Crossmodal correspondences – ecological priors or updated posteriors?

Author:

Laurits Wieslander Lyngbæk (AU702451)

202109726@post.au.dk

Faculty of Arts, Aarhus University

Bachelor of Cognitive Science

January 5th, 2023

This paper may be made public

27.423 | Character Count

Abstract:

Humans use perception to orient themselves in the world all the time. The perception we have of the world is a combination multiple senses that has been merged into one single mental representation of our environment. Our senses are very different in nature and must therefore be bound to objects as abstract features. But the brain has great trouble defining what features belong to what objects and what features are noise, this is the binding problem. It turns out that crossmodal correspondences is a principle of multi-sensory integration that helps solve the binding problem. The crossmodal correspondences come in a lot of variants, the statistical crossmodal correspondence turns out to follow a near optimal statistical model. The Bayesian Integration framework offers a model for how the brain completes this task. In this paper the theories of crossmodal correspondences will be discussed, and a speeded classification task will be used to investigate how the Bayesian integration work. The paper concludes that the brain is able to quickly update its statistical priors when interacting with novel stimuli that doesn't match expected correspondence-patterns.

Keywords: Crossmodal correspondences, the binding problem, multisensory integration, Bayesian integration, perception

Contents

1. Introduction

Brain function in general, and especially perception, have been recognized as modular in the last century- that is the brain is split into separate compartments each responsible for a singular action[1]. The modern human's cognition greatly really on their vision in everyday tasks, and the visual modality is often considered the dominant sensory modality independent of other sensory modalities. But, it turns out that the visual perception is influenced by the inputs from other sensory modalities[2].

1.1 Modular multisensory perception

It is generally believed that in the cerebral cortex there are different sensory modalities organized into unique and separate pathways that function independently of each other. The pathways operate largely independently until they reach higher level multisensory perceptual areas in the cortex, where a unified perception of the environment is produced[1]. Modular processing architecture allows for each subsystem to be optimized for its specific task, operating with maximum precision and speed, as the system doesn't have to compromise function for generalizability. A growing body of research indicates that modality-specific subsystems may not be completely independent, but rather the subsystems share information earlier in processing than generally thought. Perception of visual- and auditory sensations are proven to be processed as multisensory sensations in the intraparietal sulcus or the superior temporal polysensory area[2]. The benefit of integrating several modalities sensory input into one perception of the environment is that the variance of the final perceptual estimate is lessened.[3]

1.2 The Crossmodal Binding Problem

In many everyday situations we are bombarded with multiple unisensory inputs that are rich in information. These different unisensory inputs need to be processed and perceived as a coherent environment that we can interact with. So, to get the most reliable representation of our environment, all the stimuli need to be perceived as features belonging to objects in our environment. But, after our perception have merged these features into singular objects, we still need to be able to process and extract the different sensory features as separate items from the object[4].

This grouping of multiple sensory features into a single complete object in the perceiving mind is termed "*binding*" in modern terminology. *The binding problem* is referred to as a problem because of the difficulties that arises when attempting to decide what features come from the same object, and what features are noise that seems like it belongs to that object. The brain has increased problems with incorrectly bound features when trying to perceive another unimodal feature bound to that object.[5]

The brain needs to perceive all the information received from our senses as one singular multisensory environment in our mental representation. To overcome the problematic task of binding all the information rich features into objects the brain has developed certain principles of sensory binding. The most common principles regarding binding sensory features to mental representations, are principles such as *the temporal rule* and *the spatial rule*. According to these rules unisensory features that are perceived together in time and space, are more likely to be subjected to *multisensory integration*.[6] Recent papers have also pointed to evidence off *crossmodal correspondences* being a principle of multi-sensory integration[4,7].

1.2.1 Crossmodal correspondences

Crossmodal correspondences is a term used to describe the brain's tendency to associate certain dimensions of stimuli across sensory modalities. Crossmodal correspondences are shown to be a systematic and automatic feature. Crossmodal correspondences are systematic because they systematically apply to stimuli that can be described along an arbitrary dimension.[6] The majority of crossmodal studies investigate the relationship between stimuli from the visual modality and the auditory modality. The visual modality is investigated as it is considered our dominant sense. The auditory stimulus is often displayed in the form of varying pitch, as pitch is easy to manipulate and describable across a low to high dimension[8]. The pitch stimulus is paired with a visual stimulus that's also describable across a dimension, e.g. visual elevation[9], brightness[10], or size[9].

