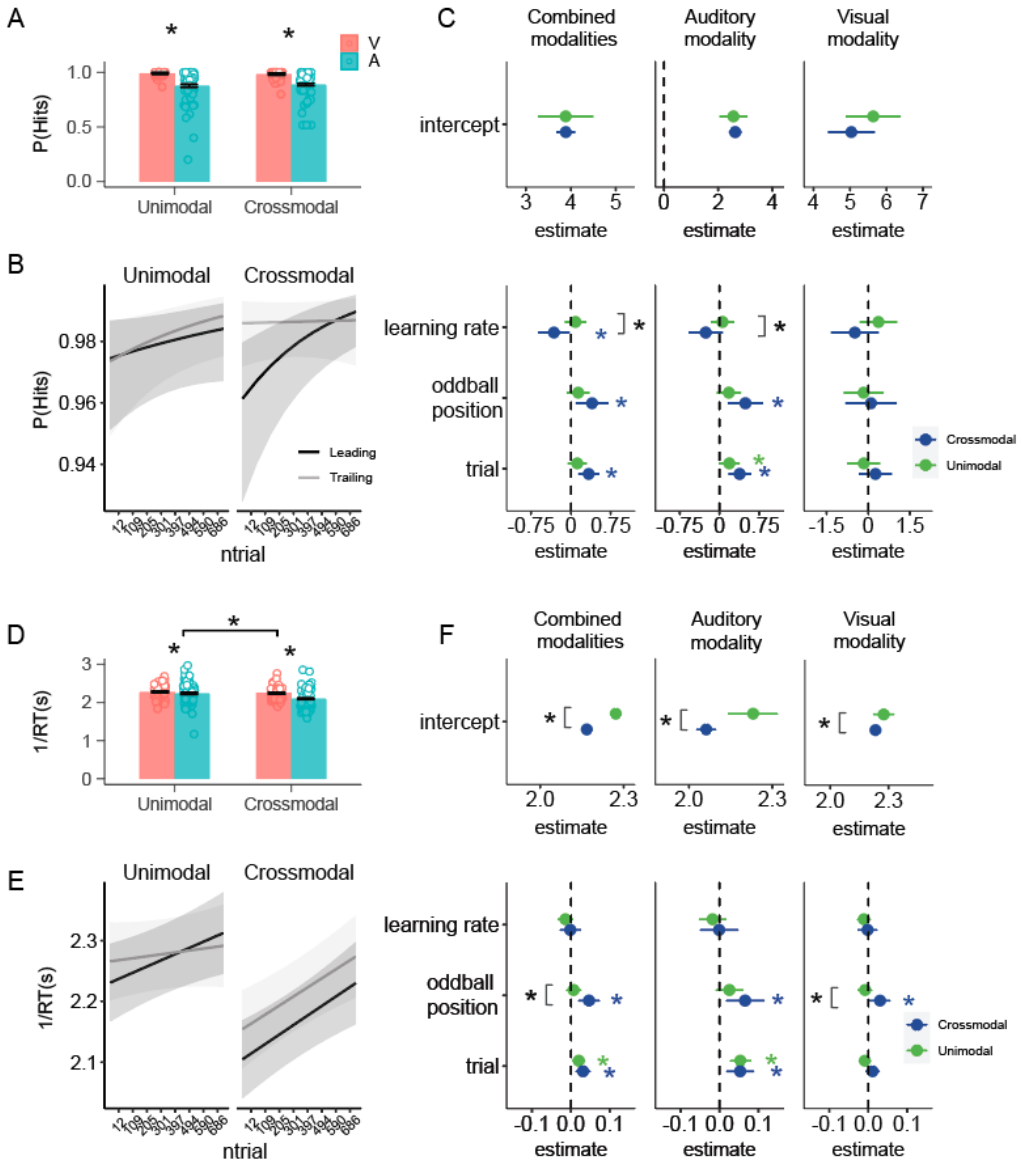

Figure 4. Performance during the implicit learning phase in experiment 2: A. Group-average proportion of hits for the unimodal and crossmodal pairs, separated by stimulus modality (V, A) and modality transition. B. Full model predictions for the unimodal and crossmodal pairs, modeling the visual and auditory modalities together and separating leading and trailing oddballs. C. Parameter estimates of the full model applied to the combined auditory and visual modalities. D, E, and F represent the same information as in the previous panels for reaction times. Confidence intervals (error bars) are calculated as the parameter estimate standard error multiplied by 1.96. Generally, an estimate whose confidence interval overlaps with a dashed vertical line indicates that the regressor modulation is not significantly different from 0. Based on model parameter contrasts, black asterisks indicate significant differences between conditions estimates, while colored asterisks indicate significant modulations of a given parameter with respect to 0. The characters "V" and "A" represent visual and auditory oddballs, respectively.

