PROBABILISTIC MODELS OF AUDIOVISUAL PERCEPTUAL DECISION MAKING

by

DAVID MEIJER

A thesis submitted to the University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

School of Psychology

College of Life and Environmental Sciences

University of Birmingham

November 2018

ABSTRACT

Multisensory integration is a fundamental component of perceptual decision making and an excellent example of how the brain deals with the abundance of sensory uncertainty in order to create a coherent understanding of its environment. In this thesis I evaluate the two most popular computational models for describing multisensory integrative processes: maximum likelihood estimation (MLE) and Bayesian causal inference (BCI). Both models predict statistically-optimal sensory integration, but in so doing MLE makes the critical assumption that the brain always fuses sensory signals under certain experimental circumstances, whereas BCI allows for flexibility by assessing whether integration is appropriate for a particular set of stimuli. In two empirical studies on audiovisual spatial integration I expose considerable limitations of MLE in explaining human behavioral results and advocate the use of BCI to evaluate multisensory integration, even under conditions that were previously thought of as optimized for MLE. In a final empirical chapter I test an important prediction that both models make: that sensory uncertainty is reduced for integrated multisensory signals. I present behavioral evidence that confirms this prediction by showing that observers' confidence levels increase as a result of audiovisual integration, thereby further validating the use of probabilistic models to describe multisensory perception.

ACKNOWLEDGEMENTS

When I started this PhD nearly four years ago I had no idea what the words in my eventual title really meant. Multisensory integration? Probabilistic models? I quickly learned that the ventriloquist effect is a very effective audiovisual spatial illusion when I became a test subject for one of my colleagues' pilot experiments and utterly failed to segregate the two sensory signals. It took me a little longer to understand what Bayes' rule was all about. But slowly, as the years passed, and as I began to comprehend that the smallest details can make a huge difference in psychophysics experiments, I recognized and have learned to utilize the explanatory power of statistical modelling to improve my own and hopefully other people's understanding of multisensory perceptual processes in the human brain. I fully realise that I would never have been able to make this exploratory journey on my own. Therefore, I would like to express my sincere gratitude to the following people:

- Uta Noppeney, my PhD supervisor, beacon of knowledge and wisdom, and occasional source of frustration because of yet another interesting research proposal that would take up all my time in the near future.
- Sam Jones, my scientific sparring partner and Goose-buddy that has become a good friend, and who I hope to keep seeing for many years to come.
- Agoston Mihalik, the chess-master with whom I spend so many late nights in the office, working, listening to ridiculously loud music and having a fabulous time.
- Steffen Bürgers, source of constructive discussions on psychometric functions and the person who dragged me out the office to enjoy climbing.

- Tom White, the ever optimistic, though sometimes sleep-deprived, colleague that unfortunately left the lab too early.
- Máté Aller, who gave me a head start by patiently explaining numerous technical issues of the laboratory equipment.
- Ambra Ferrari, the happy smiley lady of the lab, with whom I had the pleasure of working together on two experimental projects.
- Arianna Zuanazzi, who managed to raise sarcasm and irony to new heights and was sometimes a source of disturbing but silly confusion.
- Giulio Degano, the third, but definitely most pronounced, Italian in our lab whose climbing skills are an inspiration and cause of humility.
- Patrycja Delong, whose Polish wodka parties provided a fun opportunity to get away from it all, only for 'it' to kick back harder the next day.
- Michael Joannou and Alex Murphy, latecomers to the lab, but two reasons why I would happily stick around for a little longer.
- Rémi Gau and Johanna Zumer Leggett, the lab's dinosaurs, with unlimited willingness to show the younger generation how it's done.
- Martin Šmíra and Alex Krugliak who also left too early.
- All the students I worked with over the years: Harriet Clarke, Letitia Phillips, Hannah Evans, Kerstin Wenzel, Maddison Roberts, Fatima Syeda, Urmila Sudhindra, Sonal Patel, Bethany Robinson, Samina Begum, Wing Lam Chan (a.k.a. Chloe), Sebastijan Veselič, Carmelo Calafiore, and Jake Rhodes. You gave my PhD an additional dimension and showed me that teaching can be fun.

A big giant thank you to all of you!!

Last but not least, I wish to thank my mum, my dad, and my love Ivana. You have always been there for me. Supported me in difficult times and brought me joy and comfort when it was much needed. Thank you so much for being in my life, even at times when this PhD caused me to be absent and distant from yours. I hope the future will bring us closer together.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Chapter 2

Chapter 3

Chapter 4

Chapter 5

LIST OF RECURRING ABBREVIATIONS

2IFC	Two-interval forced-choice
A	Auditory
ALVR	Auditory left visual right
ANOVA	Analysis of variance
AV	Audiovisual
BCI	Bayesian causal inference
BF	Bayes factor
BOR	Bayesian omnibus risk
BVRE	Bayesian visual reliability estimates
$\boldsymbol{C}$	Causal structure
$\boldsymbol{C_{2\Delta x}}$	Type-2 criteria in terms of internal estimate difference
cd	Candela
dB	Decibel
EEG	Electroencephalography
ERC	European Research Council
fMRI	Functional magnetic resonance imaging
JND	Just noticeable difference
L	Likelihood
LL	Log-likelihood
$\boldsymbol{LL_{cv}}$	Cross-validated log-likelihood
MAP	Maximum a posteriori estimate

MCMC	Markov chain Monte Carlo
MLE	Maximum likelihood estimation
ms	milliseconds
P	Probability
PSE	Point of subjective equality
PXP	Protected exceedance probability
$\boldsymbol{r_A}$	Auditory reliability
$\boldsymbol{S}$	True spatial location of stimulus
$\boldsymbol{\widehat{S}}$	Estimate of stimulus location
SD	Standard deviation
SEM	Standard error of the mean
SPL	Sound pressure level
V	Visual
V1	Primary visual cortex
VLAR	Visual left auditory right
$\boldsymbol{w_A}$	Auditory weight
$\boldsymbol{w_S}$	Sensory weight (as opposed to prior weight)
$\boldsymbol{x_A}$	Internal auditory estimate
$\Delta\boldsymbol{AV}$	Audiovisual disparity
η	Betabinomial noise factor
λ	Lapse rate
$\boldsymbol{\mu}$	Mean of Gaussian distribution
$\boldsymbol{\sigma}$	Standard deviation of Gaussian distibution

CHAPTER 1: PREFACE

The experimental focus of my PhD throughout the years can perhaps best be described as turbulent. I started off by studying the effects of visual blinding on audiovisual integration by means of the attentional blink with electroencephalography (EEG). Although my entire first year and more was spent on that project, it did not make it into the current thesis. Instead, my interests shifted from visual awareness ratings to metacognition in general and its intricate interplay with multisensory integration in particular. More importantly, they drifted away from EEG and towards Bayesian inference.

It all began as 'a little project on the side helping out an undergrad student'. However, the seemingly simple research objective of testing whether multisensory integration leads to a boost in confidence, turned out to be a rocky road full of obstacles. First, I had to ensure empirically that multisensory integration in our experiment did indeed result in increased precision, as is often claimed, for that was the assumption which the confidence hypothesis relied upon. But my initial attempts at replicating this fundamental characteristic of multisensory integration were unsuccessful.

Over time, I began to better understand the limitations of the so-called maximum likelihood estimation (MLE) 'model' that was used to describe multisensory integration (see chapter 3). Moreover, by studying assumptions that the MLE model makes, I was able to explain why the behavioral data sets that I acquired consistently deviated from MLE predictions (chapter 4). Finally, with a renewed understanding of multisensory

integration and an improved psychophysics skills-set I was able to adequately address the 'simple' metacognitive research question. As expected, multisensory integration raised participants' confidence levels, but I had to develop a novel methodological approach to unambiguously demonstrate that (chapter 5).

The first and last chapter of this thesis bind the empirical work together by providing a literature review of the current computational models of multisensory integration (chapter 2) and an outlook towards future experimental research directions (chapter 6).

CHAPTER 2

COMPUTATIONAL MODELS OF MULTISENSORY INTEGRATION

David Meijer, Uta Noppeney

CONTRIBUTIONS

David Meijer and Uta Noppeney wrote the text

Figures and simulations by David Meijer

ACKNOWLEDGMENTS

This research was funded by ERC-2012-StG_20111109 multsens

TO BE PUBLISHED

As a book chapter in: Multisensory Perception: From Laboratory To Clinic

Editors: Ramachandran, V.S. & Sathian, K. Publisher: Elsevier

1. Introduction

Various sensory organs continuously provide our brains with uncertain information about our environment. Critically, every sensor has its specific limitations. For example, the sensitivity of our eyes' photoreceptors is optimized for use during daylight (e.g. photoreceptor sensitivity of nocturnal insects is much higher; Honkanen, Immonen, Salmela, Heimonen & Weckstrom, 2017). Our ears are specialized for detecting differences in sound pitch, but they provide only imprecise estimates for the location of a sound's source.

Imagine you are in a dimly lit bedroom at night and you hear the sound of a mosquito. To obtain the most precise estimate of the mosquito's location the brain should combine uncertain spatial information furnished by the auditory and visual senses. Critically, the brain should integrate sensory signals only when they pertain to the same event, but process them independently when they come from different events. For example, we are all familiar with those vague black spots on the wall that look annoyingly like mosquitos in the dark. These immobile black spots should not be integrated with the mosquito's buzzing sound around the head. In short, in order to generate a coherent percept of the environment, the brain needs to infer whether or not sensory signals are caused by common or independent sources. This process has been termed multisensory causal inference (Shams & Beierholm, 2010).

In this chapter we will explore the computational operations that the brain may use to solve these two challenges involved in multisensory perception, i.e. (i) how to weight and integrate signals that come from a common source into a unified percept and (ii) how to infer whether signals come from common or independent sources.

In the first section, we will introduce the normative Bayesian framework focusing on perception based on input from a single sensory channel and prior expectations. In the second section, we will describe how the brain integrates signals from multiple sensory channels pertaining to the same event into a unified percept (i.e. so-called forced fusion model). In the third section, we will explore the more general case of multisensory perception in the face of uncertainty about the world's causal structure, i.e. uncertainty about whether signals are caused by common or independent sources. Hence, this final case combines the two challenges facing the brain in a multisensory world: causal inference and weighted sensory integration. Each section first describes the normative Bayesian model and then briefly reviews the empirical evidence that shows the extent to which data from human or non-human primates are in accordance with those computational principles.

2. Combining information from a single sensory channel with prior knowledge

Any sensory signal that reaches the cerebral cortex is inevitably contaminated with various sources of noise. Let us consider how an observer can estimate the location of an event for spatial orienting from visual inputs. An observer's eyes are bombarded with photons, and each eye's lens refracts the photons such that a ray of focused light hits the retina. There, photoreceptors and ganglion cells transform the electromagnetic radiation into action potentials. This eventually, via several synapses, results in an activity pattern in the visual cortex. Importantly, noise may be introduced at each of those processing stages. The eye's view can be partially obscured by a dirty window, and its lens is unlikely to be perfectly in focus; the transformation from photons to action

potentials functions in bulk (Barlow, 1956); and synaptic transmission is a probabilistic process (Stevens, 2003). In short, the sensory organs and systems provide the brain only with an uncertain or noisy estimate of a particular property (e.g. spatial location) of events and objects in the outside world.

To constrain perceptual inference the observer can combine the noisy sensory evidence with *prior* knowledge or expectations. For example, in our natural environment it is very unlikely to observe a concave human face, where the tip of the nose faces away from the observer. When an observer is shown the inside of a mask, the brain often falsely interprets the image such that the nose is perceived to be facing the observer. The visual hollow-face illusion, as this effect was dubbed, is only one of many examples where prior knowledge affects our perception (Gregory, 1997).

The normative Bayesian framework in neuroscience posits that the brain forms a probabilistic generative model of the sensory inputs that is inverted during perceptual inference (= recognition model). Bayesian probability theory offers a precise formulation of how observers should combine uncertain information such as different sorts of noisy sensory evidence and prior knowledge to form the most reliable representation of the world. It thus sets a benchmark of a so-called 'ideal observer' or optimal performance given a particular loss function against which an organism's neural and behavioral responses can be compared.

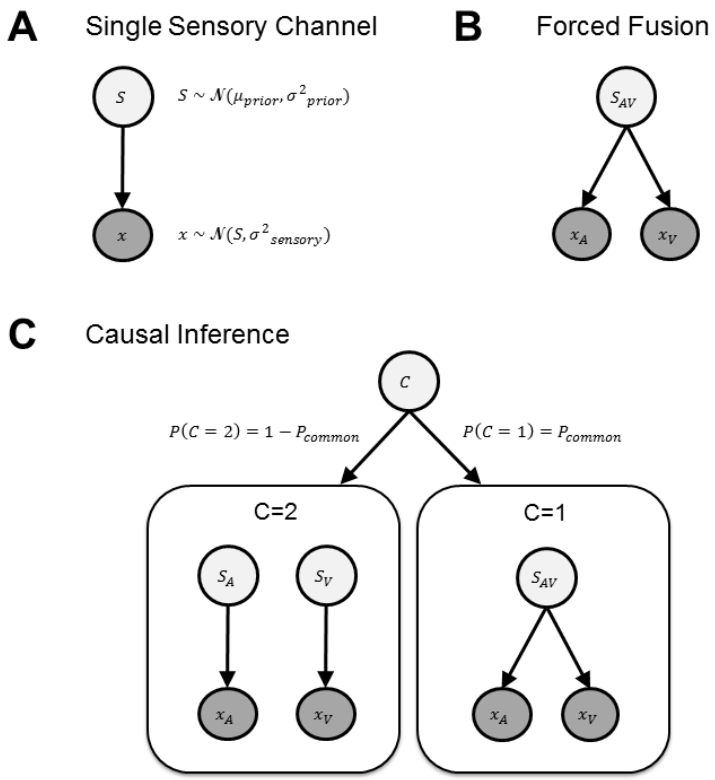


Fig 1 - Generative models corresponding to the three different cases. Fig. 1 a. Single sensory signal: A hidden source generates a sensory signal that is corrupted by noise. Fig. 1 b. Forced Fusion: A hidden source generates two sensory signals (e.g. auditory and visual) that are independently corrupted by noise. Fig. 1 c. Causal inference model explicitly models the potential causal structures that could have generated the two sensory signals (e.g. auditory and visual). In the full segregation model component (left) two independent hidden sources generate the auditory and visual signals. In the forced fusion model component, a common source generates two sensory signals (e.g. auditory and visual). A Bayesian causal inference estimate combines the estimates obtained from those two model components using a specific decision function (e.g. model averaging). Adapted from Körding, Beierholm, Ma et al. (2007).

Figure 1A shows the graphical model that illustrates the generative process for the spatial localization example above based on a single sensory channel and prior knowledge. A hidden source at the true location S generates a noisy sensory signal representation X. The true location S is sampled from a prior distribution, which is often assumed to be a Gaussian with mean μ: $S \sim \mathrm{N}(\mu_{prior}, {\sigma^2}_{prior})$. The sensory signal is corrupted by noise, i.e. sampled from a Gaussian centred on the true source location: $x \sim \mathrm{N}(S, {\sigma^2}_{sensory})$. The generative model defines the probability of each sensory input given a particular source location $P(x|S)$. During perception the observer needs to invert this generative model to compute the posterior probability $P(S|x)$, i.e. the probability of a spatial location given the sensory input x, by combining sensory evidence and prior knowledge. According to Bayes' rule, the posterior probability of a spatial location given a particular sensory input, $P(S|x)$, is proportional to the product of the likelihood $P(x|S)$ and the prior $P(S)$:

(1) $$P(S|x) = \frac{P(x|S)*P(S)}{P(x)} \propto P(x|S) * P(S)$$

The normalization constant $P(x)$ can be obtained from the product of the likelihood function and the prior by marginalizing (i.e. integrating) over all possible locations S:

(2) $$P(x) = \int P(x|S) * P(S) * dS$$

The observer then needs to minimize a particular loss function that specifies the cost of selecting the estimate $\hat{S}$ given the true location S in order to report a final point estimate. For instance, using the squared error loss function the observer would report the mean of the posterior distribution as the final spatial estimate. By contrast, using a zero-one loss function the observer would report the maximum a posteriori estimate

(MAP), i.e. the mode of the posterior distribution. Critically, under Gaussian assumptions of both prior and likelihood the posterior mean and mode are identical, i.e. both loss functions yield the same final estimate. However, asymmetric posterior distributions lead to different estimates for the posterior mean and MAP (Körding, Beierholm, Ma et al., 2007; Yuille & Bulthoff, 1996).

Priors can emerge at multiple timescales potentially ranging from seconds to evolutionary times. For instance, during evolution certain hardwired neural priors may have emerged as a result of selection pressures (Geisler & Diehl, 2002). Likewise, other hardwired priors may be fine-tuned during neurodevelopment when the immature brain is exposed to the statistics of the sensory inputs (Gopnik & Tenenbaum, 2007). Finally, the brain is thought to rapidly adjust priors to changes in the input statistics across and perhaps even within trials where the posterior of the current trial or time point forms the prior for the next trial or time point (Di Luca & Rhodes, 2016; Roach, McGraw, Whitaker & Heron, 2017). Priors are critical to constrain perceptual inference in the face of uncertainty resulting from noise, occlusion etc. As we will derive in greater detail in the next 'forced fusion' section, the influence of the prior on the final posterior estimate should be greatest if the sensory input is noisy and uncertain. This is because different sorts of evidence (e.g. prior vs. sensory evidence or different sensory evidences) should be combined in a manner weighted by their relative reliabilities (see Section 3 for details).