Most experiments conducted to research crossmodal correspondences are based on speeded classification tasks. That is, a task that investigates how quickly a participant can identify how a one-dimensional stimulus changes along that dimension. While the participants are exposed to the primary stimuli, they will also be exposed to a secondary task-irrelevant stimuli. The two modalities have crossmodal correspondences if the two stimuli can't be perceived independent of each other. The two stimuli can't be perceived independently if changing one stimulus independently causes changes in doing a task that only utilizes the other stimulus. Research documents that changing a crossmodally corresponding stimulus influences the reaction-time or accuracy of a task utilizing the other modality.[6]

If the effects of crossmodal correspondences are significant for perceptual binding in a simplified noise-free laboratory setting, then the effects must be imagined to be significant in an ecological setting as well. In ecological situations the perceptual load of crossmodal stimuli will be greatly increased, and the effects of crossmodal correspondences might therefore increase with the amount of stimuli and noise that need processing.[6]

1.3 Structural-, semantic-, and statistical correspondences

According to Spence's[4] review of crossmodal literature there exists evidence of several qualitatively different kinds of crossmodal correspondences- statistical, structural, and semantically mediated crossmodal correspondences.

Structural correspondences refer to the type of correspondences that are associated with structural neural systems we use to code sensory information[11,12]. One such system could be a generalized system for representing magnitude, like proposed in Walsh's *A Theory of Magnitude*[13]. A common structure for coding magnitude would provide the neural substrate for crossmodal correspondences between two modalities perceiving a stimulus describable by magnitude.

Semantically mediated correspondences are a result of a common linguistic term being used to describe stimuli across two different dimensions of stimuli[14]. Since these correspondences are linguistically embedded the kinds of stimuli that are affected by semantically mediated correspondences vary depending on the subject's language abilities. But certain types of descriptive adjectives are used to describe certain features in almost every language. Classically the words "low" and "high" are used to describe auditory stimuli that vary in pitch. These adjectives are also used to describe elevation of visual stimuli.

Statistical correspondences are the type of correspondences that often co-occur in nature, and therefore often are perceived together due to statistics. These correspondences are the neural-internalization of natural correlations attributes of environmental stimuli.[4] Such correspondences could be the size of an object and its resonant frequency, i.e., the larger an object is the lower is its frequency as well[15]. It seems that these types of statistical correspondences can be described favorably in terms of coupling priors in the framework of Bayesian integration theory[3,4].

1.3.1 Bayesian priors in crossmodal correspondences

The general idea of understanding crossmodal correspondences in terms of Bayesian priors is that the brain combines stimuli in a statistically optimal manner. It does this by combining prior knowledge and novel sensory information by their relative reliability[16]. The perceptual estimates of the senses are combined according to the rules of Maximum Likelihood Estimation to maximize overall perceptual precision.

Maximum Likelihood Estimation will be explained with the following explanation of Bayesian perception. When a human does a sensory measurement of an object multiple times, the estimates will have slightly different values each time. For simplicity it is expected that the measurement noise follows a Gaussian distribution, the most common distribution.

If the brain knows what types of sensory Likelihood Functions an environmental object evokes, it can match future sensory measurements to that object. The crossmodal aspect enters when two sensory Likelihood Functions are associated with the same object. The brain then integrates multiple Likelihood Functions as priors to create a crossmodal Maximum Likelihood Estimation[17], once integrated each component is no longer accessible individually without influence from the other component[16]. This crossmodal likelihood changes when a unimodal likelihood change, e.g., by an introduction of increased noise in that likelihood. At each step of estimating a maximum likelihood the internal model updates the prior Likelihood Functions based on the estimated prediction error and prior knowledge.[4,16,17]

2. Methodology of the experiment

This experiment aims to test the speed of which participants are able to update their perceptual priors of multisensory integration when a sense wrongly estimates a feature of an environmental object. This type of experiment is based on the paradigms of Rohde et. al in 2016[17] and Spence in 2011[4]. These studies are focusing on understanding special types of noise calibration of crossmodal correspondences. This study seeks to investigate how crossmodal correspondences optimize themselves when sensory Likelihood Functions have an incorrect mean due to incorrect knowledge of the environment, rather than predicting incorrectly because it is weighed incorrectly in the maximum likelihood estimation due to a sudden increase in noise.