Reaction time analysis: A repeated measures ANOVA was conducted to analyze participants' reaction times in detecting visual and auditory oddballs. The results showed that participants were faster at detecting visual oddballs than auditory ones ($F(1,63) = 11.3$, $P < 0.01$). Reaction times for crossmodal pairs were slower compared to unimodal pairs ($F(1,63) = 36.8$, $P < 0.01$), which replicates the modality switching cost that was previously reported in experiment 1. Additionally, we found an interaction between the modality and modality pair factors ($F(1,63) = 15.4$, $P < 0.01$), indicating that the modality switching cost was stronger for auditory conditions compared to visual conditions (see Fig. 4D).

A LLM was applied to the reaction times data (Fig. 4E), including an interaction between trial number, modality pair, and oddball position regressors ($\chi2 = 11.82$, $df = 1$, $P < 0.001$). The results indicated that participants' reaction times improved with the number of trials ($z = 2.48$, $P < 0.01$) for both the unimodal and crossmodal pairs. This gradual improvement can be attributed to a practice effect. Importantly, the results also showed that participants were faster on average at detecting trailing crossmodal stimuli than leading ones ($z = 4.5$, $P < 0.01$), confirming implicit learning of the crossmodal associations. No such pattern was observed for the unimodal modality pair conditions auditory ($\chi2 = 1.4$, $df = 1$, $P = 0.2$, $BF = 0.05$). Moreover, the response improvement due to learning of the pair associations was significantly greater in the crossmodal compared to the unimodal conditions ($z = 2.8$, $P < 0.01$). We replicated the same pattern when analyzing the auditory ($\chi2 = 11.82$, $df = 2$, $P < 0.001$) and visual ($\chi2 = 9.63$, $df = 3$, $P < 0.021$) modalities separately. That is, participants were faster on average in detecting auditory ($z = 3.5$, $P < 0.01$) and visual ($z = 1.68$, $P = 0.09$) crossmodal trailing than leading oddballs (Fig. 4F).

No evidence of explicit statistical learning

GLMM analyses on the proportion of correct responses showed that participants could not explicitly recognize the unimodal nor the crossmodal pairs. Specifically, we did not find

evidence of standard pairs recognition in the VV (z = 0.65, P = 0.51, BF = 0.051), AA (z = 0.81, P = 0.7, BF = 0.057), AV (z = 0.24, P = 0.8, BF = 0.043) or VA (z = -0.71, P = 0.47, BF = 0.053) conditions (Fig. 5A). Consistently, adding the “confidence level” regressor (Fig. 5B) did not improve any of the models (all BF < 0.005).

Overall, the results of this second experiment show that participants are able to learn crossmodal transitional probabilities and use them to improve their detection of trailing oddball stimuli. Interestingly, we did not observe similar effects in the unimodal conditions, which may be due to the lower transitional probability per block in this experiment compared to the first experiment. These findings suggest that crossmodal statistical learning is superior to unimodal statistical learning when modality transitions are predictable. Moreover, our results indicate that crossmodal SL is only apparent when using an implicit index of learning, whereas SL can be both implicitly and explicitly registered in the unimodal conditions.