Priors can be formed about all sorts of properties such as spatial location, shape, speed etc. Indeed, numerous studies have demonstrated how prior knowledge or expectations shape and bias perceptual inference in our natural environment or designed

experimental settings: the light-from-above prior (objects with ambiguous depth seem to face forward if the shadow is below them; Mamassian & Landy, 2001), the circularity assumption (we tend to think that an object's depth is equal to its width; Jacobs, 1999), the foveal bias (relevant objects are more likely to appear in the centre of our field of view; Kerzel, 2002; Odegaard, Wozny & Shams, 2015), the slow speed preference (most objects do not move or tend to move slowly; Stocker & Simoncelli, 2006; Weiss, Simoncelli & Adelson, 2002), and the cardinal orientation prior (vertical and horizontal orientations can be more frequently found (Girshick, Landy & Simoncelli, 2011). In the latter example the experimentally determined probabilities of the human prior distribution for orientations were shown to match the environmental statistics for orientations that were found in a large set of photographs. In addition to the long-term priors, the brain can also rapidly adapt priors to the dynamics of statistical regularities. In laboratory experiments participants may learn the distribution from which the stimuli are sampled (e.g. the range of stimulus durations in a time-interval estimation task; Jazayeri & Shadlen, 2010). In the real world they can adopt prior distributions that apply to a particular situation (e.g. the typical velocities for a ball in a game of tennis; Kording & Wolpert, 2004). Multiple studies have also shown that the biasing influence of the prior is - as expected (see above) - inversely related to the reliability of the sensory stimuli (Girshick et al., 2011; Jazayeri & Shadlen, 2010; Kording & Wolpert, 2004; Stocker & Simoncelli, 2006; Weiss et al., 2002).

At the neural level, a recent functional magnetic resonance imaging (fMRI) study has shown that the brain estimates the reliability or precision of sensory representations in primary visual cortex (V1) on a trial-by-trial basis (van Bergen, Ma, Pratte & Jehee,

2015). Participants were presented with visual gratings that varied in their orientation across trials. On each trial they indicated the perceived orientation using a rotating bar. Critically, even though no external noise was added to the stimuli, the precision of sensory representations in V1 may vary across trials because of internal neural noise. Indeed, the uncertainty estimated from the activity patterns in the visual cortex varied across trials. Moreover, it correlated positively with the variance of participants' responses, and negatively with their orientation errors. The results of this study suggest that sensory cortices represent stimulus uncertainty on a trial-by-trial basis and that this uncertainty affects behavioral performance, as predicted by probabilistic models of Bayesian inference.

3. Forced fusion: Integrating sensory signals that come from a common source

Many events and objects in the natural environment can be perceived concurrently by multiple senses that are each specialized for specific features of the outside world. Signals from different senses can provide complementary information. For instance, honey can be perceived as yellow by vision, but tastes sweet. Alternatively, multiple senses can provide redundant information about the same physical property such as spatial location. Thus, we can locate a puncture in a bicycle's inner tube by vision, audition or touch (i.e. seeing, hearing or feeling where the air flows out of the tube). In the case of redundant information across the senses, multisensory perception enables the observer to form a more precise or reliable (reliability being the inverse of variance) estimate of the environmental property in question by integrating evidence across the senses.

Figure 1B shows the generative model for spatial localization based on redundant auditory and visual information. The generative model assumes one single source at the true location S_{AV} that emits two internal sensory signals; in this case a visual and an auditory signal: x_A and x_V. As we do not allow for the two signals to be generated by two independent sources, we refer to this generative model as the forced fusion scenario, where optimal performance can be obtained by mandatory sensory integration. Again, as in the unisensory case, we assume that the auditory and visual signals, x_A and x_V are corrupted by independent Gaussian noise. Hence, we sample x_A and x_V independently according to $x_A \sim \mathrm{N}(S_{AV}, {\sigma^2}_A)$ and $x_V \sim \mathrm{N}(S_{AV}, {\sigma^2}_V)$.

During perceptual inference, the observer needs to compute the posterior probability of the spatial location given auditory and visual inputs according to Bayes' theorem:

$$(3) \qquad P(S_{AV}|x_A, x_V) = \frac{P(x_A, x_V|S_{AV}) * P(S_{AV})}{P(x_A, x_V)} \propto P(x_A, x_V|S_{AV}) * P(S_{AV})$$

Further, as auditory and visual inputs are assumed to be conditionally independent (i.e. independent noise assumption across sensory channels), we can factorize the likelihood (Oruç, Maloney & Landy, 2003):

$$(4) \qquad P(S_{AV}|x_A, x_V) \propto P(x_A|S_{AV}) * P(x_V|S_{AV}) * P(S_{AV})$$

Further, most studies in multisensory integration assume an uninformative or flat prior $P(S_{AV})$, where we can ignore the influence of the prior. As a result, the maximum a posteriori estimate turns into a maximum likelihood estimate:

$$(5) \qquad P(S_{AV}|x_A, x_V) \propto P(x_A|S_{AV}) * P(x_V|S_{AV})$$

Assuming independent Gaussian noise and uninformative priors, the optimal, most precise (i.e. most reliable or with minimum variance) audiovisual estimate $\hat{S}_{AV}$ can be computed as a reliability-weighted linear average of the two unisensory estimates (Ernst & Banks, 2002; Oruç et al., 2003):

(6) $$\hat{S}_{AV} = w_A\hat{S}_A + w_V\hat{S}_V \quad \text{with} \quad w_A = \frac{r_A}{r_A+r_V} \quad \text{and} \quad w_V = \frac{r_V}{r_A+r_V} = 1 - w_A$$

where the reliability is defined as the inverse of the Gaussian's variance: $r = \frac{1}{\sigma^2}$. Moreover, the reliability of this audiovisual estimate can be expressed as the sum of the two unisensory reliabilities:

(7) $$r_{AV} = r_A + r_V \qquad \text{which is equivalent to:} \qquad {\sigma^2}_{AV} = \frac{{\sigma^2}_A * {\sigma^2}_V}{{\sigma^2}_A + {\sigma^2}_V}$$

Hence, the reliability of the audiovisual estimate is greater than (or equal to) the maximal reliabilities of the unisensory estimates. Equation (7) shows formally that multisensory integration increases the precision of the percept. The maximal multisensory variance reduction by a factor of 2 can be obtained when the variances of the two sensory signals are equal.

In summary, the maximum likelihood estimation (MLE) model under forced-fusion assumptions makes two critical predictions for human multisensory perception performance. First, the variance associated with the multisensory percept is smaller than (or equal to) the minimal variance of the unisensory percepts (Eq. 7). Second, the multisensory percept is obtained by integrating sensory inputs weighted by their relative reliabilities (Eq. 6).

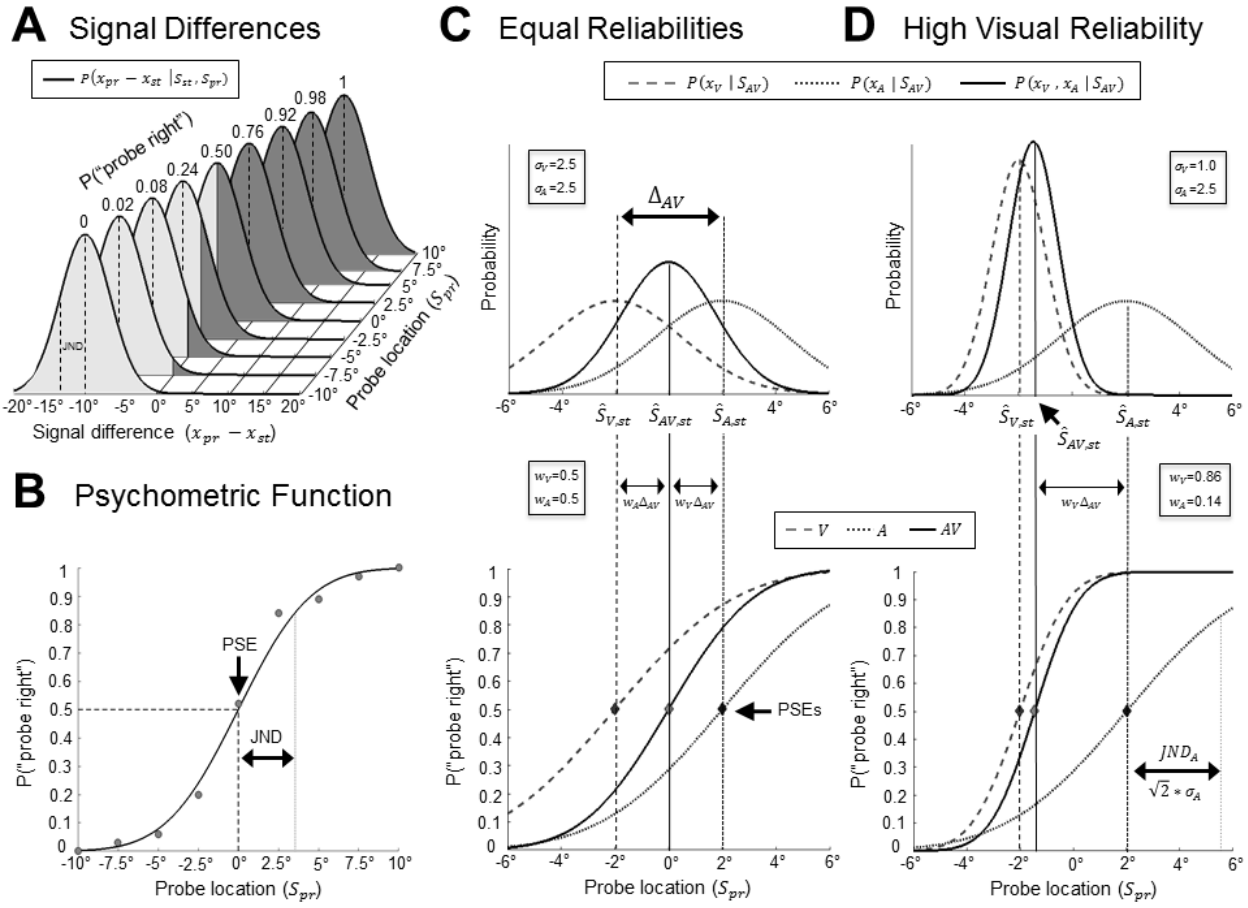


Fig 2 - Forced fusion model, maximum likelihood estimation and psychometric perturbation analysis.

Fig 2 a. Signal detection theoretic analysis of a 2IFC spatial discrimination task. For each true probe stimulus location S_{pr} (and standard stimulus location S_{st} at 0°), the observer computes a spatial estimate of the probe signal (x_{pr}) relative to the standard signal (x_{st}): i.e. the spatial signal difference $x_{pr} - x_{st}$. Because of trial-specific external and internal noise affecting both standard and probe stimuli, the signal difference is assumed to vary from trial to trial for identical true stimuli locations, S_{pr} and S_{st}, according to a Gaussian probability distribution with a standard deviation of $\sqrt{2} * \sigma_{sensory}$ that defines the summed sensory noise of the standard and probe

stimuli. The observer provides a "probe right" discrimination response when the spatial signal difference is greater than zero degrees visual angle (i.e. $x_{pr} - x_{st} > 0°$).

Fig 2 b. Psychometric function. For the data of panel A, a cumulative Gaussian shows the probability (or fraction of trials; grey circles, including measurement noise) of "probe right" responses as a function of the true probe location S_{pr}. The probability "probe right" (in B) corresponds directly to the integral (i.e. dark shaded area in A) of the Gaussian probability distribution (in A) where $x_{pr} - x_{st} > 0°$. The point of subjective equality (PSE) refers to the probe location associated with $P("probe\ right") = 0.5$. The just noticeable difference (JND) refers to the difference in probe stimulus locations at the two thresholds: $P("probe\ right") = 0.5$ and $P("probe\ right") \approx 0.84$. In a 2IFC task the JND (in B) is equal to the standard deviation of the Gaussian probability distribution of signal differences (in A): $i.e. JND = \sqrt{2} * \sigma_{sensory}$.

Fig 2 c-d. Maximum likelihood estimation (MLE) under forced fusion assumptions: The observer is presented with an audiovisual conflict stimulus (Δ_{AV}) , i.e. the visual signal is presented at $-\frac{1}{2}\Delta_{AV}$ and the auditory signal is presented at $+\frac{1}{2}\Delta_{AV}$, as the standard in the first interval, and an audiovisual congruent stimulus as the probe in the second interval. The Gaussians (top) show the likelihood functions and unbiased spatial estimates (i.e. maximum likelihood estimates; vertical lines) from the standard stimulus separately for the visual signal ($x_V = S_{V,st} = -\frac{1}{2}\Delta_{AV}$, dashed), the auditory signal ($x_A = S_{A,st} = +\frac{1}{2}\Delta_{AV}$, dotted), and the combined audiovisual signal as obtained from MLE-based integration (Eqs. (6-7), solid). The means of the Gaussian likelihood functions for the audiovisual conflict stimuli (top) can be estimated as the PSEs of the

cumulative Gaussians (bottom) obtained from auditory, visual and audiovisual 2IFC trials where the audiovisual spatial conflict stimulus is presented as the standard stimulus (i.e. see above $S_{st} = \pm\frac{1}{2}\Delta_{AV}$) and the probe stimulus is presented at variable degrees of visual angle.

Fig 2 c. For equal visual and auditory reliabilities, the means of the Gaussian likelihood functions and the PSEs of the corresponding cumulative Gaussian psychometric functions are equal to the average of the auditory and visual means or PSEs.

Fig 2 d. If the visual reliability is greater (i.e. visual variance is smaller) than the auditory one, the visual signal is assigned a greater weight. As a result, the mean of the audiovisual estimate is closer to the visual than the auditory estimate. As shown in the figure, we can estimate the sensory weights from the PSEs of the psychometric functions of the unisensory visual, unisensory auditory and audiovisual conflict conditions in a 2IFC task. Adapted from Ernst and Banks (2002).

In the following we will describe the standard psychophysical approach (Ernst & Banks, 2002; Rohde, van Dam & Ernst, 2016) that allows us to test whether human behavior is in accordance with these two MLE predictions. The main steps for testing each of the two MLE predictions involve (i) estimating the unisensory variances from perceptual performance on unisensory trials, (ii) using Eqs. (6) & (7) to make parameter-free MLE predictions about the multisensory variance and the sensory weights applied during multisensory integration and (iii) comparing these predictions with the multisensory variances and weights empirically measured during multisensory perceptual

performance. We will use an audiovisual spatial discrimination task as an example (Alais & Burr, 2004).

To investigate whether audiovisual integration of spatial inputs leads to the MLE-predicted variance reduction, we need to measure the variances associated with auditory, visual and audiovisual percepts. The empirical variances for these percepts (e.g. spatial estimates) can be estimated from participants' responses in a two-interval forced choice (2IFC) paradigm. On each trial, the observer is presented with a standard stimulus in the first interval at zero degrees ($S_{st} = 0°$) and a probe stimulus in the second interval at variable degrees of visual angle along the azimuth (S_{pr}). Standard and probe stimuli are both presented in the visual, auditory or audiovisual modalities. The observer discriminates whether the probe stimulus is on the left or right side of the standard. Next, we fit psychometric functions, i.e. a cumulative Gaussian (ψ), to the percentage 'perceived right' responses as a function of the visual angle of the presented probe separately for the visual, auditory and audiovisual conditions (e.g. using maximum likelihood estimation for fitting (Kingdom & Prins, 2016); see Figure 2A-B).

(8) $$\psi(S_{pr}) = \frac{\beta}{\sqrt{2\pi}} \int_{-\infty}^{S_{pr}} exp(-\frac{\beta^2(S_{pr}-\alpha)^2}{2})$$

where α is the point of subjective equality (PSE), i.e. the probe location where the psychometric function equals 0.5 and it is equally likely for the observer to perceive the probe left or right of the standard. Further, the just noticeable difference (JND), i.e. the difference in probe locations between the PSE and the point where the psychometric function equals ~0.84, is given by $\frac{1}{\beta}$. Importantly, as shown in Figure 2A-B, the PSE and JND obtained from the psychometric function as a cumulative Gaussian correspond

directly to the mean (μ) and standard deviation (σ) of the Gaussian distribution that describes the perceptual noise for the auditory, visual or audiovisual spatial estimates (Acuna, Berniker, Fernandes & Kording, 2015). More specifically, as we used a 2IFC paradigm in which sensory noise of both standard and probe contribute equally to the signal differences ($x_{pr} - x_{st}$), we can compute the perceptual variance for the auditory, visual and audiovisual conditions from the JNDs of their psychometric functions according to $JND^2 = 2\sigma^2$. Using Eq. (7) we can then assess whether the empirically measured AV variance is in accordance with the MLE-predicted AV variance computed from the unisensory auditory and visual variances.

To investigate whether observers integrate sensory signals weighted by their relative reliabilities as predicted by MLE we use a so-called perturbation analysis (Young, Landy & Maloney, 1993). For the perturbation analysis we need to introduce a small non-noticeable conflict between the auditory and visual signals of the audiovisual standard stimulus (n.b. no audiovisual conflict is introduced for the probe stimulus). For instance, we can shift the auditory signal by $+\frac{1}{2}\Delta_{AV}$ and the visual signal by $-\frac{1}{2}\Delta_{AV}$ relative to $S_{AV,st}$ congruent ($= 0°$). If the auditory and visual signals are equally reliable and hence equally weighted in the AV spatial estimate, the perceived AV location of the conflict AV stimulus is equal to the perceived location of the corresponding congruent AV stimulus (see Figure 2C, top panel). Yet, if the visual reliability is greater than the auditory reliability, the perceived location (i.e. spatial estimate) for the AV conflict stimulus should be biased towards the true location of the visual signal (i.e. in the above case shifted towards the left; see Figure 2D, top panel) and vice versa for greater auditory reliability. The more frequently reported visual bias on the perceived sound location has

been coined the ventriloquist effect, a perceptual illusion known since ancient times. Yet, the opposite bias from audition to vision can also emerge if the visual signal is rendered less reliable (Alais & Burr, 2004). To summarize, the cross-modal bias operating from vision to audition and vice versa provides us with information about the relative sensory weights applied during multisensory integration. Formally, we can quantify the weights applied to the auditory and visual signals from the PSEs of the psychometric functions obtained from the AV conflict conditions by rewriting Eq. (6) (see Figure 2C-D, lower panels) (Ernst & Banks, 2002):

(9) $$w_{A,emp} = \frac{PSE_{\Delta AV} - S_{V,st}}{S_{A,st} - S_{V,st}} \quad \text{with} \quad w_V = 1 - w_A$$

Note that this equation implicitly assumes that unisensory auditory and visual perception are unbiased (i.e. the PSEs of the unisensory psychometric functions are equal to zero). These empirical sensory weights can then be statistically compared with the MLE-predicted weights computed from the JNDs of the unisensory psychometric functions according to Eq. (6).