2.1 Hypothesis

In the introductory part of the paper multiple theories regarding crossmodal correspondences was described, e.g., multisensory integration, feature binding and Bayesian inference. From these theories a fitting task for the experiment can be estimated to be a simple speeded classification task, as it should be able to induce crossmodal correspondences and allow for gathering data regarding the effects. The task should only concern itself with classification of a one-modal stimuli, whilst a stimulus presented to another modality should change depending on the condition. The secondary stimuli should vary unnaturally, so that the crossmodal correspondences predict incorrectly. By ensuring that the stimulus has an overrepresentation of stimuli in contradiction to natural statistical correspondences, the participants should be able to quickly update their priors and have this change of statistical expectations recorded in their behavioural data. But to make sure the priors change at all, the primary and first hypothesis is defined as:

Hypothesis 1: Reaction time will be shorter in the congruent task since the statistical priors will update during the training trials.

Null-hypothesis 1: The mean reaction-time for the incongruent task will be equal to- or shorter than the congruent task.

This means that in a model predicting reaction time by condition, the congruent task should result in a lower reaction time than the incongruent task. If this is the case, then the calibration to updated priors of crossmodal correspondences been successful, as current research suggests that it should be the opposite relation without calibration of priors. A secondary more complicated hypothesis should then be investigated:

Hypothesis 2: The chance of picking the correct answer will changes depending on the stimuli condition and the current trial-count, as the priors will be continuously updated to the task-specific statistics after every trial.

Null-hypothesis 2: The trial-count will not affect the chance of picking the correct answer.

If the secondary hypothesis can't be disproven it would indicate that the brain is able to update its crossmodal priors quickly to adapt to novel stimuli patterns.

2.2 Participants

27 participants without self-reported auditory deficits were enrolled in the study. The participants were 17 males (mean age 24.4 ± SD 9.0) and 10 females (mean age 24.6 ± SD 9.6). The age range was 18-60 years. The participants were not discriminated between regarding their native tongue as the stimuli is expected to be statistically corresponding. The participants were informed of risks of the study and gave their willing consent and agreed to their data being anonymized. The participants received no compensation for participation.

All participants were unaware of the goal of the study before entering, but 9 of the participants had undergone a mouse tracking workshop previously and were therefore aware of what data was being tracked in these types of tasks.

2.3 Materials and stimuli

The auditory stimuli consisted of 7 different pure tones in the range of 148 Hz – 4494 Hz each lasting 500 ms. The volume of the sound levels is regulated in accordance with the ISO 226 equal-loudness curves[18] to avoid crossmodal correspondences of loudness-size or loudness-pitch dimensions. The auditory stimuli were presented through a pair of Bluetooth connected Sony 1000XM4 at a volume of 20% with noise cancelling turned on.

The visual stimuli consisted of 5 circles with a varying diameter of respectively 2.6°-3.8° of the visual angle. The stimuli were presented for 500 ms on a Huawei Matebook X Pro 2020 13,9" screen 60 cm from the participant. The participants were instructed to not lean forward during the experiment.

The experiment was designed and presented in the Python IDE OpenSesame utilizing the PsychoPy backend. OpenSesame was responsible for both stimuli presentation and recording of behavioural responses. The computer was operated by a Logitech MX Master 3S mouse at 8200 DPI.

2.4 Experiment design and procedure

The experiment is a within-subject speeded classification response task regarding relative sizes of two sequentially presented visual stimuli. The visual stimuli will be paired with an auditory stimulus that decides the condition. The participants are instructed to focus on the visual stimuli of varying circles but aren't directly told to ignore the auditory stimulus. The participants are then shown then shown two buttons respectively in the upper left and right part of the screen. Clickable buttons on the screen are used instead buttons to capture mouse-movement as well as reaction time at the expense of introducing another possible correspondence. This procedure can also be seen on figure 1. The participants were informed about the task through a visual text instruction embedded in the experiment.

As the experiment is a within-subject task the subjects will be exposed 6 trials from all three conditions for at total of 18 trials, plus 5 practice trials. The five practice trials will always be consisting of the congruent condition, that is a trial where the Hz of the tone follows size of the circle according to the chart in figure 1. The following 18 trials will then have a randomized condition order. No feedback regarding the correctness of the participant's responses was provided at any point during the experiment.