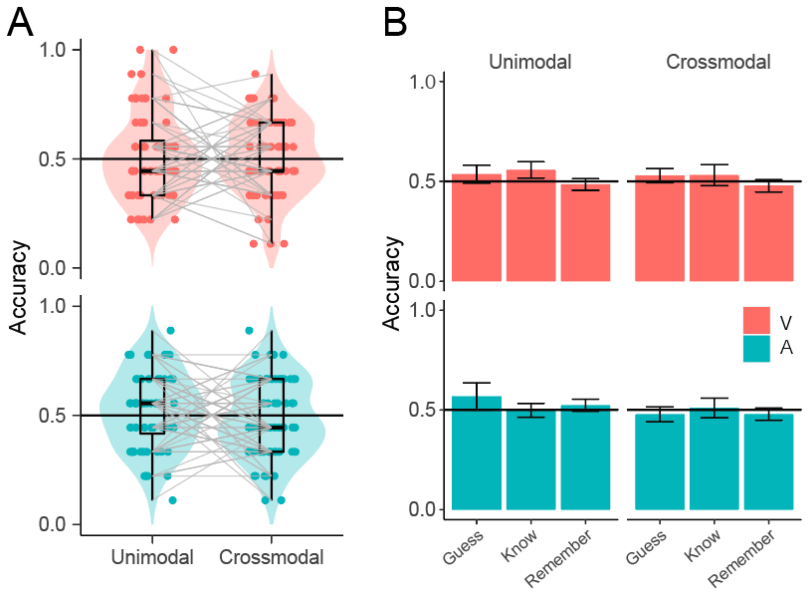

Figure 5. Explicit test performance in experiment 2: A. The accuracy in discriminating standard from deviant pairs was not significantly above chance level for any modality nor pair comparison. B. The analysis of participants accuracy as a function of the confidence in their responses demonstrates that participants could not explicitly 20ecognize any stimuli associations. “V” and “A” characters stand for visual and auditory oddballs.

Experiment 3

In our previous statistical learning experiments, participants were presented with meaningless fractal images and synthetic sounds to avoid encoding of information in higher-

level representations that are detached from their sensory components. Notably, in experiment S1 (see supplementary materials) and experiment 2, we observed that utilizing an online measure of participants' performance disclosed implicit learning of statistical regularities between meaningless crossmodal stimuli. Furthermore, we noted that such learning might be impeded by a predisposition to group information from the same sensory modality. In experiment 3, our objective was to test the hypothesis that crossmodal regularities can be learned explicitly if the information is represented at a supramodal level, such as semantically. In agreement with (Frost et al., 2015) hypothesis, we predict that if statistical learning can occur between high-level stimuli representations, participants should be able to learn crossmodal contingencies, despite biases at the perceptual level that group stimuli from the same modality. To test this prediction in the new experiment, we used visual and auditory stimuli with semantic content to facilitate the formation of sensory-independent stimulus representations.

Methods

Fifty university students participated in experiment 3 (7 males, mean age 21.9, range 18-45). Participant's recruitment criteria and the stimuli were identical to the previous experiments.

Procedure

Similar to the first experiment, we randomly intermixed visual, auditory, audiovisual, and visuo-auditory pairs in each experimental block while maintaining the TPs between stimuli within a standard pair at 100% and the transitional probabilities between non-standard pairs at 8%. However, unlike the previous experiments, we used meaningful stimuli to represent visual and auditory objects. The visual images were composed of 6 animate objects (e.g., crocodile) and 6 inanimate objects (e.g., piano) overlaid on a black background. Similarly, auditory stimuli consisted of 500 ms clips corresponding to 6 animate (e.g., dog barking) and 6 inanimate objects (e.g., phone ringing). All stimuli were downloaded from the internet and pre-processed in the same manner as the fractals and synthetic stimuli used in the previous experiments to generate the oddballs. Prior to beginning of the experiment, we ensured that participants could identify the stimuli by sequentially presenting them with the images and sounds and asking them whether they could recognize their semantic referents. We found that all the stimuli referents could be correctly identified by the participants.

Results

Implicit learning: oddball detection analysis

Accuracy: During the implicit learning phase, participants accurately detected both the visual and auditory oddball stimuli, with group averages of $P(Hits)_V = 0.97$ and $P(Hits)_A = 0.84$ (Fig. 6A). To explore how participants' oddball detection changed as a function of modality and modality transition, we conducted a repeated measures ANOVA. This analysis revealed that participants detected more visual than auditory oddballs ($F(1,48) = 75.6$, $P < 0.01$) and more crossmodal than unimodal oddballs ($F(1,48) = 75.6$, $P < 0.01$). We also observed a significant interaction between modality and modality transition ($F(1,48) = 15.5$, $P < 0.01$), indicating that the benefit of detecting crossmodal over unimodal oddballs was greater in the auditory modality. We had not observed an effect in this direction in the previous experiments, and we do not have a definitive explanation for it. We speculate that the improvement in detecting crossmodal auditory oddballs compared to unimodal ones might be explained by a more effortful and therefore prolonged processing of auditory semantic stimuli compared to visual stimuli, which could induce a "blink" effect in processing subsequent information in the same sensory modality.