Critically, measuring the sensory weight requires a difference in the location of unisensory component signals, i.e. the presentation of incongruent audiovisual signals. While a greater inter-sensory conflict may enable a more reliable estimation of sensory weights, it progressively violates the forced fusion assumption and makes it less likely that observers assume a common source for the sensory signals. As a rule of thumb, a Δ_{AV} equal to the JND of the more reliable sensory signal has been proposed to be adequate (Rohde et al., 2016).

Numerous psychophysical studies have suggested that human observers integrate two sensory signals near-optimally, i.e. as predicted by the forced fusion model outlined above. For instance, near-optimal integration has been shown for visual-tactile size estimates in a seminal study by Ernst and Banks (2002). Four participants judged, by looking and/or feeling, whether the height of a raised ridge stimulus was taller than a standard comparison height. The true height of the ridge varied with small deviations from the standard height on a trial-by-trial basis. The used apparatus allowed the researchers to independently decrease the visual reliability by addition of visual noise at four different levels. Psychometric functions were fit to the unisensory and bisensory responses such that MLE-predicted and empirical weights and variances could be compared (as described above). Results indicated that the visual variance increased and visual weights decreased with increasing visual noise levels (as predicted by Eq. 6). Importantly, the empirical visual weights and visual-haptic variances were similar to the MLE-predicted weights and variances for all four noise levels (with a notably clear bisensory variance reduction when the visual and haptic perceptual reliability were similar; Eq. 7); thus suggesting that visual and haptic sensory signals were integrated in (near-) optimal fashion (Ernst & Banks, 2002). A follow-up experiment by the same group, using similar stimuli and apparatus, replicated the finding of an optimal variance reduction for visual-tactile size estimates (in conditions with negligible spatial disparity between the two sensory-specific cues; Gepshtein, Burge, Ernst, & Banks, 2005). Other examples of multisensory integration for which human behavior was shown to be in line with maximum likelihood estimation include audiovisual location estimates (Alais & Burr, 2004), audiovisual frequency discrimination (Raposo, Sheppard, Schrater & Churchland, 2012; Sheppard, Raposo & Churchland, 2013), visual-tactile object-shape

judgments (Helbig & Ernst, 2007), audiovisual duration estimates (Hartcher-O'Brien, Di Luca & Ernst, 2014), and audiovisual motion-speed discrimination (Mendonça, Santos & Lopez-Moliner, 2011).

At the neural level, neurophysiological studies in non-human primates have shown that neural populations (Fetsch, Pouget, DeAngelis & Angelaki, 2011) and single neurons (Gu, Angelaki & Deangelis, 2008; Morgan, Deangelis & Angelaki, 2008) integrate sensory signals weighted by their reliabilities in line with MLE predictions in visual-vestibular motion discrimination tasks. Further, Fetsch, Pouget, DeAngelis and Angelaki (2011) showed that the variances and sensory weights obtained from decoding spiking rates in a population of multisensory neurons were qualitatively comparable to the variances and weights observed at the behavioral level. At a more implementational level, these authors have proposed the divisive normalization model (Ohshiro, Angelaki & DeAngelis, 2011, 2017). This normalization model mediates reliability-weighted sensory integration, because the activity of each neuron is normalized by the activity of the entire pool of neurons.

Additional evidence in support of reliability-weighted multisensory integration at the neural level comes from several human fMRI studies showing that the connectivity between unisensory regions and association regions such as the superior temporal sulcus depends on the relative audiovisual reliabilities in speech recognition tasks (Beauchamp, Pasalar & Ro, 2010; Nath & Beauchamp, 2011). Likewise, the blood oxygenation level-dependent (BOLD) response induced by somatosensory inputs in parietal areas was modulated by the reliability of concurrent visual input during a visuohaptic size discrimination task (Helbig et al., 2012).

Despite considerable evidence in support of MLE-optimal integration in human and non-human primates, accumulating research has also revealed situations where the sensory weights and reduction in multisensory variance are not fully consistent with the predictions of maximum likelihood estimation. These findings highlight assumptions and limitations of the standard MLE forced fusion model for multisensory perception.

Focusing on the sensory weights, numerous studies have shown that human observers overweight a particular sensory modality in a range of tasks. Most prominently, in contrast to the classical study by Alais and Burr (2004) showing MLE-optimal auditory and visual weights in spatial localization, Battaglia, Jacobs and Aslin (2003) reported that observers rely more strongly on visual than auditory signals for spatial localization. Likewise, a series of studies have shown auditory overweighting in audiovisual temporal judgment tasks (Burr, Banks & Morrone, 2009; Maiworm & Röder, 2011), vestibular overweighting in visual-vestibular self-motion tasks (Butler, Smith, Campos & Bülthoff, 2010; Fetsch, Turner, DeAngelis & Angelaki, 2009), visual overweighting in a visual-vestibular self-rotation task (Prsa, Gale & Blanke, 2012), and haptic overweighting in a visual-haptic slant discrimination task (Rosas, Wagemans, Ernst & Wichmann, 2005). In all of those studies the sensory modality that is overweighted was the modality that is usually more reliable for this particular task in everyday experiences. One may therefore argue that the brain adjusts the weights of the sensory inputs not only based on the input's current reliability but also imposes a modality-specific reliability prior that reflects the modality's reliability for a particular property or task in everyday life (Battaglia et al., 2003; Maiworm & Röder, 2011).

With respect to the second MLE prediction of multisensory variance reduction, numerous studies, covering a variety of sensory modalities and tasks, have also shown a decrease in multisensory variance that is smaller than predicted by the forced fusion model (Eq. 7). For example, this was shown for audiovisual interval duration judgments (Burr et al., 2009), audiovisual speed discrimination (Bentvelzen, Leung & Alais, 2009), visual-haptic slant discrimination (Rosas et al., 2005), and visual-haptic size and depth estimation tasks (Battaglia, Kersten & Schrater, 2011; Gepshtein & Banks, 2003). This 'sub-optimal' integration performance can be explained by several key assumptions of the forced fusion model that may not hold in our natural environment. First, the forced fusion model assumes that two signals are necessarily generated by one single source. However, in the real world sensory signals can be generated by common or independent sources, leading to uncertainty about the world's causal structure (see next section). Likewise, in some experimental settings the observer may take into account this causal uncertainty, in particular if conflict trials are included or artificial stimuli are used that do not enhance the observer's forced fusion or common source assumptions (Bentvelzen et al., 2009; Gepshtein & Banks, 2003). Second, the MLE model assumes that the sensory noise is independent between sensory modalities (Oruç et al., 2003). This assumption may be violated in some multisensory estimation tasks where dependencies exist between sensory modalities as a result of cross-modal adaptive calibration (e.g. auditory spatial estimates can be recalibrated by synchronous visual signals through a process that is different from multisensory integration; Ernst, 2012; Gepshtein & Banks, 2003; Gori, Sciutti, Burr & Sandini, 2011; Jacobs, 2002; Wozny & Shams, 2011). Third, the MLE model does not include additional sources of noise that may be added after integration,

e.g. during decision making and response selection (Battaglia et al., 2011; Burr et al., 2009).

4. Causal Inference: Accounting for observers' uncertainty about the world's causal structure

The forced fusion model presented in the previous section accommodates only the special case of where two signals come from a common source. As a result, it can only model that two signals are integrated in a mandatory fashion. Yet, in our natural environment our senses are bombarded with many different signals. In this more naturalistic scenario an observer should bind signals into one coherent and unified percept only when they come from a common source, but he needs to treat them separately when they come from independent sources. Critically, the observer does not know the causal structure underlying the sensory signals. Instead, he needs to infer whether signals come from common or independent sources from the signals themselves. A range of correspondence cues such as temporal coincidence and correlations, spatial co-location and higher-order cues such as semantic, phonological, metaphoric, etc. correspondences (Adam & Noppeney, 2010; Bishop & Miller, 2011; Kanaya & Yokosawa, 2011; Lee & Noppeney, 2011, 2014; Maier, Di Luca & Noppeney, 2011; Noppeney, Josephs, Hocking, Price & Friston, 2008; Parise & Spence, 2009; Parise, Spence & Ernst, 2012; Soto-Faraco & Alsius, 2009; Stevenson, Fister, Barnett, Nidiffer & Wallace, 2012; van Wassenhove, Grant & Poeppel, 2007; Warren, Welch & McCarthy, 1981) are critical cues informing observers about whether signals come from a common source and should thus be integrated. Hence, multisensory perception in our natural

environment relies on solving the so-called causal inference problem (Shams & Beierholm, 2010). It requires observers not only to deal with uncertainty about perceptual estimates, but also with causal uncertainty, i.e. their uncertainty about the world's causal structure.

Spatial ventriloquism is a prominent audiovisual perceptual illusion that illustrates not only reliability-weighted integration (see Section 3), but also how the brain arbitrates between integration and segregation in the face of uncertainty about the causal structure of the world. At small spatial disparities, the perceived location of an auditory event (e.g. the voice of a puppeteer) shifts towards the location of a temporally correlated but spatially displaced visual event (e.g. the facial movements of the puppet) and vice versa depending on the relative auditory and visual reliabilities as described in the forced fusion section (Alais & Burr, 2004). This spatial biasing (i.e. the ventriloquist effect) breaks down or is at least attenuated at large spatial disparities and audiovisual asynchronies when it is unlikely that auditory and visual signals are caused by a common source (Hairston et al., 2003; Lewald & Guski, 2003; Odegaard, Wozny & Shams, 2017; Slutsky & Recanzone, 2001; Wallace et al., 2004).

Initial modelling approaches introduced coupling priors to allow signals from different senses to bias each other without being integrated into one single unified percept (Bresciani, Dammeier & Ernst, 2006; Roach, Heron & McGraw, 2006). More recently, Körding, Beierholm, Ma et al. (2007) (and simultaneously Sato, Toyoizumi and Aihara, 2007) proposed a Bayesian causal inference model that explicitly models the potential causal structures (i.e. common source or independent sources) that could have

generated the sensory signals. Figure 1C shows the generative model for Bayesian causal inference in an audiovisual spatial ventriloquist paradigm and localization task.

The generative model of Bayesian causal inference assumes that common ($C = 1$) or independent ($C = 2$) sources are determined by sampling from a binomial distribution with $P(C = 1)$ equal to the common-source prior P_{common}. The common source prior thus quantifies the observers' 'unity assumption' (Chen & Spence, 2017) or prior tendency to integrate signals from different sensory modalities into one unified percept.

For a common source, the "true" location S_{AV} is drawn from the spatial prior distribution $N(\mu_{prior}, {\sigma^2}_{prior})$. For two independent causes, the "true" auditory (S_A) and visual (S_V) locations are drawn independently from this spatial prior distribution. The spatial prior distribution models an observer's prior expectations of where events may happen (see Section 2). For instance, we can model a central bias or expectation that events happen in the centre of the visual field (Kerzel, 2002; Odegaard et al., 2015) by setting $\mu_{prior} = 0°$ and adjusting its strength in terms of the variance ${\sigma^2}_{prior}$.

Finally, exactly as in the unisensory and the forced fusion cases, noise is introduced independently for each sensory modality by drawing the sensory inputs x_A and x_V independently from normal distributions centered on the true auditory (or visual) locations with parameter σ_A (or σ_V). Thus, σ_A and σ_V define the noise (i.e. reliability) of the inputs in each sensory modality.

In total, the generative model includes the following free parameters: the common-source prior P_{common}, the spatial prior standard deviation σ_{prior}, the auditory standard deviation σ_A, and the visual standard deviation σ_V.

Given this probabilistic generative model, the observer needs to infer the causal structure that has generated the sensory inputs (i.e. common source or causal judgment) and the location of the auditory and/or visual inputs (i.e. spatial localization task). Critically, as we will see below, an observer's spatial estimates inherently depend on his strategy of how to deal with his uncertainty about the underlying causal structure. In other words, the observer's implicit causal inference co-determines his spatial estimate during a localization task.

The posterior probability of the underlying causal structure can be inferred by combining the common-source prior with the sensory evidence according to Bayes' rule (Körding, Beierholm, Ma et al., 2007):

(10) $$P(C = 1|x_A, x_V) = \frac{P(x_A,x_V|C=1)*P_{common}}{P(x_A,x_V)}$$

In explicit causal inference tasks (e.g. common source or congruency judgments), observers may thus report common or independent sources by applying a fixed threshold (e.g. $CI_{Th} = 0.5$) to the posterior probability of a common source:

(11) $$\hat{C} = \begin{cases} 1 & if \quad P(C = 1|x_A, x_V) \geq CI_{Th} \\ 2 & if \quad P(C = 1|x_A, x_V) < CI_{Th} \end{cases}$$

As expected and shown in Figure 3A, the posterior probability for a common source decreases with increasing spatial disparity between the auditory and visual signals. Indeed, numerous studies have demonstrated that participants are less likely to perceive signals as coming from a common source for large inter-sensory conflicts such as audiovisual spatial disparity or temporal asynchrony (Bosen et al., 2016; Hairston et al., 2003; Lewald & Guski, 2003; Rohe & Noppeney, 2015a; Slutsky & Recanzone, 2001;

Soto-Faraco & Alsius, 2009; Stevenson et al., 2012; van Wassenhove et al., 2007; Wallace et al., 2004).

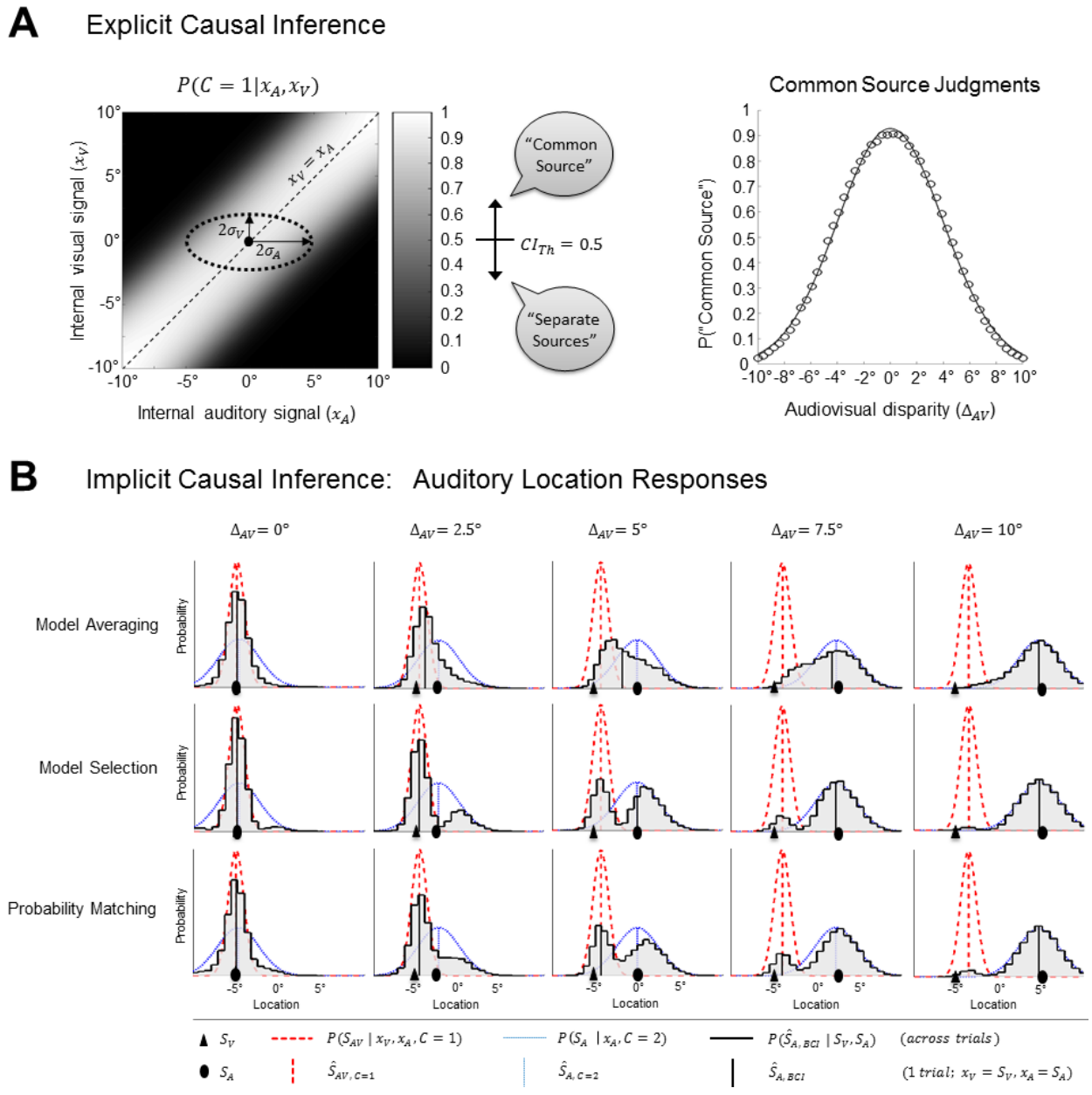


Fig 3 - Explicit and implicit Bayesian causal inference.