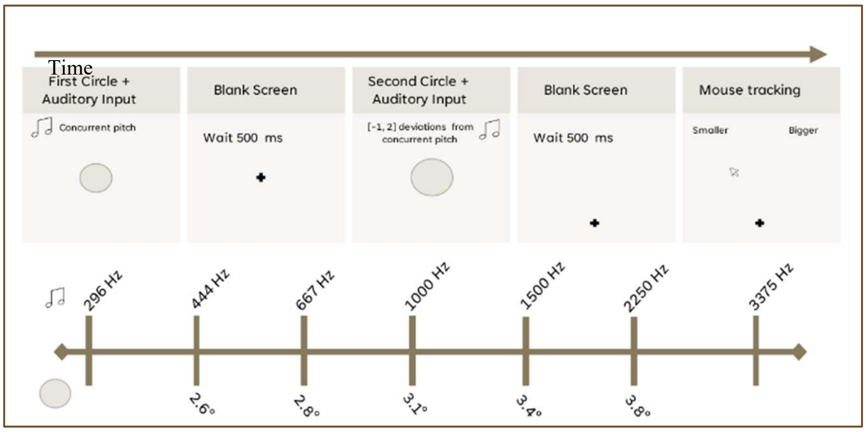

Figure 1.

The conditions of the experiment can be seen in figure 2. The Exaggerated condition is alike to the congruent condition, but the pitch increase/decrease is double the size.

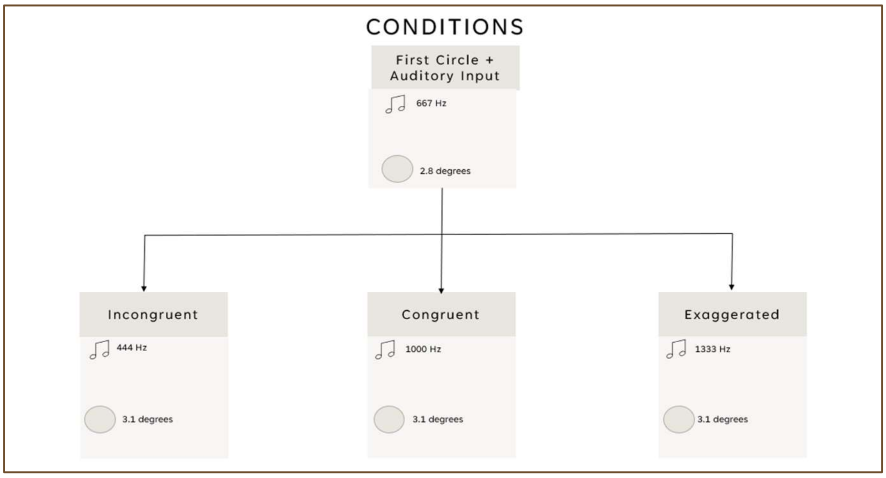

Figure 2.

The incongruent condition is only incongruent in the setting of this experiment. That is, the incongruent condition will only be incongruent in relation to the rest of the stimuli. The prior expectation documented by other studies indicates that the Hz should decrease when the visual angle increases. This condition is statistically considered the congruent stimuli parring, when looking at encountered stimuli in the natural world.

2.5 Analysis and results.

The mean accuracy of the participants was 85 % correct, ranging from 70% to 100%. The mean reaction times across participants was between 0.9-1.0 seconds (SD 0.40-0.55 seconds). The mean response time and accuracy has been visualized in figure 3.