Based on the GLMM analysis of detection performance (Fig. 6B), we found no empirical evidence supporting the notion that participants implicitly learned the item associations in either the unimodal or crossmodal modality pairs ($\chi2 = 0.41$, $df = 1$, $P = 0.5$; $BF = 0.01$). Indeed, the model that better explained the combined auditory and visual data only included trial number and modality transition as regressors (without an interaction between them; $\chi2 = 10.3$, $df = 1$, $P < 0.01$), indicating that participants' performance improved gradually with the number of trials ($z = 3.24$, $P < 0.1$). We re-ran the GLMM on each sensory modality independently (Fig. 6C) and reproduced a null effect of oddball position in both the auditory ($\chi2 = 0.007$, $df = 1$, $P = 0.93$, $BF = 0.13$) and visual modalities ($\chi2 = 0.58$, $df = 1$, $P = 0.58$, $BF = 0.17$). We only found that participants' performance in the auditory condition improved with the number of trials for both leading and trailing oddballs due to practice effects. However, their performance in the visual conditions decreased with the number of trials, possibly due to fatigue.

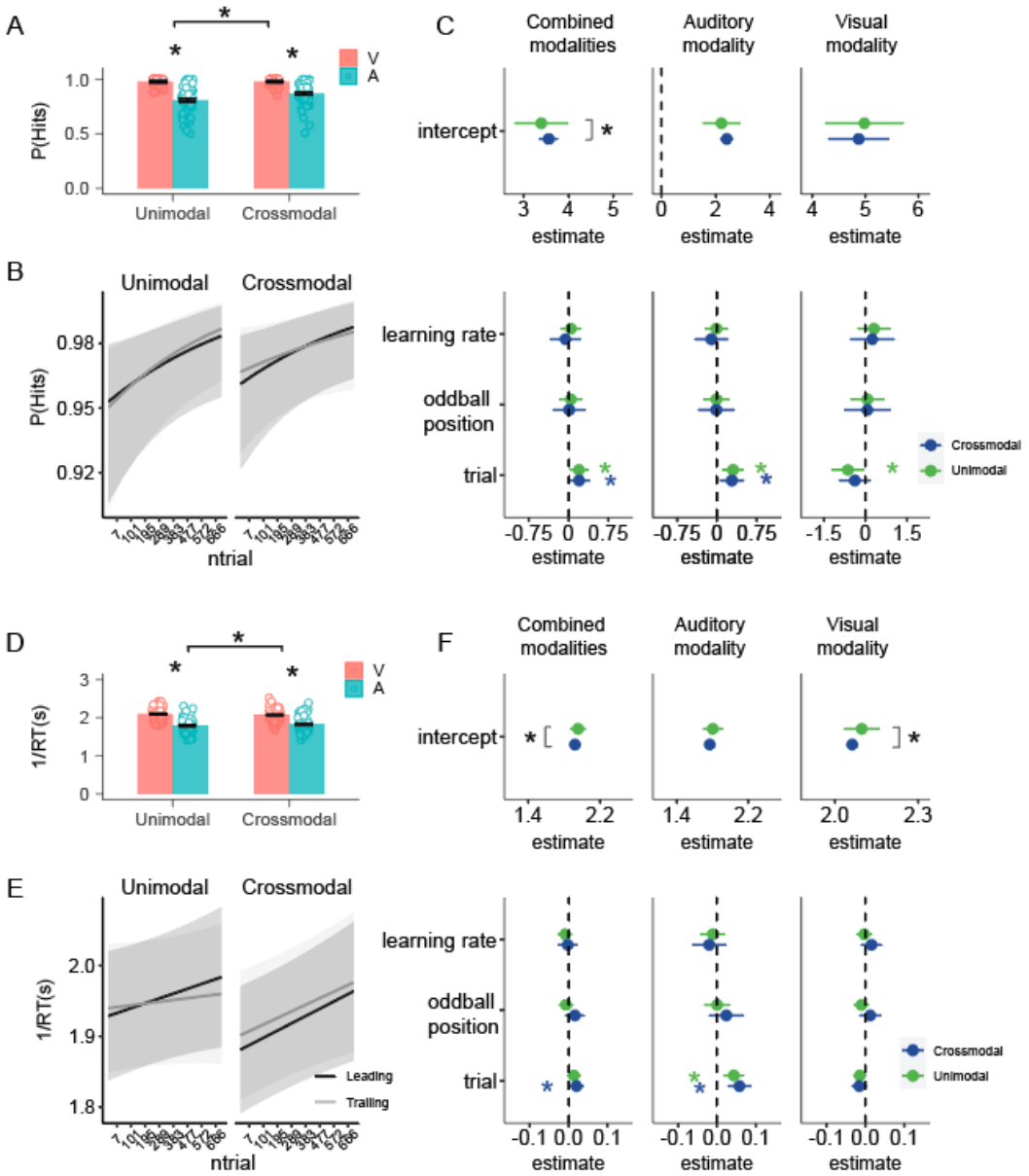

Figure 6 Performance during the implicit learning phase in experiment 3: A. Group-average proportion of hits for the unimodal and crossmodal pairs, separated by stimulus modality (V, A) and modality transition. B. Full model predictions for the unimodal and crossmodal pairs, modeling the visual and auditory modalities together and separating leading and trailing oddballs. C. Parameter estimates of the full model applied to the combined auditory and visual modalities. D, E, and F represent the same information as in the previous panels for reaction times. Confidence intervals (error bars) are calculated as the parameter estimate standard error multiplied by 1.96. Generally, an estimate whose confidence interval overlaps with a dashed vertical line indicates that the regressor modulation is not significantly different from 0. Based on model parameter contrasts, black asterisks indicate significant differences between conditions estimates, while colored asterisks indicate significant modulations of a given parameter. The characters "V" and "A" represent visual and auditory oddballs, respectively.

Reaction time analysis: A repeated measures ANOVA on the reaction times (Fig. 6D) revealed that participants were faster at detecting visual oddballs than auditory oddballs ($F(1,48) = 15.5$, $P < 0.01$). There was no significant effect of modality transition ($F(1,48) = 0.16$, $P = 0.68$). However, we did find an interaction between modality and modality transition, indicating a specific improvement in reaction times for unimodal visual oddballs relative to crossmodal oddballs ($F(1,48) = 5.3$, $P < 0.01$). This interaction is consistent with a modality switching cost that is primarily observed in transitions from auditory to visual oddball stimuli.