Fig 3 a. Explicit causal inference. The posterior probability of a common source $P(C = 1|x_A, x_V)$ is shown as a function of the internal auditory and visual signals (x_A and x_V). It decreases for increasing spatial disparities between the internal audiovisual

signals. The observer is assumed to report a common source if the posterior probability for a common source is greater than a threshold CI_{th} (e.g. if $P(C = 1|x_A, x_V) > 0.5$). Critically, even if the true auditory and visual source locations are identical (i.e. $S_A = S_V$) the internal visual and auditory signals can differ because of internal and external noise (e.g. the area circumscribed by the dashed black circle covers 95% of the bivariate Gaussian probability distribution $P(x_A, x_V|S_A = 0°, S_V = 0°)$. Right panel: Probability of a common source judgment (across trials) as a function of spatial disparity Δ_{AV} between the auditory and visual sources (S_A and S_V) as predicted by the Bayesian causal inference model (see text).

Fig 3 b. Implicit causal inference – Auditory location responses: Simulated auditory location responses as a function of audiovisual spatial disparity (Δ_{AV}, columns 1 to 5) according to Bayesian causal inference for the three decision functions: model averaging (top row), model selection (middle row) and probability matching (bottom row). The black triangles indicate the true visual source location S_V and the black disks the true auditory source location S_A. For one trial per panel with $x_A = S_A$ and $x_V = S_V$: The dashed lines show the audiovisual posterior probability distributions $P(S_{AV}|x_A, x_V, C = 1)$ and audiovisual spatial estimates $\widehat{S}_{AV,C=1}$ (i.e. maximum a posteriori estimates; vertical lines) for the forced fusion model component. The dotted lines show the auditory posterior probability distributions $P(S_A|x_A, C = 2)$ and auditory spatial estimates $\widehat{S}_{A,C=2}$ for the full segregation model component. Finally, the vertical solid lines indicate the Bayesian causal inference estimate $\widehat{S}_{A,BCI}$. The solid lines delineating the grey shaded area define the probability distributions (i.e. normalized histograms) of the Bayesian causal inference estimates across many trials

$P(\widehat{S}_{A,BCI}|S_A, S_V)$. The distributions were generated from 10,000 randomly sampled x_A, x_V for each combination of S_A, S_V, with the parameters for visual noise: $\sigma_V = 1°$, auditory noise: $\sigma_A = 2.5°$, central spatial prior distribution: $\mu_{prior} = 0°$ and $\sigma_{prior} = 10°$, and common source prior: $P_{common} = 0.5$ (n.b. the same parameter values were used in panel A). Adapted from Wozny, Beierholm and Shams (2010).

Critically, the estimate of the auditory and visual source location needs to be formed depending on the underlying causal structure: In the case of a known common source ($C = 1$), the optimal estimate of the audiovisual location is a reliability-weighted average of the auditory and visual percepts and the spatial prior (i.e. this is the forced fusion estimate of Section 3, Eq. (6), with addition of the spatial prior):

$$\hat{S}_{AV,C=1} = \frac{\frac{x_A}{\sigma^2{}_A} + \frac{x_V}{\sigma^2{}_V} + \frac{\mu_{prior}}{\sigma^2{}_{prior}}}{\frac{1}{\sigma^2{}_A} + \frac{1}{\sigma^2{}_V} + \frac{1}{\sigma^2{}_{prior}}} \quad (12)$$

In the case of known independent sources ($C = 2$), the optimal estimates of the auditory and visual signal locations (for the auditory and visual location report, respectively) are independent from each other (i.e. the so-called full segregation estimates).

$$\hat{S}_{A,C=2} = \frac{\frac{x_A}{\sigma^2{}_A} + \frac{\mu_{prior}}{\sigma^2{}_{prior}}}{\frac{1}{\sigma^2{}_A} + \frac{1}{\sigma^2{}_{prior}}} \quad \text{and} \quad \hat{S}_{V,C=2} = \frac{\frac{x_V}{\sigma^2{}_V} + \frac{\mu_{prior}}{\sigma^2{}_{prior}}}{\frac{1}{\sigma^2{}_V} + \frac{1}{\sigma^2{}_{prior}}} \quad (13)$$

Critically, the observer does not know the underlying causal structure and hence needs to provide a final estimate of the auditory and visual locations that accounts for this causal uncertainty. More specifically, the observer can combine the estimates under the

two causal structures using various decision functions such as “model averaging,” “model selection,” or “probability matching” (Wozny, Beierholm & Shams, 2010), as described below.

According to the “model averaging” strategy, the observer accounts for his causal uncertainty by combining the integrated, forced fusion spatial estimate with the segregated, task-relevant unisensory spatial estimate (i.e., either auditory or visual; whichever needs to be reported) weighted in proportion to the posterior probability of the underlying causal structures. This strategy minimizes the error about the spatial estimates under the assumption of a squared loss function (Körding, Beierholm, Ma et al., 2007).

$$\hat{S}_A = P(C = 1|x_A, x_V) * \hat{S}_{AV,C=1} + \big(1 - P(C = 1|x_A, x_V)\big) * \hat{S}_{A,C=2} \tag{14}$$

$$\hat{S}_V = P(C = 1|x_A, x_V) * \hat{S}_{AV,C=1} + (1 - P(C = 1|x_A, x_V)) * \hat{S}_{V,C=2} \tag{15}$$

According to the ‘model selection’ strategy, the observer reports the auditory ($\hat{S}_A$) or visual ($\hat{S}_V$) spatial estimate selectively from the more likely causal structure. This strategy minimizes the error about the inferred causal structures, as well as the error about the spatial estimates given the inferred causal structures.

$$\hat{S}_A = \begin{cases} \hat{S}_{AV,C=1} & if \quad P(C = 1|x_A, x_V) \geq 0.5 \\ \hat{S}_{A,C=2} & if \quad P(C = 1|x_A, x_V) < 0.5 \end{cases} \tag{16}$$

$$\hat{S}_V = \begin{cases} \hat{S}_{AV,C=1} & if \quad P(C = 1|x_A, x_V) \geq 0.5 \\ \hat{S}_{V,C=2} & if \quad P(C = 1|x_A, x_V) < 0.5 \end{cases} \tag{17}$$

According to ‘probability matching’, the observer reports the spatial estimate of one causal structure stochastically selected in proportion to its posterior probability.

(18) $$\hat{S}_A = \begin{cases} \hat{S}_{AV,C=1} & if \quad P(C=1|x_A, x_V) \geq \alpha \\ \hat{S}_{A,C=2} & if \quad P(C=1|x_A, x_V) < \alpha \end{cases} \quad \text{with} \quad \alpha \sim Uniform(0,1)$$

(19) $$\hat{S}_V = \begin{cases} \hat{S}_{AV,C=1} & if \quad P(C=1|x_A, x_V) \geq \alpha \\ \hat{S}_{V,C=2} & if \quad P(C=1|x_A, x_V) < \alpha \end{cases} \quad \text{with} \quad \alpha \sim Uniform(0,1)$$

As illustrated in Figure 3B, Bayesian causal inference transitions gracefully between sensory integration and segregation as a function of inter-sensory conflict irrespective of the specific decision function. In other words, while the forced fusion model allows only for a linear combination of the sensory signals ($\hat{S}_{AV,C=1}$ in Figure 3B), Bayesian causal inference models ($\hat{S}_{A,BCI}$) combine sensory signals non-linearly as a function of inter-sensory conflict. They predominantly integrate sensory signals approximately in line with forced fusion models, when the conflict is small, but attenuate integration for large conflicts. Numerous studies since the inception of multisensory integration as a research field in its own right have provided qualitative evidence for the computational principles governing Bayesian causal inference. For instance, several studies have demonstrated an inverted U-shape function for % perceived synchronous or the McGurk effect as a function of audiovisual synchrony of speech signals (Lee & Noppeney, 2011; Maier et al., 2011; Soto-Faraco & Alsius, 2009; van Wassenhove et al., 2007).

Over the past decade, accumulating research has also quantitatively compared human behavior with the predictions of Bayesian causal inference in a range of tasks including audiovisual spatial localization (Beierholm, Quartz & Shams, 2009; Bosen et al., 2016; Körding, Beierholm, Ma et al., 2007; Natarajan, Murray, Shams & Zemel, 2009; Odegaard & Shams, 2016; Odegaard et al., 2015, 2017; Rohe & Noppeney, 2015a, 2015b, 2016, Sato et al., 2007; Wozny et al., 2010), audiovisual temporal discrimination (Magnotti, Ma & Beauchamp, 2013; McGovern, Roudaia, Newell & Roach, 2016; Odegaard & Shams,

2016), visual-vestibular heading estimation (de Winkel, Katliar & Bülthoff, 2017), audiovisual speech recognition (Magnotti & Beauchamp, 2017), audiovisual distance perception (Mendonça, Mandelli & Pulkki, 2016) and audio-visuo-tactile numerosity judgments (Wozny, Beierholm & Shams, 2008). In the following, we discuss the role of (i) reliability of the sensory inputs, (ii) common source prior and (iii) the decision function in Bayesian causal inference.

To investigate the influence of sensory reliability on how human observers arbitrate between sensory integration and segregation, Rohe and Noppeney (2015a) presented participants with auditory and visual spatial signals at multiple spatial disparities and visual reliabilities. In a dual task, observers performed Bayesian causal inference implicitly for auditory spatial localization and explicitly for common source judgment. The study showed that visual reliability shapes multisensory integration not only by determining the relative sensory weights, but also by defining the spatial integration window. As expected by Bayesian causal inference, highly reliable visual signals sensitized observers to audiovisual disparity thereby sharpening the spatial integration window.

In addition to bottom-up sensory signals, Bayesian causal inference depends on the so-called "common source prior", embodying an observer's prior expectations that two signals are caused by a common source. This raises the question whether these common source priors are hardwired in an individual, specifically for a particular task and stimulus characteristics. For instance, in a conversational setting with a single speaker, we should be more inclined to integrate his/her facial movements with the syllables s/he is uttering for improved speech comprehension. By contrast, in a busy pub where we are

bombarded with many conflicting auditory and visual speech signals, unconstrained information integration would be detrimental. In a first study, Odegaard and Shams (2016) showed that common source priors are relatively stable across time (also see Beierholm et al., 2009), yet task-specific. More specifically, they did not generalize from a spatial ventriloquism task to a double flash illusion task. Yet, in a follow-up study where they dynamically manipulated the probability of audiovisual signals being synchronous and co-located, in a ventriloquist paradigm, they demonstrated that observers dynamically adapt their common source priors to the environmental statistics (Odegaard et al., 2017). Indeed, dynamic adjustment of common source priors had also previously been shown during audiovisual speech perception (Gau & Noppeney, 2016; Nahorna, Berthommier & Schwartz, 2012, 2015).

Finally, Wozny, Beierholm and Shams (2010) investigated in a large cohort of more than 100 observers, whether observers are more likely to use model averaging, model selection or probability matching as decisional functions in Bayesian causal inference. Surprisingly, they demonstrated that human observers predominantly use probability matching in audiovisual spatial localization. While probability matching may be thought of as being sub-optimal for static environments, humans have been shown to use this strategy in a variety of cognitive tasks (e.g., reward learning; Erev & Roth, 2014; Vul, Goodman, Griffiths & Tenenbaum, 2014). The authors proposed that probability matching may be a useful strategy to explore potential causal structures in a dynamic environment. In summary, accumulating psychophysical research has shown that human perception is governed qualitatively and to some extent quantitatively by the principles

of Bayesian causal inference, raising the question of how the brain may compute Bayesian causal inference.

At the neural level, extensive neurophysiological and neuroimaging evidence has demonstrated that multisensory integration, as indexed by multisensory response enhancement or suppression relative to the unisensory responses, depends on a temporal and spatial window of integration (Meredith, Nemitz & Stein, 1987; Meredith & Stein, 1996). Spatial windows of integration may be related to neuronal receptive field properties. By contrast, temporal windows of integration may rely on computation of temporal correlations (e.g. see recent model using the Hassenstein-Reichardt detector; Parise & Ernst, 2016) and have recently been associated with brain oscillations (Cecere, Rees & Romei, 2015; Samaha & Postle, 2015; Thakur, Mukherjee, Sen & Banerjee, 2016). Models for the neural implementations of Bayesian causal inference have been proposed, but their biological plausibility needs to be shown as yet (Cuppini, Shams, Magosso & Ursino, 2017; Ma & Rahmati, 2013; Spratling, 2016; Yu, Chen, Dong & Dai, 2016).

At the neural systems level, two recent neuroimaging studies by Rohe and Noppeney (2015b, 2016) investigated how the brain accomplishes Bayesian causal inference by combining psychophysics, fMRI, Bayesian modeling and multivariate decoding. On each trial participants localized audiovisual signals that varied in spatial discrepancy and visual reliability. The studies demonstrated that the brain computes Bayesian causal inference by encoding multiple spatial estimates across the cortical hierarchy. At the bottom of the hierarchy, in auditory and visual cortical areas, location is represented on the basis that the two signals are generated by independent sources (= segregation). At the next

stage, in posterior intraparietal sulcus, location is estimated under the assumption that the two signals are from a common source (= forced fusion). It is only at the top of the hierarchy, in anterior intraparietal sulcus, that the uncertainty about whether signals are generated by common or independent sources is taken into account. As predicted by Bayesian causal inference, the final location is computed by combining the segregation and the forced fusion estimates, weighted by the posterior probabilities of common and independent sources.

5. Conclusions

Bayesian models of perceptual inference define how an observer should integrate uncertain sensory signals to provide an accurate and reliable percept of our environment. They thus set a benchmark of an ideal observer against which human perceptual performance can be compared. Forced fusion models and psychophysical studies have highlighted that human observers integrate sensory signals that come from a common source weighted approximately in proportion to their relative reliabilities. More recent models of Bayesian causal inference account for an observer's uncertainty about the world's causal structure by explicitly modelling whether sensory signals come from common or independent sources. A final Bayesian causal inference estimate is then obtained by combining the estimates under the assumptions of common or independent sources according to various decision functions. Accumulating psychophysical and neuroimaging evidence has recently suggested that human observers perform spatial localization and speech recognition tasks in line with the principles of Bayesian causal inference.

CHAPTER 3

THE LIMITS OF MAXIMUM LIKELIHOOD ESTIMATION IN EXPLAINING AUDIOVISUAL SPATIAL INTEGRATION

David Meijer, Sebastijan Veselič, Carmelo Calafiore, Uta Noppeney

CONTRIBUTIONS

David Meijer and Uta Noppeney designed the experiment and wrote the introduction and methods section for stage-1 submission of this registered report. David Meijer prepared the MATLAB scripts for stimulus presentation. Data acquisition was performed by Sebastijan Veselič and Carmelo Calafiore. David Meijer analysed the results and wrote the results and discussion sections.

ACKNOWLEDGMENTS

We thank Maddison Roberts for acquiring the pilot data.

This research was funded by ERC-2012-StG_20111109 multsens

TO BE PUBLISHED

As a registered report in CORTEX

Stage 1 in-principle accepted. Stage 2 submission in preparation

ABSTRACT

Multisensory perception is regarded as one of the most prominent examples where human behaviour conforms to the computational principles of maximum likelihood estimation (MLE). In particular, observers are thought to integrate auditory and visual spatial cues weighted in proportion to their relative sensory reliabilities into the most reliable and unbiased percept consistent with MLE. Yet, evidence to date has been inconsistent. The current pre-registered, large-scale (N=36) replication study aimed to investigate the extent to which human behavior for audiovisual localization is in line with maximum likelihood estimation. The acquired psychophysical data show that while observers were able to reduce their multisensory variance relative to the unisensory variances in accordance with MLE, they weighed the visual signals significantly stronger than was predicted when locating spatially-incongruent audiovisual stimuli. We conclude that maximum likelihood estimation does not adequately describe human multisensory integration and we discuss three potential extensions of the model for a better match with empirical data: (i) Bayesian causal inference, (ii) prior beliefs for sensory reliability estimates, and (iii) cross-modal bottom-up salience and/or top-down cognitive factors.

1. Introduction

Sensory organs provide the brain with information about the outside world. Information from different senses can be complementary (e.g. an object's shape viewed from the front but haptically explored from the back) or redundant (e.g. the object's location). For example, both visual and auditory modalities provide uncertain information about the

spatial position of a mosquito flying in a dimly lit room. In order to obtain the most reliable and unbiased estimate (i.e. an estimate that is associated with the least variance or uncertainty) an observer should integrate redundant sensory information weighted in proportion to their relative reliabilities according to Maximum Likelihood Estimation (MLE) (Ernst & Banks, 2002). Reliability weighted integration according to MLE (i.e. the 'ideal observer' model; here simply called 'MLE model') thus sets a benchmark of statistically optimal performance against which human behaviour can be compared (Ernst & Bülthoff, 2004).

In their seminal study, Alais and Burr (2004) showed that human audiovisual localization conforms to the predictions of the MLE model. In a 2-interval forced choice (2IFC) localization task, participants were presented a so-called standard stimulus in the middle in one interval and a so-called probe stimulus at various locations along the azimuth in the other interval. Standard and probe were either both auditory or visual or audiovisual. Participants indicated which of the two stimuli (standard or probe) was located more on the left. The reliability of the visual stimuli, a low contrast Gaussian blob, was manipulated by blurring (i.e. increasing its size), whereas the reliability of the auditory stimuli, short click sounds, was kept constant. By introducing a small, unnoticeable spatial conflict between the auditory and visual components of some of the audiovisual stimuli, Alais and Burr were able to determine the relative weights that participants assigned to the auditory and visual signals during audiovisual integration. As predicted by the MLE model, observers integrated auditory and visual signals in proportion to their relative reliabilities that were computed from the unisensory auditory and visual conditions (n.b. the reciprocal of response variance corresponds to

the perceived reliability). They assigned a weight to the visual signal that increased with the visual reliability. Moreover, the variance (i.e. unreliability) of the audiovisual spatial estimates was smaller than the variances of the unisensory auditory and visual spatial estimates. Again, the audiovisual variance was closely predicted by the MLE model based on the variance of unisensory percepts.