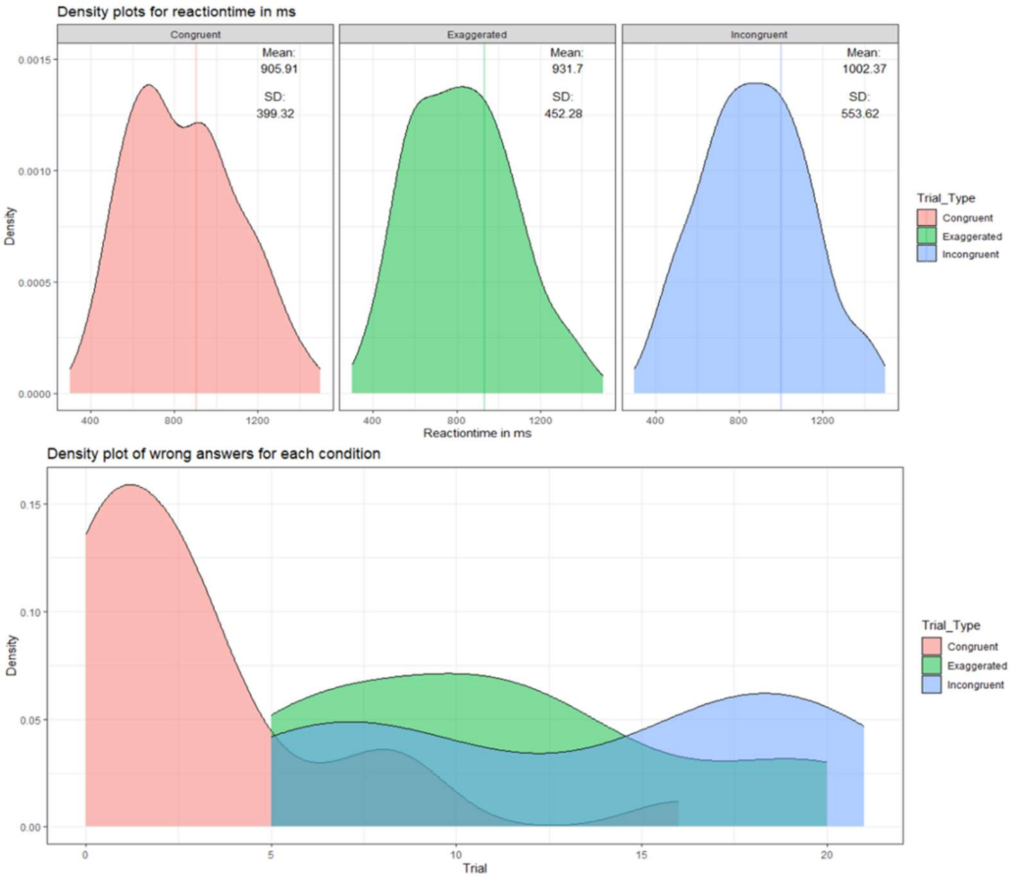

Figure 3 – Density plots of answers and reaction times

2.5.1 – Lognormal linear mixed effect model

To test the first hypothesis, the lme4 package (Bates et al., 2014[19]) for R (R Core Team[20]) was used to fit a linear mixed effect model to the data. The model was defined as such:

$$ReactionTime \sim Condition + (1 \mid ID)$$

The model was fitted on the data from the trials after the 5 training trials. This was done to avoid mistakenly interpreting the data as behavioural perceptual processes, when it might just be participants trying to understand the experimental flow. Participants ID was modelled as a random intercept as it was expected that each participant had a different baseline of reaction time. Visual inspection found no deviations from normality or homoscedasticity after the data had ben log transformed to account for the skewed nature of reaction time. The reported β-values are on the log-scale.

The mixed effect model revealed that congruence was a significant predictor of reaction time ($\beta = 6.809$, $SE = .0417$, $t = 163.4$, $p < .001$). The model showed no real effect of the exaggerated condition as a predictor of reaction time ($\beta = 0.028$, $SE = .0584$, $t = 163.4$, $p = .63$). The exaggerated condition isn't necessary for the first hypothesis to pass, but only a leftover from an earlier stage of the experiments design, as will be discussed later. Using the incongruent condition as a predictor showed a greater difference than the exaggerated condition. The incongruent condition had a minor effect, ($\beta = 0.101$, $SE = .056$, $t = 1.781$, $p = .075$).

2.5.2 – Interacting effects

The study therefore moved on to investigate the second hypothesis by fitting a model of interactions between trial-count and condition. The R package lme4 was also used here with the binomial family. An interaction effect was used as the Bayesian theory hypothesis indicates that the effects will be crossing as the trial count increases. All trials was included in this model as changes in correspondences are expected to also happen during training trials.

*Correct ~ Trial-count * Condition + (1 | ID)*

Given the comprehensive nature of interaction effects, the results of the model will be displayed in table 1 for simplicity. The raw data will also be displayed graphically in the lower part of figure 3.