We conducted a linear mixed-effects (LMM) analysis to determine whether participants learned the pair associations implicitly. We modelled the reaction times for both visual and auditory modalities together (Fig. 6E). Our analysis revealed that the best model for explaining participants' responses only included trial and modality pair as fixed-effect regressors ($\chi2 = 5.8$, $df = 1$, $P < 0.05$). However, we did not find evidence of implicit statistical learning as the oddball position regressor did not improve the model fit ($\chi2 = 3.03$, $df = 1$, $P = 0.08$, $BF < 0.01$).

Further investigation of the regression estimates (Fig. 6F) showed that participants' performance improved with increased practice in the crossmodal conditions ($z = 2.14$, $P = 0.02$), indicating a possible practice effect. Separate LMM analyses for the auditory and visual modalities also did not reveal improved model fits with the inclusion of oddball position ($\chi2 = 1.2$, $df = 1$, $P = 0.26$, $BF < 0.01$; $\chi2 = 0.004$, $df = 1$, $P = 0.98$, $BF < 0.01$, respectively). These findings did not provide empirical evidence that participants implicitly learned the transitional probabilities in either modality.

Explicit learning of auditory, visual and audiovisual transitional probabilities

GLMM analyses on the proportion of correct responses in the explicit learning test phase showed that participants recognized the unimodal AA and VV standard pairs above chance level ($\chi2 = 16.3$, $df = 1$, $P < 0.001$; $\chi2 = 9.53$, $df = 1$, $P < 0.005$, Fig. 7A). Interestingly we also found that participants explicitly recognized the AV standard pairs ($\chi2 = 6.08$, $df = 1$, $P < 0.02$). However, we did not find evidence of standard pairs recognition in the VA ($\chi2 = 1.2$, $df = 1$, $P = 0.27$; $BF = 0.09$) condition. The confidence level regressor significantly improved the model's fit in the VV ($\chi2 = 7.4$, $df = 2$, $P < 0.03$) and AV conditions ($\chi2 = 7.9$, $df = 2$, $P < 0.03$). Specifically, the participants reported a larger proportion of correct responses in the "know" than in the "remember" (VV, $z = 2.42$, $P < 0.02$; AV, $z = 2.76$, $P < 0.01$) and "guess" (VV, $z = 2.5$, $P < 0.05$; AV, $z = 2.48$, $P < 0.02$) confidence levels (Fig. 7B). These results

together indicate that participants can explicitly recognize unimodal and crossmodal standard pairs when these contain semantic information. Moreover, they could explicitly recognize the stimuli associations, as participants were more accurate in those decisions rated with a confidence level “know”. The same pattern did not apply to the “remember” confidence level.