However, the conclusions of Alais and Burr (2004) are not supported in a related study of audiovisual spatial integration by Battaglia, Jacobs and Aslin (2003). In this study, participants' integrated sensory signals weighted by their reliability, yet the visual weights were significantly higher than predicted by the MLE model. In our own lab, we have recently observed similar visual overweighting during audiovisual spatial integration (here described as our pilot data, see Appendix A). Battaglia et al. (2003) have argued that visual overweighting may result from human observers imposing a prior on the sensory reliabilities based on their everyday experiences: i.e. in most situations the visual spatial signal is far more reliable than the auditory spatial signal. Such priors are not incorporated in the MLE model. Alais and Burr (2004) briefly mention in the discussion that their participants were trained extensively on the auditory localization task, which may potentially have taught participants to trust their auditory sense more leading to a stronger auditory weight. Yet, a life-long prior on the sensory modalities is just one of many possible accounts of why human behavior diverges from MLE predictions (for a recent review, see Rahnev & Denison, 2018). Most importantly, in the multisensory and wider perception literature the findings by Alais and Burr are interpreted and generally cited as evidence that human observers integrate sensory signals or cues in line with the MLE predictions. Multisensory integration according to

MLE predictions is considered a generic and fundamental mechanism of how human observers integrate information from multiple sources. Therefore, it is important to ascertain that naïve human observers indeed integrate sensory signals from vision and audition weighted in proportion to their relative sensory reliabilities as predicted by the MLE model.

In line with previous research the current study investigated whether human behaviour is consistent with predictions of the MLE model in two steps: First, we investigated whether participants integrate the auditory and visual signals in proportion to their unisensory reliabilities (i.e. we compared empirical and predicted sensory weights). Second, we investigated whether the variance reduction of the audiovisual percept is equal to the MLE predicted variance reduction. Since we found the empirical sensory weights to be significantly different from the MLE-predicted weights we conclude that audiovisual spatial integration for untrained participants is not adequately described by the MLE model.

2. Method

2.1 Maximum Likelihood Estimation model

The MLE model makes two key quantitative predictions for observers' integrated audiovisual location estimates. First, an observer should integrate the unisensory location estimates $\hat{S}_A$ and $\hat{S}_V$ weighted in proportion to their relative sensory reliabilities:

(1) $\hat{S}_{AV} = w_A\hat{S}_A + w_V\hat{S}_V$ with $w_A = \frac{r_A}{r_A+r_V} = \frac{\frac{1}{{\sigma_A}^2}}{\frac{1}{{\sigma_A}^2}+\frac{1}{{\sigma_V}^2}}$ and $w_V = \frac{r_V}{r_V+r_A} = \frac{\frac{1}{{\sigma_V}^2}}{\frac{1}{{\sigma_V}^2}+\frac{1}{{\sigma_A}^2}}$

where w_V and w_A are the sensory weights and reliability (r) is the inverse of the sensory variance (σ^2).

Second, the sensory variance of the integrated estimate is predicted to be lower than the sensory variance of either of the unisensory estimates:

(2) ${\sigma_{AV}}^2 = \frac{{\sigma_A}^2{\sigma_V}^2}{{\sigma_A}^2+{\sigma_V}^2} < \min({\sigma_A}^2, {\sigma_V}^2)$

This second equation is generally considered the more stringent test for the MLE model, as it confirms that the two unisensory signals are truly integrated on a trial-by-trial basis (i.e. the forced fusion assumption); whereas the first equation may also hold (on average) if $\hat{S}_{AV}$ is fully determined by either $\hat{S}_A$ or $\hat{S}_V$, but when the choice for either is made probabilistically in proportion to the sensory weights (i.e. 'cue switching'; Ernst & Bülthoff, 2004).

2.2. Experiment overview

This study aimed to examine whether the MLE model accurately predicts the results of untrained participants in an audiovisual localization task that was designed to be nearly identical to the study by Alais and Burr (2004). The most striking difference is that we use only one visual reliability level (but see Section 2.6.2.2.), which is individually adjusted for each participant to match his/her auditory reliability level (see Section 2.6.1.3). Matching of the unisensory reliabilities is important in order to maximize the MLE-predicted variance reduction for audiovisual stimuli relative to the most reliable

unisensory stimuli (Eq. 2, Section 2.1). This experimental choice was made to optimize the chances of arbitrating between MLE-based integration and 'cue switching'.

2.3. Sample characteristics

The primary outcome measures were two group-level one-sided paired t-tests that assessed the two key MLE predictions (Eq. 1 and Eq. 2, Section 2.1) by testing for differences between the empirically determined and MLE-predicted sensory weights and audiovisual variances (see Section 2.9). The null hypothesis stated that the MLE model describes participants' audiovisual integration adequately (i.e. in line with the findings of Alais and Burr (2004) there is no significant difference between MLE predicted and empirical weights or AV variances). Any significant ($p < 0.05$) difference between predicted and empirical weights/variances indicated that the data were not consistent with the MLE model; as previously reported by Battaglia et al. (2003). For Battaglia et al.'s average effect size (Cohen's d) of 0.58 (estimated across different stimulus reliability levels; their figure 7) an a-priori power analysis revealed that 36 participants were required to obtain high statistical power ($1-\beta = 0.96$, $\alpha = 0.05$, $d_z \geq 0.58$; as computed with G*Power 3.1; Faul, Erdfelder, Buchner & Lang, 2009; www.gpower.hhu.de). Based on this power analysis, we decided to include thirty-six participants in the final analysis and results (i.e. excluded participants were replaced until 36 complete data sets were obtained, see Section 2.11).

All participants were university students with reportedly normal hearing, (corrected to) normal vision and no history of neurological or psychiatric disorder. Participants

provided informed consent and were compensated by means of study credits or cash[1]. The study was approved by the human research review committee of the University of Birmingham (approval numbers ERN_11-0470AP4 & ERN_15-1458P[2]).

2.4. Stimuli

The visual stimulus was a greyscale circular blob with a bivariate Gaussian amplitude envelope. Its size (defined by the 2D Gaussian's standard deviation, σ_{blob}; symmetrical in all directions) was adjusted individually for each observer to equate the unisensory spatial uncertainties for visual and auditory spatial estimates (Section 2.6.1.3). Visual stimuli were presented for a duration of 16.7 milliseconds (ms) in low-contrast (20 cd/m^2 in its centre) on a darker grey background (15 cd/m^2)[3].

The auditory stimulus was a 16.7 ms burst of white noise (70 dB SPL)[4], which included a 5 ms on/off ramp. To create virtual spatial sound sources along the azimuth, the auditory signal was convolved with standardised head-related transfer functions (Gardner & Martin, 1995; http://sound.media.mit.edu/resources/KEMAR.html).

[1] The option for compensation by cash was added after stage 1 in-principle-acceptance (IPA) of the manuscript in order to recruit from a larger pool of potential participants.
[2] The second ethics code was approved after IPA and is a replacement of the first ethics code.
[3] In the stage 1 IPA version of this manuscript low contrast visual stimuli were presented on a black background (1.9 cd/m^2 on 0.12 cd/m^2). The background was changed to grey after pilot testing in order to reduce hurting participants' eyes because of the many sudden brightness changes.
[4] Sound pressure level (SPL) was increased from 60 to 70 dB after post-IPA pilot testing to ensure that all participants could easily hear the auditory stimuli.

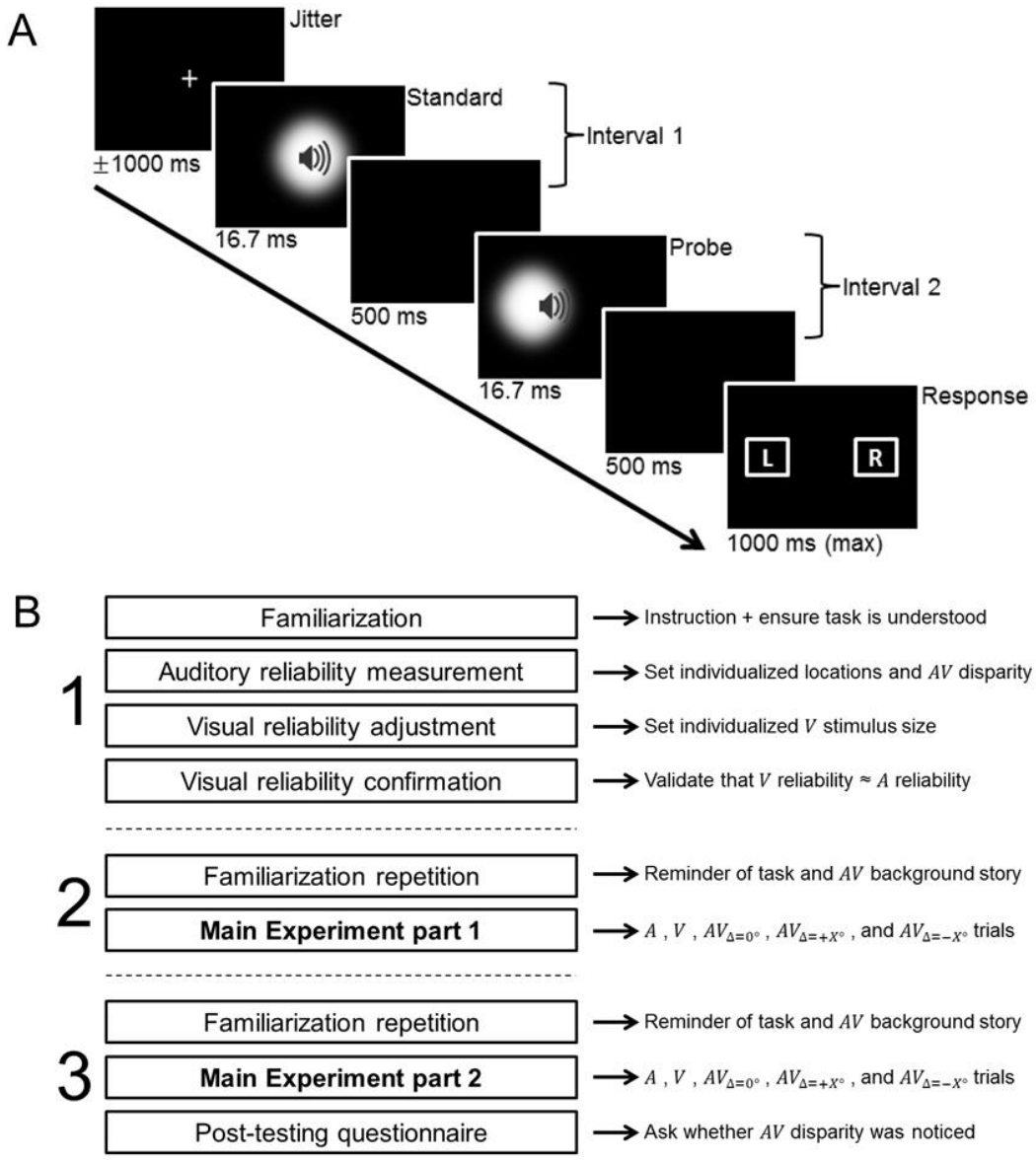


Fig 1 – Trial structure for the audiovisual localization task (Panel A) and full experimental procedure (Panel B). Fig. 1 a. A jittered pre-stimulus time period, in which participants fixated a cross in the middle, was followed by two intervals, each of which consisted of a stimulus and a subsequent blank period. The stimuli in the two intervals were either both auditory or visual or audiovisual (the latter is shown here). The first stimulus, the 'standard', was always presented in the middle. The second stimulus, the 'probe', was presented at one of thirteen locations along the azimuth. An audiovisual probe could be spatially congruent, or incongruent (with a small spatial

conflict between the auditory and visual signals; as shown here). After the second interval, two rectangles appeared on the screen to prompt participants to indicate via a two choice key press whether the location of the probe was left or right of the standard. Fig 1 b. The experiment included three sessions on three separate days (vertically numbered in the figure). In session 1 we individually (for each participant) adjusted the probe locations and AV spatial disparity (Section 2.6.1.2.) and the spatial perceptual reliability of the visual signal to match the spatial perceptual reliability of the auditory signal. The visual reliability was adjusted by changing the size of the visual stimulus (i.e. σ_{blob}; see Section 2.6.1.3.). At the end of session 1, we validated that auditory and visual reliabilities were approximately equal (Section 2.6.1.4.). In session 2 and 3, the probe locations, AV spatial disparity and visual stimulus size were set to the levels defined in session 1 and they were not further adjusted during the main experiment (but see Section 2.6.2.2.). All tasks, throughout all three sessions made use of the trial structure as described in Panel A.

2.5. Two interval forced choice paradigm

Unless otherwise stated, all tasks presented auditory, visual or audiovisual stimuli in a two-interval-forced choice (2IFC) paradigm. Fig. 1a provides a trial overview.

Participants were presented on each trial with a standard in the first interval and a probe in the second interval to avoid sequential order effects that may have affected the estimation of the slope parameters (Dyjas, Bausenhart & Ulrich, 2012). The interstimulus interval was 500 ms. Probe and standard within a trial were of the same sensory modality, i.e. both auditory (A), visual (V) or audiovisual (AV). The standard was always

presented at 0° visual angle along the azimuth, whereas the probe was presented at a location that is selected with equal probability from thirteen possible locations that were determined individually for each participant based on his/her auditory JND (see Section 2.6.1.2.), unless mentioned otherwise (see Sections 2.6.1.1 - 2.6.1.3). Critically, while the AV standard was always spatially congruent, the AV probe was either spatially congruent (i.e. $AV_{\Delta=0^\circ}$) or spatially incongruent with a small, so-called non-noticeable audiovisual spatial disparity ΔAV (i.e. $AV_{\Delta=+X^\circ}$ or $AV_{\Delta=-X^\circ}$, where the visual stimulus' location was moved by $+\frac{1}{2}\Delta AV$ and the auditory stimulus was moved by $-\frac{1}{2}\Delta AV$; n.b. the size of ΔAV was adjusted individually, see Section 2.6.1.2.). 500 ms after probe offset two rectangles were presented to prompt participants to report whether the probe was left or right of the standard. Observers indicated their response by pushing a button with their left or right index finger (maximum response time = 1 second; the prompt disappeared after a response was given). The trial onset asynchrony was jittered between 750 – 1250 milliseconds. Prior to standard onset, participants fixated a central grey cross (1° diameter) with luminance equal to the centre of the visual stimuli.

The sensory modality of the trials was blocked (A, V or AV, in pseudorandom order) and indicated to participants prior to block begin. In AV blocks, the congruent and incongruent trials were randomized.

2.6. Experimental procedure

The study consists of three 2.5 hour sessions that were performed on three separate days. In the following we will briefly describe the series of experimental parts in the first, second and third sessions; as shown in Fig. 1b.

2.6.1. First session:

2.6.1.1. (Familiarization). Brief familiarization runs were introduced at the beginning of each session to ensure that participants understood and were familiar with the task. They also mitigated learning effects and reduced variability of perceptual reliability across sessions. Participants were provided with the background story that the AV stimulus was to be considered as the result of somebody hitting the back of the screen with a metal stick (the visual blob representing the stick's imprint during the hit) in order to enhance observer's so-called 'forced fusion' assumptions that the AV's auditory and visual component signals come from a common source (Alais & Burr, 2004)[5]. Participants then completed a short familiarization series that included A, V and $AV_{\Delta=0^\circ}$ trials (In session one: 5 trials x 3 conditions x 12 locations: ±1°, ±4°, ±7°, ±10°, ±13°, ±15°; with highly reliable visual stimuli: σ_{blob} was pseudorandomized between 2° and 8°). After every response participants were given immediate corrective feedback, i.e. a green/red circle was presented on the screen to indicate a correct/incorrect response (200 ms duration).

2.6.1.2. (Auditory reliability measurement). This experimental part consisted of two parts[6]. 1. Participants first completed a series of A trials (20 trials x 13 locations: 0°, ±1°, ±2°, ±3°, ±5°, ±7°, ±10°). Participants that obtained an accuracy of less than 90% for those forty trials on which the probe was presented at ±10° azimuth were excluded at this stage (i.e. they did not participate in the main experiment). For each participant we fitted a psychometric function to the fractions of 'perceived right' across the thirteen

[5] The background story was changed after IPA to better fit the type of sounds that were presented. Originally we told participants to think of a ball being thrown at the back of the screen.
[6] This task was split into two parts after IPA and initial pilot testing. Performing the auditory reliability measurement using individualized probe locations (part 2) results in better estimates (see also main text).

probe locations (see Section 2.8). The auditory spatial uncertainty, expressed as the just noticeable difference (JND), is given by the inverse of the fitted slope parameter ($JND = \frac{1}{\beta}$). These individual auditory JNDs were used at three levels in the experiment: (i) probe locations, (ii) visual reliability and (iii) spatial disparity.

i. Probe locations: We set the probe locations for all subsequent parts of the study in a subject-specific fashion according to $locations = (0, \pm 0.5, \pm 1, \pm 1.5, \pm 2, \pm 2.5, \pm 3) * JND$ (rounded to 0.5° under the constraint that the 13 locations were unique). This procedure ensures that the psychometric functions of each participant were sampled at comparable probabilities of 'right' responses thereby providing more reliable estimates of slope, PSE and lapse rate parameters (Wichmann & Hill, 2001a).

ii. Visual reliability: We adjusted the reliability, i.e. size of the visual Gaussian blob individually for each participant to match the auditory perceptual reliability (see Section 2.6.1.3).

iii. AV spatial disparity: Previous studies have demonstrated that observers' sensitivity to detecting whether or not sensory signals come from a common source and should be integrated according to forced fusion assumptions depends on sensory reliability (Rohe and Noppeney, 2015a). Based on a power analysis simulation (see Appendix B), we set AV disparity equal to one auditory JND individually for each participant (i.e. $\Delta AV = \pm \mathrm{JND}$; conform recommendations by Rohde, van Dam & Ernst, 2016). The power analysis simulation (see Appendix B) suggested that a spatial disparity of one auditory JND allows one to reveal with high statistical power (1-β=0.95) that the empirical weight deviates from the MLE-predicted weight by approximately 0.06 or more. Yet, this limited spatial disparity also ensured that participants integrate sensory signals into one unified

audiovisual percept according to forced fusion strategies rather than take into account the causal structure of the sensory signals as accommodated by more complex Causal Inference models (see Körding, Beierholm, Ma et al., 2007; Shams & Beierholm, 2010; Rohe and Noppeney, 2015a, 2015b, 2016; and further discussions in Appendix B).