Fixed effects	*Estimate*	*Std. Error*	*Z value*	*p-value*
(intercept)	1.418	0.337	4.207	p < .001 ***
Trial-count	0.199	0.058	3.462	p < .001 ***
Exaggerated	0.595	0.929	0.640	p = 0.52
Incongruent	1.497	0.807	1.856	p = 0.06 (*)
Trial:Exaggerated	-0.120	0.092	-0.311	p = 0.189
Trial:Incongruent	-0.231	0.078	-2.950	p = 0.003 **

Table 1.

3. Discussion

The aim of this pilot study was to investigate how crossmodal correspondences update their priors when the unimodal stimuli repeatedly is the opposite of what is expected. To do so, the chosen experiment investigates the crossmodal correspondences of pitch and size, as this is one of the best documented effects in crossmodal research. A single experiment was enough to answer both hypotheses satisfactorily given the scope of the paper.

3.1 Models and hypotheses

To answer the research question two hypothesis were created. The first hypothesis H_1 aimed to test if the unnaturally skewed stimuli combination and the 5 training trials were sufficient to update the prior expectations of statistical correspondences that is documented by other studies. The linear mixed effect model fitted to predict reaction time had a very significant intercept predicting the congruent condition. The incongruent condition almost reached a significant p-value of < 0.05. But given the small sample size of this study, this p-value is judged to be enough to continue with the hypotheses as proofs of concept. According to the model, the incongruent condition had an increased mean reaction time, and H_1 is therefore not falsified.

The second hypothesis H_2 regards the cognitive systems ability to update incorrect sensory integration priors. To test this, a model with interacting effects between Trial and Condition was fitted to the data. This model used data from all trials, including the training trials, as the training trials was expected to contain information about the prior's ability to update. As seen in table 1, the four values that atleast reached a p-value of 0.1, was the intercept, trial-count, incongruent condition and the interaction effect between trial-count and incongruent condition. A possible interpretation of this model that tries to answer the hypothesis could be as follows.

All trials have an average chance of being answered correctly, and it seems that the congruent condition starts of being least likely to be correct. But as the trial count increases, the congruent condition's chance of being correct will increase, as Trial-count is a positive number. The incongruent condition starts of with a much higher probability of correct trials, but as seen in the Trial:Incongruent estimate, the chance of guessing correct will fall as the Trial-count increases for the incongruent condition. This gives us a very simple model indicating that the statistical crossmodal correspondences will be updated to novel stimuli relations. The hypothesis therefore confirmed, and H^2 can't be disproven.

3.2 Mouse-tracking, exaggeration, and confusing names.

A few considerations for future experiments became noticeable while creating the experiment and through post-hoc analysis of the data. In the first stages of planning this experiment and collecting data it was meant to be an experimental design that could answer mouse tracking hypothesis about congruent and incongruent crossmodal correspondences. But during the data collection it became noticeable that the congruent and incongruent tasks had been swapped by a mistake, this meant that the participants was been trained on naturally-incongruent statistical correspondences. Instead of recreating the experiment and recollecting the data, the Bayesian inference H_2 hypothesis was constructed around the new experiment before looking at the data. A few archaeological remains are therefore still influencing the experimental structure, like the use of left-right clickable digital buttons that may influence correspondences. A future version of this experimental design would probably choose two buttons on the keyboard. It is also this history that is the reasoning for the exaggerated condition, as the study was originally trying to answer a way simpler question, and therefore had room for an added nuance of congruency.

Future research should investigate statistical congruencies' adaptability to changes in modular perception, to understand the differences between adapting to noise and adapting to changes in the environment. An improvement of the experimental design could be achieved by introducing an in-between participant's study where the training section is varying in trial length depending on condition, and then showing 50 % congruent and 50% incongruent stimuli.

It was however outside the resources and scope of this proof-of-concept paper to create a function or general model of crossmodal correspondences' adaptability to change. This paper however did indicate that the brain seems to quickly be able to change what direction of dimensional correspondences it considers congruent.

4. Conclusion

This study aimed to investigate how statistical crossmodal correspondences update priors when an object is repeatedly perceived differently than a unimodal stimulus expects. The auditory stimuli used was inverted in relation to previous research's description of congruent pitch-size correspondences. Using a combination of linear- and binomial regression-models it was shown that the brain is able to quickly recalibrate to novel statistical correspondences of bimodal sensory information. The paper was therefore unable to falsify any of its hypotheses but had a few methodological problems that should be resolved in future studies. Future research should investigate how well statistical irregularities linger after exposure. A methodological consideration for future studies is to avoid having two conditions moving along the direction of a single stimuli dimension when attempting to calculate change in the direction of crossmodal correspondences.