The results of experiment 3 provide additional evidence for crossmodal statistical learning. However, unlike previous experiments where SL was revealed through the detection of oddballs in the implicit learning phase, we only found evidence of unimodal and crossmodal SL during the explicit test phase. This difference in the pattern of results may be associated with how participants process meaningless and meaningful stimuli. We speculate that the processing of meaningless stimuli may favor the detection of sensory changes, which could facilitate SL benefits in detecting oddball sensory noise. In contrast, processing of meaningful stimuli may emphasize the neural encoding of high-level referents of the presented inputs, reducing the SL impact on sensory feature processing and favoring deeper and more elaborate processing. This, in turn, may facilitate the explicit recognition of the associations between unimodal and crossmodal pairs.

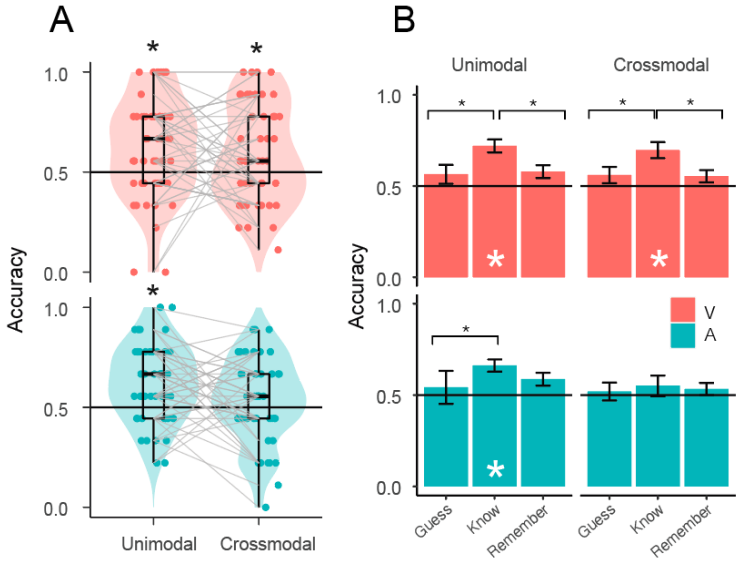

Figure 7. Explicit test performance in experiment 3: A. The accuracy in discriminating standard from deviant pairs is significantly above chance level for the auditory and visual unimodal pairs, and for the crossmodal audiovisual pairs. B. The analysis of participants accuracy as a function of the confidence in their responses demonstrates that participants could explicitly recognise the associations between the auditory, visual and audiovisual stimuli pairs. “V” and “A” characters stand for visual and auditory oddballs. Black and white asterisks indicate significant differences between conditions and significant difference against chance level respectively.

Discussion

Previous studies have failed in showing multimodal interactions in SL (Conway & Christiansen, 2006; Seitz et al., 2007; Walk & Conway, 2016). This has led to propose that, although SL relies on domain-general computational principles, the learning of statistical dependencies between multimodal (low level) stimulus features (e.g. tones, shapes, vibrations) is constrained by the representational characteristics of its correspondent sensory modality (Frost et al., 2015).

Here, in three separate experiments we aimed to investigate alternative explanations that could have impeded the ability of previous studies to detect crossmodal statistical learning. To achieve this, we utilized an implicit measure of SL by examining participants' performance in detecting leading and trailing oddball stimuli during the exposure phase. Using generalized linear mixed models (GLMMs), we assessed how SL evolved during the exposure phase and included the trial number of the oddball as an additional predictor in the model.

We observed implicit crossmodal statistical learning (SL) between low-level non-semantic information, as shown in experiment S1 (see supplementary materials) and experiment 2. This effect was more pronounced in experiment 2, when stimuli were paired in modality-predictable transitions. This result suggests that there is a perceptual bias to group stimuli from the same modality together. In consistence with this idea, in experiment 1 in which participants were unable to predict the next modality we could not find significant crossmodal SL. Conversely, in experiment 2, when modality alternations were predictable and participants could anticipate the next modality input, SL of crossmodal pairs was superior to the SL of unimodal transitional probabilities. This enhanced performance of crossmodal SL when perceptual grouping biases are controlled may be interpreted from an attentional framework. Previous research has shown that attention plays a crucial role in unimodal (Treisman & Gelade, 1980) and multimodal (Talsma et al., 2010) information binding, and has been demonstrated to be determinant for SL (López-Barroso et al., 2016; Richter & de Lange, 2019; Toro et al., 2005). Therefore, although our results suggest that orienting attention across sensory modalities is costly (i.e. modality-switching cost), previous research showed that attentional resources are independent for each sensory modality when processing low-level information (Alais et al., 2006). We speculate that this increased availability of attentional resources in predictable crossmodal transitions compared to unimodal transitions enhances the SL of crossmodal statistical associations. Our findings also suggests that crossmodal associations are less explicit, as they are primarily reflected by an implicit measure of SL.