2. The second part of the 'auditory reliability measurement' is a refinement of the first part's measurement by using the individualized JND-based locations (see point i above), thereby ensuring an adequately measured auditory JND. Participants completed a second series of A trials (20 trials x 13 individualized locations). The new auditory JND that was obtained from a second fitted psychometric function replaced the JND from the first measurement. This second auditory JND was used in all further tasks (see points i, ii, and iii above).

2.6.1.3. (Visual reliability adjustment). Using adaptive staircases we adjusted the size of the Gaussian blobs (defined by σ_{blob}, Section 2.4) such that the reliability of the V and A spatial perceptual estimates were equated individually for each subject. First, we obtained observer's auditory localization performance for locations at $(\pm 0.5, \pm 0.85, \pm 1.2) * JND$ from the fitted psychometric function (see Section 2.6.1.2.). Using two unisensory visual interleaved adaptive staircases for each of these three location pairs we adjusted the size of the Gaussian blob such that the fraction 'perceived right' in the visual trials matched the target fractions estimated from the psychometric function of the auditory condition (σ_{blob} starting values: 2° and 40°; σ_{blob} decreased after each incorrect response and increased after U consecutive correct responses (U = 1, 2, 4 for the three location pairs, respectively) with up/down step sizes (Δ^+/Δ^-) weighted according to: $fraction\ correct = (\frac{\Delta^-}{\Delta^- + \Delta^+})^{\frac{1}{U}}$; Kingdom & Prins, 2016). The

adaptive staircases were terminated after 30 reversals. For each staircase σ_{blob} was computed pooled over the last 20 reversals. For each participant we identified which of the six staircases provided the estimate that was most distant from the pooled σ_{blob} across all six staircases. To attenuate effects of potential outliers, we discarded this estimate and then computed the final pooled σ_{blob} across the remaining five staircases (i.e. $\sqrt{\frac{1}{n} * \sum({\sigma_{blob}}^2)}$, with n = 5 staircases * 20 reversals)[7].

2.6.1.4. (Visual reliability confirmation). To validate that V and A variances were successfully equated, participants completed a series of 260 V trials (20 trials x 13 individualized locations) with a constant visual stimulus size (σ_{blob} as determined in Section 2.6.1.3.) and 260 V trials (20 trials x 13 individualized locations) with variable visual stimulus sizes (selected pseudo-randomly between $\frac{1}{2} * \sigma_{blob}$ and $2 * \sigma_{blob}$)[8]. The V trials with constant stimulus size were presented interleaved with the variably sized V trials. Importantly, the variably sized V trials were not analysed (i.e. trials of 'no interest') and only served to ensure similar conditions as in the main experiment (see Section 2.6.2.2.).

For each participant we fitted a psychometric function to the fractions of 'perceived right' for the V trials with constant stimulus size, across the thirteen probe locations, and the variance was computed from the fitted psychometric function (see Section 2.8). If (i) the difference between the variances obtained from this V and the previous A (Section 2.6.1.2.) psychometric functions was too large (i.e. if it led to a MLE-predicted

[7] Exclusion of the most distant staircase result was added to the protocol after post-IPA pilot tests had shown that it was fairly common for one of the six σ_{blob} values to be an outlier.
[8] The number of trials for visual reliability confirmation was reduced from 2x520 to 2x260 after IPA to reduce the overall duration of the first session.

multisensory variance reduction of less than one third of the smallest unisensory variance: ${\sigma_{AV,mle}}^2 > \frac{2}{3} * \min({\sigma_A}^2, {\sigma_V}^2)$ according to Eq. 2, Section 2.1), or if for either of the two psychometric functions (A or V) (ii) the lapse rate was greater than 0.06 (Wichmann & Hill, 2001a) or (iii) the goodness-of-fit was insufficient (see Section 2.10), then participants were considered to be unreliable with respect to their localization performance and therefore excluded (and replaced) from the study at this stage.

2.6.2. Second and third session:

2.6.2.1. (Familiarization repetition). At the beginning of session 2 and 3, participants were reminded of the background story (as described in greater detail in Section 2.6.1.1) and took part in a short familiarization run (with feedback after every trial, see Section 2.6.1.1) to minimize variability in perceptual reliability and task performance across sessions (5 trials x 3 conditions x 12 locations: $(\pm 0.5, \pm 1, \pm 1.5, \pm 2, \pm 2.5, \pm 3) * JND$ with visual reliability similar to the main experiment (Section 2.6.2.2.): σ_{blob} was pseudorandomized between $\frac{1}{2} * \sigma_{blob}$ and $2 * \sigma_{blob}$).

2.6.2.2. (Main experiment). Participants completed 520 trials (40 trials x 13 individualized locations) for each of the 5 main conditions (A, V, $AV_{\Delta=0^\circ}$, $AV_{\Delta=+X^\circ}$, $AV_{\Delta=-X^\circ}$; where X° is the individualized audiovisual disparity ΔAV, see Section 2.6.1.2.; i.e. 520 x 5 = 2600 'trials of interest') as well as an additional 520 V trials and (3 x 520 =) 1560 $AV_{\Delta=0^\circ}$ trials (i.e. 2080 trials of 'no interest'). Critically, in half of the V and AV trials (i.e. 'trials of interest') the visual stimulus size (σ_{blob}, see Section 2.6.1.3.) was constant and defined based on the results of session 1, such that visual and auditory reliabilities were equated. In the other half of the V and AV trials (i.e. 'trials of no interest') the

visual stimulus size was variable and selected pseudo-randomly between $\frac{1}{2} * \sigma_{blob}$ and $2 * \sigma_{blob}$. These latter 'trials of no interest' were not analysed. They were included to ensure that observers could not rely on a stored set of sensory weights, but needed to compute the sensory weights on a trial-by-trial basis. The AV trials of no interest were all spatially congruent.

The main experiment (4680 trials spread over two days) was divided into 20 short A, V and AV blocks. The A blocks included 26 trials, the V blocks 52 trials and the AV blocks 156 trials. The number of trials varied across sensory modalities because the A reliability level was fixed for auditory stimuli. By contrast, for half of the trials (i.e. trials of interest) the visual reliability was fixed, while it was variable for the other half of the visual trials (i.e. trials of no interest). Further, AV stimuli were presented three times as frequent as V stimuli, because AV stimuli were presented without audiovisual conflict (i.e. spatially congruent), with a positive audiovisual conflict and with a negative audiovisual conflict. The blocks of the different sensory modalities were presented in pseudorandom order and equally split across the second and third session (i.e. main experiment part 1 and part 2, see Fig. 1b). Importantly, only data from this main experiment was used to assess whether participants integrated the AV signals as predicted by MLE. Thus, A, V and AV conditions were controlled for stimulus exposure and experimental duration (n.b. the unisensory localization performances in session 1 or familiarization tasks were not used in the final analysis).

2.6.2.3. (Post-testing questionnaire). At the end of the third session participants completed a short questionnaire. Embedded in general questions about participants' subjective performance (e.g. "Did you get tired during the experiment and do you think

this affected your accuracy?" and "Rate the difficulty of the task (scale 1-10) for the three different stimuli: auditory only, visual only, and audiovisual") the following important question was asked: "For audiovisual stimuli, did you ever have the impression that the auditory and visual signals did not come from the same location?" Responses to this question served as subjective reports on whether the audiovisual spatial conflict was indeed non-noticeable (i.e. the forced-fusion assumption).

2.7. Experimental Setup

Participants were seated behind a table in a dark room with their chin on a chinrest placed at a distance of 75 cm from a grey screen (opaque fine PVC fabric; 127.5 cm width x 170cm height). The visual stimuli were back-projected onto the screen using a 60Hz DLP projector (BenQ MW529). The sounds were presented by means of headphones (Sennheiser HD 280 Pro) with a playback frequency of 192 kHz. Auditory and visual stimulus presentation was controlled using Psychtoolbox 3.0.12 (Brainard, 1997; Kleiner, Brainard & Pelli, 2007; www.psychtoolbox.org) running on MATLAB R2016a (www.mathworks.com) with maximum audiovisual asynchronies < 2 ms (100 stimulus presentations, 0.03 ms mean, 0.5 ms standard deviation).

Fixation was monitored using a desktop mount Eyelink 1000 eye tracker (www.sr-research.com) that was calibrated before the start of each block of trials[9]. Trials on which the participant failed to fixate within a 3° radius during a 1 second period prior to probe onset, or in which blinks were recorded during either of the stimuli presentations, were excluded from further analysis.

[9] The type of eye tracker was changed after IPA. The low-cost Tobii EyeX gaming eye tracker (https://tobiigaming.com) that we had initially planned to use was upgraded to the Eyelink 1000 when it became available.

2.8 Fitting psychometric functions

For each observer, we computed the fraction of 'perceived right' for each of the thirteen probe locations (on the horizontal axis x), separately for each condition. These thirteen data points per condition can be described by the psychometric function (ψ), a model with three parameters (α, β, λ):

(3) $\psi(x; a, \beta, \lambda) = \lambda + (1 - 2\lambda)F(x; \alpha, \beta)$ with

$$F(x; \alpha, \beta) = \frac{\beta}{\sqrt{2\pi}} \int_{-\infty}^{x} exp(-\frac{\beta^2(x-\alpha)^2}{2})$$

where $F(x; \alpha, \beta)$ is the cumulative normal distribution, α is the mean of the normal distribution (i.e. point of subjective equality, PSE), the so-called slope parameter β is the reciprocal of the participant's spatial uncertainty (i.e. just noticeable difference, JND), and λ is the lapse rate (i.e. the probability of an incorrect response independent of probe location x) (Kingdom & Prins, 2016). N.b. The JND (i.e. $\frac{1}{\beta}$) in this 2IFC localization task is related to the sensory variance of the stimuli (σ^2) according to: $JND^2 = 2\sigma^2$ (i.e. the sensory noise of standard and probe both contribute to the JND).

A psychometric function is 'fit' to observers' fraction of 'perceived right' responses by adjusting its parameter values (α, β, λ) such that the likelihood of the data is maximized. For this we used the Nelder-Mead optimization algorithm, as implemented in Palamedes toolbox 1.8.2 (www.palamedestoolbox.org). N.b. Likelihood (L) is computed as the following product:

(4) $L = \prod_{i=1}^{N} \bar{p}_i^{k_i} * (1 - \bar{p}_i)^{(n_i - k_i)}$

where $\bar{p}_i = \psi(x_i; a, \beta, \lambda)$ is the expected probability of observing a 'right' response given probe location x_i and parameter values for a, β and λ (Eq. 3); k_i is the empirical number of 'right' responses out of n_i trials, and N is the total number of probe locations[10].

For analysis of the main experiment, we simultaneously fitted five psychometric functions to the five different conditions: A, V, $AV_{\Delta=0^\circ}$, $AV_{\Delta=+X^\circ}$ and $AV_{\Delta=-X^\circ}$ (i.e. the product of the five likelihoods is maximized), individually for each observer. To avoid biases in the slope parameters (β) by inaccuracies of the estimated lapse rate parameters (λ) for the individual conditions, we constrained the lapse rate parameters to be equal across all five conditions (Kingdom & Prins, 2016; see specifically their Box 4.6 and Section 9.2.5). In other words, we assumed that observers' miss-responses in their left and right hemifield for non-specific reasons such as blinking, inattention etc. would be comparable across conditions. Furthermore, we assumed one common slope parameter for the three (i.e. one congruent, one positive and one negative spatial conflict) AV conditions (N.b. the MLE model predicts equality of the slopes across the AV conditions; it is therefore standard to compute a single AV variance estimate by averaging the variances of the AV conditions; e.g. see Alais & Burr, 2004). Given those parameter constraints we obtained 5 Gaussian means (i.e. $\alpha = \mathrm{PSE}$), 3 Gaussian variances (i.e. $0.5 * \left(\frac{1}{\beta}\right)^2 = \sigma^2$), 1 lapse rate parameter (λ) for each observer.

Critically, the results of these psychophysics experiments rely on participants' maintaining attention and being willing to perform this audiovisual location task in a

[10] The in-principle-accepted version of this manuscript contained a different but equivalent equation for the likelihood as a product across all trials (instead of the current product across all locations). The change enables us to use the same parameters in equation 5 for computation of the likelihood using the betabinomial model.

reliable fashion. However, it is unrealistic to expect from participants that their performance, vigilance and internal criteria remain absolutely constant over the duration of five hours of psychophysical testing (divided over session 2 and 3, see Section 2.6.2.2). In line with the most recent psychophysics literature we have therefore adopted the betabinomial model (Fründ et al., 2011; Schütt et al., 2016). For a full rationale of this decision, please see Appendix C.[11] The betabinomial model introduces one additional parameter η that provides a normalized measure (between 0 and 1) for the amount of random deviance away from the psychometric functions' predicted probabilities of 'right' responses (i.e. η increases as $\sum_{i=1}^{N}\left|\frac{k_i}{n_i}-\psi(x_i;a,\beta,\lambda)\right|$ gets larger). In order to fit the betabinomial model, including parameter η, we have used the following equation for the likelihood (i.e. instead of Eq. 4; Schütt et al., 2016):

(5) $L=\prod_{i=1}^{N}\frac{B(k_i+\eta'\bar{p}_i,\ n_i-k_i+\eta'(1-\bar{p}_i))}{B(\eta'\bar{p}_i,\ \eta'(1-\bar{p}_i))}$ with $\eta'=\frac{1}{\eta^2}-1$

where B denotes the beta function and the other parameters are the same as in Eq. 4. To clarify, only one η parameter is fitted per participant (similar to the shared lapse rate parameter λ); i.e. N denotes the number of probe locations across all five psychometric functions ($N=5*13=65$).

To ensure adequate performance and model fits, we excluded (and replaced) participants if (i) the lapse rate was greater than 0.06 (Wichmann & Hill, 2001a) or (ii) the goodness-of-fit was insufficient (see Section 2.10).

2.9 Sensory weights and AV variances

[11] For transparency reasons we note that the decision to adopt the betabinomial model was taken after all thirty-six datasets had been acquired. Please see Appendix C for a full rationale.

The normal distributions' variances of the unisensory conditions (A and V) were used to compute the MLE predictions for the auditory weight (w_A in Eq. 1, Section 2.1) and for the variance of the AV percept (${\sigma_{AV}}^2$ in Eq. 2, Section 2.1). The empirical auditory weight was computed from the audiovisual conditions with a small spatial cue conflict (i.e., $AV_{\Delta=+X^\circ}$ and $AV_{\Delta=-X^\circ}$, with X° equal to the auditory JND, see Section 2.6.1.2.):

$$(6)\ w_{A,emp} = \frac{PSE_{\Delta AV=+X^\circ} - PSE_{\Delta AV=-X^\circ}}{2*|\Delta \mathrm{AV}|} + \frac{1}{2}$$

where the PSEs serve as the means of the location estimates $\hat{S}$ (c.f. Eq. 1, Section 2.1)[12]. Please note that in consistency with previous work Eq. 6 makes the additional assumption that $PSE_{\Delta AV=0^\circ} = PSE_A = PSE_V$; i.e. that the spatial bias is equal for AV congruent, A and V conditions (Fetsch, Pouget, DeAngelis & Angelaki, 2011).

The primary outcome measures of this study were the results of statistical comparisons that investigated whether the i. empirical auditory weight and ii. empirical AV variance were significantly different from the MLE predictions (i.e. $w_{A,mle}$ vs. $w_{A,emp}$ and ${\sigma^2}_{AV,mle}$ vs. ${\sigma^2}_{AV,emp}$). To allow for generalization to the population level, empirical and MLE-predictions for each participant were entered into one-sided paired t-tests (or one-sided Wilcoxon signed-rank tests if Kolmogorov-Smirnov tests indicated non-normal distributions) at the random effects group level. The tests were one-sided because (in addition to the fact that Alais and Burr (2004) also reported one-tailed tests) given our pilot data (Appendix A) and previously published reports (Battaglia et al., 2003), we expected that any difference would have been in the following direction: $w_{A,mle} > w_{A,emp}$ and/or ${\sigma^2}_{AV,emp} > {\sigma^2}_{AV,mle}$. Further assessments were made by computing one-

[12] The in-principle-accepted version of this manuscript contained an error in the weights equation (Eq. 6). The signs of the PSEs of the incongruent conditions have been changed.

sided Bayes factors using a Jeffreys prior on variance and a Cauchy prior on positive effect sizes for the alternative hypothesis (the prior is zero for negative effect sizes, interval $c = [0, \infty]$; scaling factor $r = \sqrt{2}/2$) and a point prior on zero effect size for the null hypotheses (Rouder, Speckman, Sun, Morey & Iverson, 2009; Morey and Rouder, 2011; BayesFactor Package 0.9.12 in R 3.4.1; http://bayesfactorpcl.r-forge.r-project.org/). Bayes factor BF_{01} expresses evidence in favor of the null-hypothesis (no difference). $BF_{01} > 3$ indicates a good fit of the MLE model, whereas $BF_{01} < \frac{1}{3}$ indicates a significant difference between the empirical and MLE-predicted parameter values. Effect size index d_z was computed as: (G*Power 3.1; www.gpower.hhu.de):

$$(7)\ d_z = \frac{|\mu_x - \mu_y|}{\sqrt{{\sigma_x}^2 + {\sigma_y}^2 - 2\rho_{xy}\sigma_x\sigma_y}}$$

where μ_x, μ_y and σ_x, σ_y are the population means and standard deviations, and ρ_{xy} denotes the correlation between the two measures.