5. References

1. Shams L. Early Integration and Bayesian Causal Inference in Multisensory Perceptio. Published online 2012. Accessed December 22, 2022. https://www.ncbi.nlm.nih.gov/books/NBK92847

2. Foundations of Sensation and Perception. Routledge & CRC Press. Accessed December 25, 2022. https://www.routledge.com/Foundations-of-Sensation-and-Perception/Mather/p/book/9781848723443

3. Ernst MO. A Bayesian view on multimodal cue integration.

4. Spence C. Crossmodal correspondences: A tutorial review. *Atten Percept Psychophys*. 2011;73(4):971-995. doi:10.3758/s13414-010-0073-7

5. Burwick T. The binding problem. *WIREs Cognitive Science*. 2014;5(3):305-315. doi:10.1002/wcs.1279

6. Klapetek A, Ngo MK, Spence C. Does crossmodal correspondence modulate the facilitatory effect of auditory cues on visual search? *Atten Percept Psychophys*. 2012;74(6):1154-1167. doi:10.3758/s13414-012-0317-9

7. Calvert GA, Thesen T. Multisensory integration: methodological approaches and emerging principles in the human brain. *Journal of Physiology-Paris*. 2004;98(1):191-205. doi:10.1016/j.jphysparis.2004.03.018

8. Spence C, Deroy O. How automatic are crossmodal correspondences? *Consciousness and Cognition*. 2013;22(1):245-260. doi:10.1016/j.concog.2012.12.006

9. Evans KK, Treisman A. Natural cross-modal mappings between visual and auditory features. *Journal of Vision*. 2010;10(1):6. doi:10.1167/10.1.6

10. Martino G, Marks LE. Perceptual and Linguistic Interactions in Speeded Classification: Tests of the Semantic Coding Hypothesis. *Perception*. 1999;28(7):903-923. doi:10.1068/p2866

11. Mondloch CJ, Maurer D. Do small white balls squeak? Pitch-object correspondences in young children. *Cognitive, Affective, & Behavioral Neuroscience*. 2004;4(2):133-136. doi:10.3758/CABN.4.2.133

12. Pinel P, Piazza M, Le Bihan D, Dehaene S. Distributed and Overlapping Cerebral Representations of Number, Size, and Luminance during Comparative Judgments. *Neuron*. 2004;41(6):983-993. doi:10.1016/S0896-6273(04)00107-2

13. Walsh V. A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*. 2003;7(11):483-488. doi:10.1016/j.tics.2003.09.002

14. Multisensory synesthetic interactions in the speeded classification of visual size | SpringerLink. Accessed January 2, 2023. https://link.springer.com/article/10.3758/BF03193720

15. Coward SW, Stevens CJ. Extracting Meaning from Sound: Nomic Mappings, Everyday Listening, and Perceiving Object Size from Frequency. *Psychol Rec*. 2004;54(3):349-364. doi:10.1007/BF03395478

16. Brunel L, Carvalho PF, Goldstone RL. It does belong together: cross-modal correspondences influence cross-modal integration during perceptual learning. *Frontiers in Psychology*. 2015;6. Accessed January 3, 2023. https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00358

17. Rohde M, van Dam LCJ, Ernst MO. Statistically Optimal Multisensory Cue Integration: A Practical Tutorial. *Multisens Res*. 2016;29(4-5):279-317. doi:10.1163/22134808-00002510

18. Equal-loudness-level contours for pure tones: The Journal of the Acoustical Society of America: Vol 116, No 2. Accessed January 2, 2023. https://asa.scitation.org/doi/abs/10.1121/1.1763601?casa_token=0K5OSJ1y6J0AAAAA:y4qeUd6TsB0PIsAmjhqX8IiWwZuPiuxOUJAnnt-78PplV02k4M8f6KX6gt3CTDmShes1PlRTBqoIJw

19. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. 2015;67:1-48. doi:10.18637/jss.v067.i01

20. R Core Team (2020). — European Environment Agency. Accessed January 4, 2023. https://www.eea.europa.eu/data-and-maps/indicators/oxygen-consuming-substances-in-rivers/r-development-core-team-2006