This reinforces the idea that evaluating whether participants have learned a statistical association should not solely rely on an explicit recognition task.

An additional interesting result in our experiments was that on average, explicit statistical learning was found to be superior in the auditory modality as most of the experiments showed explicit recognition of the auditory pairs, but not the visual pairs. However, these results do not indicate that participants were unable to learn the visual pairs. In fact, experiment 1 showed implicit learning of visual transitional probabilities, while experiment S1 and 3 showed explicit learning of visual transitional probabilities. A weaker statistical learning effect was observed for visual stimuli sequences compared to auditory stimuli sequences, which is consistent with previous literature demonstrating a statistical learning superiority for auditory over visual sequential contingencies (Conway & Christiansen, 2005; Saffran, 2002).

Finally, we addressed whether there might be different mechanisms involved in SL between crossmodal low-level and high-level semantic information. Experiment 3 suggests that when information can be encoded in amodal representations (i.e., semantic stimuli), the learning of crossmodal contingencies is also explicit, despite perceptual grouping biases. Previous research has shown that SL is enhanced when associated items represent semantic compared to non-semantic referents (Cunillera, Laine, et al., 2010). A higher-level processing of the stimuli might facilitate the explicit recognition of the associated pairs in both unimodal and crossmodal conditions. Our behavioral results do not allow us to draw definitive conclusions about the neural mechanisms involved in crossmodal SL. However, they do support the idea that SL can occur at multiple levels of representation, encompassing both low-level and high-level information. While learning of low-level associations is constrained by perceptual phenomena (i.e., perceptual grouping bias), learning of semantic associations bypasses these limitations potentially occurring at a supra-modal level.

The results of this study compels revision of the current neurobiological model of SL in which learning of statistical regularities between low-level stimuli features relies on hard-wired learning computations that take place in their respective sensory cortex. Here, we show that visual and auditory synthetic stimuli that are difficult to be represented in a supramodal format (e.g. semantic representations), can be bound together although with perceptual limitations that can be removed through attention mechanisms. In addition, the use of an implicit measure was crucial to characterize crossmodal SL.

Author contributions: D.D., R.D.B., and A.P.B. designed the experiment. D.D. and F.G. conducted the experiment. A.P.B. analyzed the data. F.G. and A.P.B. wrote the original draft. D.D. and F.G have contributed equally to the development of this project.

Acknowledgments: We thank Alicia Maciá-Mendoza for help in data collection. This work was supported by the AGAUR (BP 00213) to A.P.B. and through the European Regional Development Fund (PID2021-127146NB-I00 project) and the SGR (2021 SGR 00352) from AGAUR, Generalitat de Catalunya to R.D.B.

Conflict of interest statement: The authors declare no competing financial interests.

References

Alais, D., Morrone, C., & Burr, D. (2006). Separate attentional resources for vision and audition. *Proceedings. Biological Sciences / The Royal Society*, *273*(1592), 1339–1345. https://doi.org/10.1098/rspb.2005.3420

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting Linear Mixed-Effects Models using lme4*. *1*. https://doi.org/10.18637/jss.v067.i01

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(1), 24–39. https://doi.org/10.1037/0278-7393.31.1.24

Conway, C. M., & Christiansen, M. H. (2006). Statistical Learning Within and Across Modalities: Abstract versus Stimulus-Specific Representations. *Psychological Science*, *17*(10), 905–912.

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Speech segmentation is facilitated by visual cues. *Quarterly Journal of Experimental Psychology*, *63*(2), 260–274. https://doi.org/10.1080/17470210902888809

Cunillera, T., Laine, M., Càmara, E., & Rodríguez-Fornells, A. (2010). Bridging the gap between speech segmentation and word-to-world mappings: Evidence from an audiovisual statistical learning task. *Journal of Memory and Language*, *63*(3), 295–305. https://doi.org/10.1016/j.jml.2010.05.003

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499–504. https://doi.org/10.1111/1467-9280.00392

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117–125. https://doi.org/10.1016/j.tics.2014.12.010

Glicksohn, A., & Cohen, A. (2013). The role of cross-modal associations in statistical learning. *Psychonomic Bulletin & Review*, *20*(6), 1161–1169. https://doi.org/10.3758/s13423-013-0458-4