2.10 Goodness of Fit

The validity of the analysis method described above relies on the assumption that the data for each condition can be accurately fitted by a cumulative Gaussian function. In order to validate this assumption we performed a goodness of fit test. This test compares i. the likelihood of participants' responses given the model that is constrained by the cumulative Gaussian function(s) to ii. the likelihood given a so-called 'saturated' model that models observers' responses with one parameter for each stimulus location in each condition. The likelihood ratio for the original data set is then compared with a null-distribution of likelihood ratios that is generated by parametrically bootstrapping data (5000x) from the model constrained by the cumulative Gaussian distribution

(Kingdom & Prins, 2016; Wichmann & Hill, 2001a) and where additionally the expected probabilities of 'right' responses ($\bar{p}_i$) are drawn at random from beta distributions with mean $\psi(x_i; a, \beta, \lambda)$ and variance $\eta^2\psi(x_i; a, \beta, \lambda)(1 - \psi(x_i; a, \beta, \lambda))$ (Schütt et al., 2016). If fewer than 5% of the parametrically bootstrapped likelihood ratios were smaller than the likelihood ratio for the original data set (i.e. $p < 0.05$), then insufficient goodness of fit was inferred and the data set excluded (i.e. the participant was replaced). This exclusion criterion is required as parameters from psychometric functions that do not adequately fit observers' responses cannot be interpreted.

2.11 Summary of participant exclusion criteria

To ensure that our results and conclusions were based only on data sets from participants that maintain attention and provide reliable responses we have excluded participants prior to the final test session if i. their A localization performance was not adequate (accuracy < 90% for ±10° azimuth; Section 2.6.1.2), ii. the difference between unisensory auditory and visual variances was so large that the MLE predicted multisensory variance reduction was smaller than a third of smallest unisensory variance (Section 2.6.1.4), iii. the lapse rate was greater than 0.06 or the goodness-of-fit was insufficient for either of the two unisensory psychometric functions obtained during the first session: A (Section 2.6.1.2) or V (Section 2.6.1.4). It is important to emphasize that participants were excluded from the study because of the above criteria prior to the main experiment which compares participants' audiovisual integration with the MLE predictions. In addition, we have excluded participants after the main experiment in the third session, if the lapse rate was greater than 0.06 or the goodness-of-fit was

insufficient for the psychometric functions obtained during the main experiment (Section 2.9 - 2.10).

Excluded participants were replaced such that the final number of included participants was thirty-six (Section 2.3).

2.12 Summary of outcome-neutral conditions

The following criteria ensured that the data are of good quality, so that they enabled us to test the null-hypothesis that observers integrated audiovisual signals in line with MLE prediction: i. We included only participants with adequate auditory localization ability and performance (accuracy ≥ 90% for ±10° azimuth, Section 2.6.1.2). This will have excluded participants that may overweight the visual sense in AV trials because auditory localization (over an extended period of time) is too demanding. ii. We have only included participants where we adjusted V reliability individually such that A and V perceptual reliability were approximately equated (Section 2.6.1.4). This criterion ensured that flooring/ceiling effects were avoided. It rendered our experimental design powerful for revealing a robust multisensory variance reduction if participants indeed integrated audiovisual signals according to MLE predictions (Eq 2, Section 2.1) and thus allowed us to dissociate whether or not human performance is in line with MLE predictions. iii. We have only included participants with lapse rates smaller than 0.06 (Wichmann & Hill, 2001a) and adequate goodness of fit ($p > 0.05$). This criterion ensured that data sets were included only from participants that consistently maintained attention and motivation throughout the entire experiment.

2.13 Post-hoc exploratory analyses

The fitted parameters for empirical weights and variances (see Section 2.9) are only estimates of observer's true weights and variances, because any psychometric function fit is inevitably affected by experimental noise (Kingdom & Prins, 2016). In order to visualize the amount of uncertainty that is associated with each estimate we made use of the parameter estimates that were fit during the goodness-of-fit parametric bootstrap procedure (N = 5000, see Section 2.10). 95% confidence intervals were computed as the distance between the 2.5 and 97.5 percentiles of the bootstrapped parameter distributions. Furthermore, using the bootstrapped distributions, we could also test for significant differences between parameters at the level of the individual observer: This was done by comparison of the empirically determined parameter difference with a null distribution of differences (i.e. expected differences due to noise when the two parameters are actually equal) which was constructed by subtracting the empirical difference from all bootstrapped parameter differences (thus ensuring that the null distribution is approximately centred at zero). Significance was inferred when the empirical difference exceeded 95% of the null-distribution (absolute values were used for two-sided tests).

In the above-described method of computing empirical auditory weights (Eq. 6) we have made the assumption that systematic left-right biases (as expressed by the PSE) in the unisensory auditory or visual condition do not affect the PSEs of the spatially-incongruent audiovisual conditions, and any such bias for audiovisual trials is instead best captured by the PSE of the spatially congruent condition. N.b. in 2IFC tasks one does not generally expect any biases, but because of the intuitive nature of our

experimental design, in which the probe is always presented after the standard, participants may be more likely to respond to where they perceive the probe independent of where they perceive the standard, thus allowing biases to be observed. One could argue that a participant with unisensory biases in opposite directions (e.g. auditory is perceived more leftward: $PSE_A > 0°$, and visual more rightward: $PSE_V < 0°$) could experience a larger audiovisual disparity when audiovisual stimuli are presented with a spatial conflict in the same direction (e.g. A on left, V on right). The forced fusion assumption in this condition may thus be violated, leading to a reduced number of trials in which integration occurs and potentially resulting in differences with the MLE-predictions for both audiovisual variance and weights. In order to test whether this is the case, we performed a second psychometric function fit to all datasets, using the betabinomial model as described above. The only difference was that in this case five (instead of three) slope parameters were fitted, one for each condition (i.e. we fit $\sigma_{AV,emp}$ separately for each of the three audiovisual conditions). Moreover, using the parameter estimates of this second fit, we computed the auditory weights separately for the two incongruent conditions while taking unisensory biases into account:

(8) $$w_{A,ALVR} = \frac{PSE_{AV,ALVR} - (PSE_V - \frac{1}{2}X°)}{(PSE_A + \frac{1}{2}X°) - (PSE_V - \frac{1}{2}X°)} \quad \text{and} \quad w_{A,VLAR} = \frac{PSE_{AV,VLAR} - (PSE_V + \frac{1}{2}X°)}{(PSE_A - \frac{1}{2}X°) - (PSE_V + \frac{1}{2}X°)}$$

where the abbreviation ALVR is used for the condition where the audiovisual spatial conflict is imposed as Auditory Left, Visual Right; i.e. $\Delta AV = +X°$. Likewise VLAR is used for $\Delta AV = -X°$.

3. Results

3.1 Participant exclusions

Five participants were excluded during/after the first session for the following reasons: (i) Two participants did not pass the unisensory auditory performance threshold (>90% at $\pm 10°$; Section 2.6.1.2.). (ii) One participant was excluded because the difference between unisensory visual and auditory variances was too large even after they were supposedly matched using a staircase procedure (${\sigma_{AV,mle}}^2 > \frac{2}{3} * \min({\sigma_A}^2, {\sigma_V}^2)$; Section 2.6.1.4.). (iii) One participant was excluded because the unisensory visual lapse rate was too high ($\lambda = 0.11$, Section 2.6.1.4.). (iv) One participant was excluded in session 1 because the eye tracker failed to calibrate successfully even after multiple tries (for unknown reasons this participant's eyes could not be tracked at all). Furthermore, two participants chose to withdraw from the study voluntarily after successful completion of session 1. All seven participants were replaced such that thirty-six participants (26 women, 10 men; 21.8 mean age, ± 2.6 years SD) completed all three sessions. All of these datasets were included for analyses (i.e. no dataset had to be excluded because the goodness of fit was inadequate or because the lapse rate was too high in the main experiment; Section 2.11, but also see Appendix C).

3.2 Trial exclusions

Trials were excluded from analyses of the main experiment when we could determine with certainty that the participant did not fixate within a $3°$ radius around the fixation cross or when a participant blinked during either standard or probe stimulus presentation (Section 2.7). Unfortunately, for six participants the collected eye tracker data proved unreliable (it contained many sudden jumps in gaze location and occasional

time gaps in which no data was collected at all, possibly due to difficulties with the eye-tracker's focus on participants' pupils). No trials were excluded for these six participants. For the other thirty participants we excluded on average 3% (maximally 12%) of the 2600 trials of interest of the main experiment (Section 2.6.2.2).

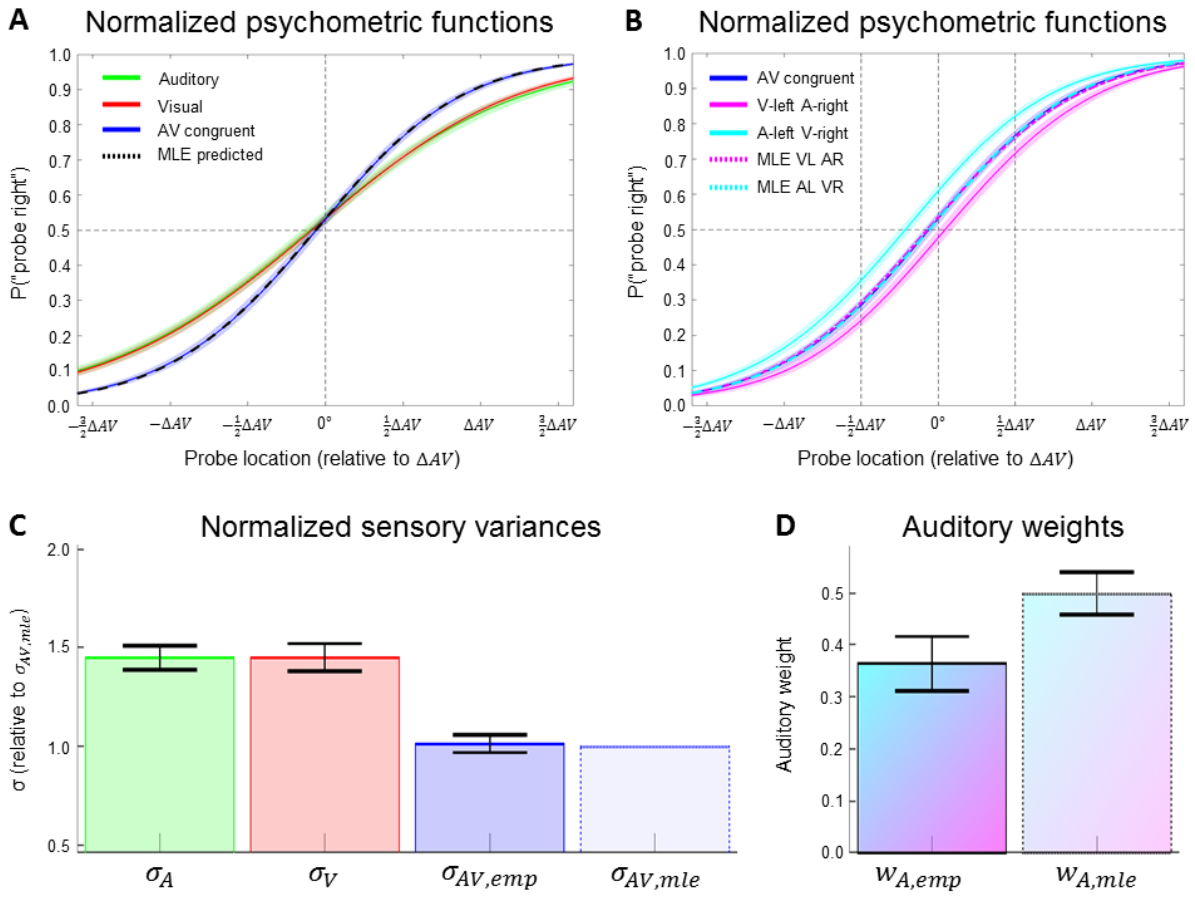


Fig 2 – Main outcomes at the group level. Fig. 2 a-b. Psychometric functions were fit to responses for A, V and AV (congruent and incongruent) conditions of each participant with each stimulus level (x-axis) expressed relative to the individual's ΔAV ($\approx JND_A$, see Section 2.6.1.2). Shown here are the group-level means (+/- SEM in shaded area) of each condition (solid lines: A = green, V = visual, $AV_{\Delta=0^\circ}$ = blue, $AV_{\Delta=-X^\circ}$ = magenta, and $AV_{\Delta=+X^\circ}$ = cyan). Similarly, using MLE-predicted parameters (see Eqs. 1-2)

hypothetical psychometric functions were constructed (dashed lines) to illustrate the MLE-predictions for $AV_{\Delta=0^\circ}$ (panel A, in black), $AV_{\Delta=-X^\circ}$ (panel B, in magenta) and $AV_{\Delta=+X^\circ}$ (cyan). The MLE-predicted psychometric functions all largely overlap with the empirical AV congruent condition. N.b. If participants were to completely ignore one of the two sensory modalities then their incongruent PSEs are expected to be near $\pm\frac{1}{2}\Delta AV$ (vertical dashed black lines). Fig. 2 c. Sensory noise parameters σ were first normalized with respect to $\sigma_{AV,mle}$ (Eq. 2) per participant before the group-level means (colored bars) and 95% confidence intervals (i.e. $\pm$1.96 SEM; black error bars) were computed. Note that $\sigma_{AV,mle} = 1$ by construction. Fig. 2 d. Group-level mean ($\pm$1.96 SEM) of empirical and MLE-predicted auditory weights (Eqs. 6 and 1).

3.3 Main outcomes at the group level

We jointly fitted (using the betabinomial model, Eq. 5) five psychometric functions to each participant's dataset of the main experiment, one for each condition: Unisensory auditory and visual, audiovisual spatially congruent and spatially incongruent (A-left V-right, and V-left A-right) (Section 2.8). Using the unisensory variances we then predicted (using MLE) the audiovisual variance and the PSEs of the incongruent audiovisual conditions (Eq. 1-2). Figure 2A-B summarizes these fits at the average group level (n.b. for illustration purposes we have normalized the x-axis with respect to $\Delta AV \approx JND_A$). It is immediately clear that the audiovisual slope (constrained to be equal for all three AV conditions) is steeper than both unisensory slopes (which seem almost perfectly matched, as we had intended; Section 2.6.1.3). The audiovisual slope is nearly identical to the MLE-predicted slope (the two curves overlap entirely). Moreover, the PSEs (i.e.

location at which $P(\text{"probe right"}) = 0.5$) for unisensory and AV congruent conditions are very similar (with a small bias for responding "probe right", i.e. PSEs < 0°, for all three conditions). While the MLE-model predicts that the PSEs of the spatially incongruent conditions also coincide with the AV congruent condition (Figure 2B), the empirical PSEs of the incongruent conditions actually modestly deviate from the MLE-predicted PSEs. Both incongruent PSEs suggest that the visual stimulus component was weighted stronger than expected based on MLE-predictions: e.g. when the visual and auditory probe were displaced by $+\frac{1}{2}\Delta AV$ and $-\frac{1}{2}\Delta AV$, respectively, this resulted in more "probe right" responses (thus a negative PSE shift; solid cyan line).

For all participants and conditions we then expressed the psychometric functions' slope parameters as the standard deviation of the sensory noise, σ (see Section 2.8), and we used Eqs. 1 & 6 to compute the MLE-predicted and empirical auditory weights. Figure 2C-D summarize these parameters of interest at the average group-level (n.b. for illustration purposes we have normalized σ with respect to $\sigma_{AV,mle}$). In support of the MLE model we find no evidence that $\sigma_{AV,emp} > \sigma_{AV,mle}$ at the group level: t(35) = 0.33, p = 0.37, BF_{01}= 4.24, d_z = 0.06. However, in contradiction to the MLE model, we do find that the auditory weights are significantly smaller than predicted, $w_{A,emp} < w_{A,mle}$: t(35) = 6.25, p < 0.0001, BF_{10} > 10000, d_z = 1.04. (Throughout this article we report results from one-sided t-tests because none of the Kolmogorov-Smirnov tests indicated that a non-parametric test was required.)

3.4 Exploratory analyses at the individuals level

While the results from the above-described group-level statistical tests are unambiguous and decisive for drawing conclusions regarding the validity of the MLE model, they do

not reveal the amount of uncertainty that is inherent with each individual's parameter estimate. We used parametric bootstrapping to obtain 95% confidence intervals for the estimates and to perform statistical tests for differences between various parameters at the individuals level (Section 2.13). Figure 3 shows the results of this exploratory analysis. Panel A illustrates the extent to which we were successful at matching the visual reliability to the auditory reliability (by adjusting the size of the visual blob; Section 2.6.1.3). Please note that the unisensory data that is depicted here is collected during the main experiment only (sessions 2-3). The reliability match has deteriorated somewhat since the first session for seven participants and with these scores they would not have passed criterion as set for the first session (i.e. ${\sigma_{AV,mle}}^2 > \frac{2}{3} * \min({\sigma_A}^2, {\sigma_V}^2)$, the dotted blue line indicates the limit). However, deviances from equality were only moderate, so sensitivity for detecting differences with MLE-model predictions should still be high. (B) Indeed, a significant multisensory variance reduction is demonstrable for the majority of our participants ($\sigma_{AV} < min(\sigma_A, \sigma_V)$, N = 24). (C) Although there are five participants for whom the one-sided bootstrap tests show that the empirical audiovisual variance is greater than what was predicted ($\sigma_{AV,emp} > \sigma_{AV,mle}$), we note that there is an equal number of participants who show a substantial deviation into the other direction. The fact that so few participants deviate from the MLE-predictions despite having optimized experimental conditions for finding such differences is strong evidence in support of near-optimal multisensory integration. However, (D) we also find that twenty participants significantly overweighted the visual stimuli during audiovisual integration ($w_{A,emp} < w_{A,mle}$), which clearly demonstrates that the group-level result for visual overweighting is not an accidental finding that can be explained by uncertainty of the parameter estimates.