Gondan, M., Lange, K., Rösler, F., & Röder, B. (2004). The redundant target effect is affected by modality switch costs. *Psychonomic Bulletin and Review*, *11*(2), 307–313. https://doi.org/10.3758/BF03196575

López-Barroso, D., Cucurell, D., Rodríguez-Fornells, A., & de Diego-Balaguer, R. (2016). Attentional effects on rule extraction and consolidation from speech. *Cognition*, *152*, 61–69. https://doi.org/10.1016/j.cognition.2016.03.016

Mitchel, A. D., Christiansen, M. H., & Weiss, D. J. (2014). Multimodal integration in statistical learning: Evidence from the McGurk illusion. *Frontiers in Psychology*, *5*(MAY), 1–6. https://doi.org/10.3389/fpsyg.2014.00407

Mitchel, A. D., & Weiss, D. J. (2011). Learning across senses: Cross-modal effects in multisensory statistical learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(5), 1081–1091. https://doi.org/10.1037/a0023700

Orpella, J., Mas-Herrero, E., Ripollés, P., Marco-Pallarés, J., & de Diego-Balaguer, R. (2021). Language statistical learning responds to reinforcement learning principles rooted in the striatum. *PLOS Biology*, *19*(9), e3001119. https://doi.org/10.1371/journal.pbio.3001119

Otto, T. U., & Mamassian, P. (2010). Divided Attention and Sensory Integration: The Return of the Race Model . *Journal of Vision* , *10*(7), 863. http://www.journalofvision.org/content/10/7/863.abstract

Otto, T. U., & Mamassian, P. (2012). Noise and Correlations in Parallel Perceptual Decision Making. *Current Biology: CB*, 1–6. https://doi.org/10.1016/j.cub.2012.05.031

Redington, M., & Chater, N. (1996). Transfer in Artificial Grammar Learning: A Reevaluation. *Journal of Experimental Psychology: General*, *125*(2), 123–138. https://doi.org/10.1037/0096-3445.125.2.123

Richter, D., & de Lange, F. P. (2019). Statistical learning attenuates visual activity only for attended stimuli. *ELife*, *8*, 1–27. https://doi.org/10.7554/eLife.47869

Richter, D., Ekman, M., & de Lange, F. P. (2018). Suppressed sensory response to predictable object stimuli throughout the ventral visual stream. *Journal of Neuroscience*, *38*(34), 7452–7461. https://doi.org/10.1523/JNEUROSCI.3421-17.2018

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, *47*(1), 172–196. https://doi.org/10.1006/jmla.2001.2839

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. In *Science (New York, N.Y.)* (Vol. 274, Issue 5294, pp. 1926–1928). https://doi.org/10.1126/science.274.5294.1926

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52. https://doi.org/10.1016/S0010-0277(98)00075-4

Seitz, A. R., Kim, R., Van Wassenhove, V., & Shams, L. (2007). Simultaneous and independent acquisition of multisensory and unisensory associations. *Perception*, *36*(10), 1445–1453. https://doi.org/10.1068/p5843

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2017). Redefining "Learning" in Statistical Learning: What Does an Online Measure Reveal About the Assimilation of Visual Regularities? *Cognitive Science*, 1–36. https://doi.org/10.1111/cogs.12556

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001

Spence, C., Nicholls, M. E. R., & Driver, J. (2001). The cost of expecting events in the wrong sensory modality. *Perception and Psychophysics*, *63*(2), 330–336. https://doi.org/10.3758/BF03194473

Spotworks, S., & Berthoud, E. (2008). The Fractal Flame Algorithm. *Http://Citeseerx. Ist. Psu. Edu/Viewdoc/Summary*.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, *14*(9), 400–410. https://doi.org/10.1016/j.tics.2010.06.008

Thiessen, E. D. (2010). Effects of Visual Information on Adults' and Infants' Auditory Statistical Learning. *Cognitive Science*, *34*(6), 1093–1106. https://doi.org/10.1111/j.1551-6709.2010.01118.x

Toro, J. M., Sinnett, S., & Soto-faraco, S. (2005). *Speech segmentation by statistical learning depends on attention*. *97*, 25–34. https://doi.org/10.1016/j.cognition.2005.01.006

Treisman, a M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. http://www.ncbi.nlm.nih.gov/pubmed/7351125

Walk, A. M., & Conway, C. M. (2016). Cross-Domain Statistical–Sequential Dependencies Are Difficult to Learn. *Frontiers in Psychology*, *7*(February), 1–9. https://doi.org/10.3389/fpsyg.2016.00250

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*, 475–482.