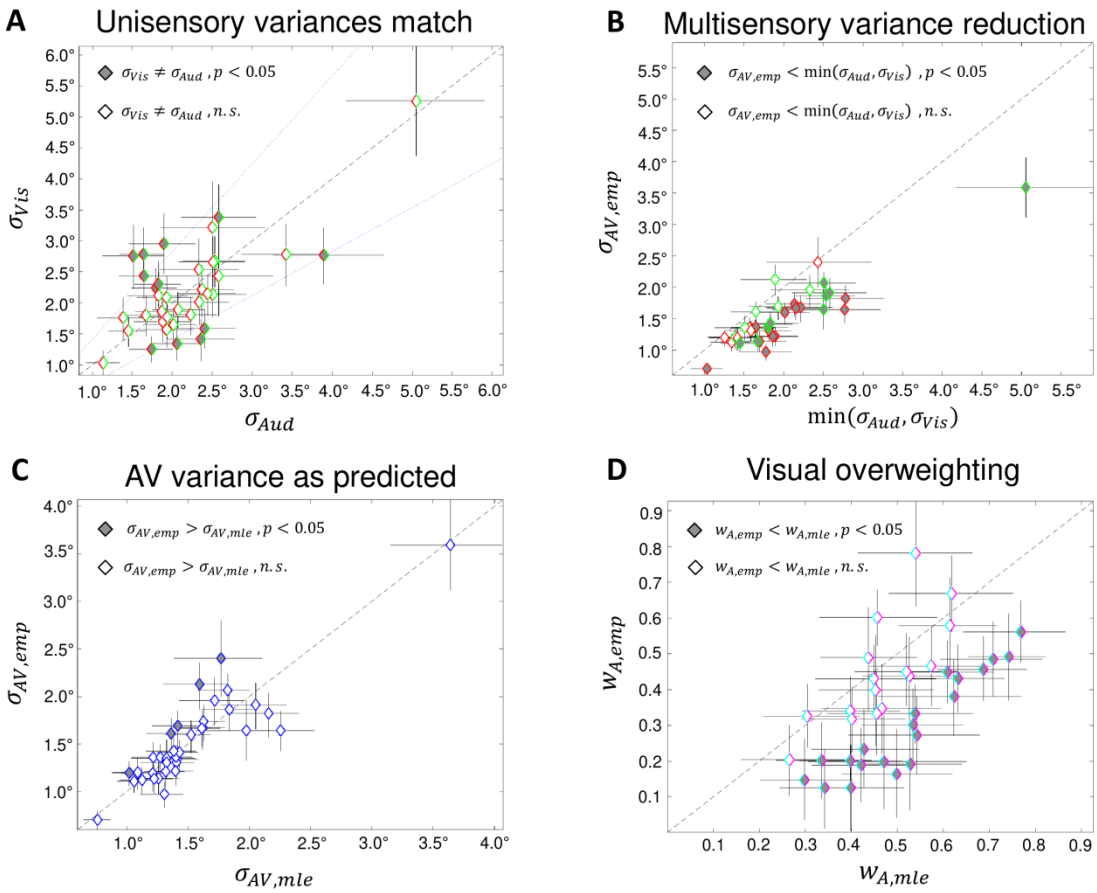


Fig 3 – Results at the individuals level. For all panels (A-D): Diamonds represent combinations of two parameter estimates for the individual participants as obtained in the main experiment (Section 2.6.2.2.). 95% confidence intervals for each parameter (Section 2.13) are illustrated by horizontal (x-axis parameter) or vertical (y-axis parameter) black lines. Black dashed lines are drawn to represent equality at x=y. Dark shaded diamonds indicate that bootstrap tests demonstrated a significant difference between the two parameters (two-sided tests in panel A, one-sided tests in panels B-D). Fig. 3 a. Most unisensory noise parameters are similar to each other, within the limits that were used as exclusion criteria in session 1 (shown here as blue dotted lines; Section 2.6.1.4.). Fig. 3 b. A significant multisensory variance reduction, relative to the most-reliable unisensory modality, is found for a majority of participants (dark

shaded; N=24). The color of the diamonds' edges indicate the most-reliable unisensory modality for each participant (A = green, V = red). Fig. 3 c. Most empirical audiovisual variances are not significantly different from the MLE-predicted audiovisual variances. Fig. 3 d. Significant visual overweighting relative to MLE-prediction is found for a majority of participants (N=20).

3.5 Control analyses for the effects of unisensory biases and audiovisual spatial disparity

The methods that we used to acquire the above-mentioned analyses results made two important assumptions: 1. that there is no meaningful difference between the slope parameters of the three audiovisual conditions (independent of spatial congruence), and 2. that any left-right bias detected in the unisensory auditory and/or visual condition is irrelevant when computing empirical auditory weights because left-right biases for audiovisual trials are best described by the PSE of the AV congruent condition. As a control analysis we tested these two assumptions by 1. fitting psychometric functions using independent slope parameters for each of three audiovisual conditions, and 2. by computing the empirical auditory weights separately for the two incongruent conditions while taking unisensory left-right bias into account (see Eq. 8, Section 2.13). Importantly, as was already suggested in Fig. 2B, the results of this control analysis show that visual overweighting is independently significant in both incongruent conditions ($w_{A,VLAR} < w_{A,mle}$: t(35) = 3.30, p = 0.001, BF_{10} = 31, d_z = 0.55 and $w_{A,ALVR} < w_{A,mle}$: t(35) = 2.12, p = 0.02, BF_{10} = 2.56, d_z = 0.35). The Bayes factors are much smaller than in the main analysis (c.f. BF_{10} > 10000, see above), because the auditory weight estimates computed in this manner are affected by random deviations of the unisensory PSEs: e.g.

if $PSE_A < 0°$, while the other PSEs are unaffected by biases, then $w_{A,ALVR}$ increases whereas $w_{A,VLAR}$ decreases (and vice versa for $PSE_A > 0°$; see Eq. 8). Indeed, the Bayes factor is greater if we compare the mean of these two empirical weights for each participant against the MLE-predicted weights: t(35) = 3.82, p = 0.0003, BF_{10} = 112, d_z = 0.63.

Regarding the fitted sensory noise parameters for the audiovisual conditions $\sigma_{AV,emp}$ we find a significant effect of condition (repeated-measurements ANOVA with 'AV condition' as within-subjects factor: F(2) = 5.97, p = 0.004) with the order of the group means $\sigma_{AV,VLAR} < \sigma_{AV,congruent} < \sigma_{AV,ALVR}$. In pairwise comparisons with $\sigma_{AV,mle}$ the only significant difference was found for $\sigma_{AV,ALVR} > \sigma_{AV,mle}$: t(35) = 2.3, p = 0.01, BF_{10} = 3.67, d_z = 0.39 (BF_{01} ≈ 10 for both $\sigma_{AV,VLAR} > \sigma_{AV,mle}$ and $\sigma_{AV,congruent} > \sigma_{AV,mle}$). This may suggest that the forced-fusion assumption has been violated in the spatially incongruent condition ALVR (but there could be other explanations too; see Discussion Section). One may hypothesize that it is more likely for participants to notice the audiovisual disparity (thus not integrate A and V) in the ALVR condition if unisensory biases exist in the same direction as the audiovisual conflict: i.e. when $PSE_A > 0°$ and/or $PSE_V < 0°$. However, contrary to that hypothesis, we found no correlation between the relative difference $\sigma_{AV,ALVR}/\sigma_{AV,mle}$ and the unisensory PSE difference $PSE_A - PSE_V$ (Pearson's r = 0.14, p = 0.42). From these control analyses we conclude that unisensory biases are unlikely to have affected the results, but we did observe some evidence suggesting that participants' behavioral results deviate from MLE-predictions not only in auditory weights but in audiovisual variance too.

3.6 Results from the post-testing questionnaire

Finally, we turned to the post-testing questionnaires to further guide our post-hoc exploratory analyses (Section 2.6.2.3). Thirteen participants reported that they occasionally experienced the auditory and visual signals as coming from two separate locations on audiovisual trials (i.e. the auditory and visual component signals of the AV probe were perceived on opposite sides of the AV standard). These responses indicate a breakdown of the forced-fusion assumption in these thirteen participants, at least on some trials. To investigate whether these thirteen participants showed a different pattern of results than the other twenty-three participants we performed a group level repeated-measurement ANOVA on the two main-outcome measures together: (i) relative sensory noise differences $\sigma_{AV,emp}/\sigma_{AV,mle}$ and (ii) weights differences $w_{A,mle} - w_{A,emp}$ (as obtained using the primary analysis pipeline; Sections 2.8-9). The factor 'σ_{AV}-or-w_A measure' was defined for within-subjects effects and a between-subjects factor created two groups based on the questionnaire responses (i.e. 13 with- and 23 participants without experience of an AV disparity). The analysis demonstrated a significant main effect of the between-subjects factor: F(1) = 5.09, p = 0.027. Both deviations from the MLE-predictions ($\sigma_{AV,emp} > \sigma_{AV,mle}$ and $w_{A,emp} < w_{A,mle}$) are exacerbated in the group of thirteen participants who claimed to have experienced occasional AV disparities; although neither of two independent one-sided two-sample t-tests for differences between the two groups reached significance by themselves (t(34) = 1.67, p = 0.052, Cohen's d = 0.58 for σ_{AV}; and t(34) = 1.49, p = 0.073, Cohen's d = 0.52 for w_A). Notably, using only the subset of thirteen datasets from participants with reported experience of AV disparity, we did find a marginally significant difference for $\sigma_{AV,emp} >$

$\sigma_{AV,mle}$ in accordance with what would be expected if the forced-fusion assumption is violated: t(12) = 1.89, p = 0.041, BF_{10} = 2.09, d_z = 0.52 (c.f. BF_{01}= 4.24 for the same test on all 36 datasets; see main outcomes above). Importantly, however, we also do still find significant visual overweighting in the subgroup of 23 participants who did not report to have experienced any audiovisual disparity (t(22) = 3.70, p = 0.0006, BF_{10} = 58.5, d_z = 0.77), thus suggesting that visual overweighting is a general mechanism that is not exclusive to observers with awareness of the experimentally induced audiovisual disparity.

4. Discussion

The aim of this study was to investigate the extent to which human behavior for audiovisual localization is in line with maximum likelihood estimation. Specifically, we have attempted to replicate the results of Alais and Burr's (2004) seminal study in which they showed evidence that human observers integrate audiovisual spatial signals according to MLE. However, utilizing carefully designed methods that were peer-reviewed and registered before data collection, we have presented evidence that naïve observers' audiovisual responses cannot be fully explained by MLE predictions. While the data show a significant variance reduction for audiovisual relative to unisensory conditions in agreement with MLE, they also unambiguously show that participants weighted the visual signals significantly stronger than is predicted by MLE-predicted reliability-weighted integration. Before providing three alternative but not mutually-exclusive explanations for these findings, we first discuss some differences with the

study of Alais and Burr (2004) and highlight the robustness of the observed visual overweighting.

The design of this study has been optimized to create conditions in which we could test for deviations from the MLE predictions with high statistical power. (i) Relative to previously published psychophysics studies on audiovisual spatial localization (Alais & Burr, 2004; Battaglia et al., 2003) we have recruited six times more participants (N=36). (ii) Experimental sensitivity for detecting deviations from MLE predictions was maximized by matching the reliability of auditory and visual stimuli for each participant. (iii) Individualized audiovisual spatial disparity sizes ensured an adequate trade-off between the risk of violating the forced-fusion assumption (small disparities are preferred) and high statistical power for potential weights differences (large disparities are preferred; Appendix B). (iv) Stimulus locations were adjusted to each individual's performance level such that the parameters of the psychometric functions could be reliably estimated based on a high number of trials at relevant stimulus levels. (v) High quality data was ensured by excluding participants that showed signs of inadequate or inconsistent performance and by using eye tracking to control for proper gaze fixation as well as removing missed trials due to blinks.

The beneficial effects of these optimized experimental conditions are best illustrated by the exploratory analyses at the level of the individuals (Figure 3). Contrary to Alais and Burr (2004), we were able to demonstrate a significant audiovisual variance reduction (i.e. multisensory behavioral benefit) in two thirds of our participants, most likely because of smaller confidence intervals for our parameter estimates (i.e. more reliable results). Such variance reductions suggest that participants based their responses on

integrated audiovisual signals rather than probabilistically responding to either of the unisensory signals (i.e. they were not 'cue-switching'; Ernst & Bülthoff, 2004). Moreover, significant visual overweighting was observed at the individuals-level in the majority of participants, thus excluding the possibility that this group-level deviation from MLE predictions was due to noisy parameter estimates. Post-hoc control analyses further consolidated that conclusion by showing that visual overweighting was independently present in both audiovisual conflict conditions and unlikely to have been influenced by unisensory biases. Finally, visual overweighting was significant at the group-level even when we constrained the analysis to a subset of participants who explicitly claimed to have had no experience of the audiovisual disparity (in a post-testing questionnaire).

However, the fact that a third of our participants reported that they had occasionally perceived the auditory and visual probe on opposing sides of the standard indicates that the forced-fusion assumption may have been violated (even though we had taken great care to avoid that by using individualized conflict sizes). It is well-known that MLE-type integration of multisensory signals breaks down as a function of their spatial disparity (Gepshtein et al., 2005). Bayesian causal inference (BCI) is able to model the extent of this break down and it has successfully been used to predict human multisensory perception in conditions of partial integration (Körding, Beierholm, Ma et al., 2007; Rohe & Noppeney, 2015a, 2015b, 2016). One particular decision function of the BCI model (so-called 'model-averaging') even allows for partial integration on single trials without participants being aware of the multisensory conflict (Wozny et al., 2010; Rohe & Noppeney, 2015a). If our participants based their responses on such partially integrated visual spatial estimates (as opposed to partially integrated auditory spatial estimates),

then this could well explain the visual overweighting that we observed. Indeed, seven participants reported (in the post-testing questionnaire) that they had based their responses on the visual signals whenever in doubt. (We do not know whether more participants used this strategy, as this is unfortunately not something that we asked for directly in the questionnaire; the seven participants reported this voluntarily to the question about experiencing spatial disparities).

An alternative explanation for visual overweighting was suggested by Battaglia et al. (2003). They had also observed visual overweighting in their audiovisual localization study (though we note that their visual stimuli, random-dot stereograms of bumps, are substantially different from the blurred blobs that were used in the current study and by Alais and Burr, 2004). Battaglia et al. (2003) hypothesized that observers may have developed a prior probability distribution to estimate a visual signal's reliability in a Bayesian way. This prior would have higher probabilities for lower visual variances, in agreement with the fact that vision normally provides more reliable spatial estimates. Using such biased estimates of the visual reliability observers would overweight visual signals when applying reliability-weighted audiovisual integration.

Importantly, both of the above explanations for visual overweighting, BCI and biased visual reliability estimates, would result in increased audiovisual variance relative to MLE-predictions. However, in proof-of-principle simulation studies we have demonstrated that it is likely for this variance increase to go unnoticed due to experimental noise; see Appendix D. By establishing that our design is more sensitive for detecting group-level weights differences than audiovisual variance differences, these

simulations thus confirm that BCI and biased visual reliability estimates form two plausible explanations for the main outcomes of the current study.

A third explanation for visual overweighting is based on the hypothesis that it is us researchers who use the wrong reliability estimate to base MLE predictions on (as opposed to assuming that observers use a biased estimate of the actual visual reliability, as above). This is the case if the visual reliability during visual-only trials is lower than the visual reliability during audiovisual trials. For example, visual reliability might be higher in an audiovisual context because of low-level cross-modal salience boosting effects (Aller, Giani, Conrad, Watanabe & Noppeney, 2015) or because of top-down cognitive factors such as sound-induced increased attentional levels (Talsma, Senkowski, Soto-Faraco & Woldorff, 2010). Such an increase of the visual reliability would lead an ideal observer to weigh the visual signals stronger during multisensory integration. Relative to MLE predictions by the researcher (based on the reliability as observed in the visual-only condition) this would manifest as visual overweighting and a small *decrease* of the audiovisual variance. Similar to the other two explanations discussed above, such a deviation of the predicted variance may have been obscured by experimental noise. (While this third hypothesis provides another plausible explanation for the main outcomes of the current study, we note that it fails to explain the results of our pilot study: there we observed an audiovisual variance that was *higher* than the most reliable unisensory variance (auditory) for all seven participants (Appendix A). Both the BCI model and a Bayesian model with a prior on visual reliability estimates would better match the pilot study's results.)

In conclusion, we state that the here presented data illustrate the limitations of MLE in explaining human multisensory perception. While MLE describes a simple, elegant, and statistically optimal model for multisensory integration that provides an intuitive understanding for the behaviorally advantageous bisensory variance reduction that can be observed under some ideal experimental conditions (including the current experiment), it has failed to quantitatively predict the sensory weights that participants used when integrating audiovisual spatial signals. As with any model, MLE presents an incomplete description of reality and it makes assumptions that can be violated. In fact, the current experiment demonstrates that it is extremely hard, if not impossible, to design laboratory experiments that would not violate these assumptions or otherwise expose the limitations of MLE (provided the design also allows for sufficient statistical sensitivity). This statement is further supported by numerous previously published reports that human multisensory perception diverges from MLE predictions (Battaglia et al., 2003; Burr, Banks & Morrone, 2009; Fetsch, Turner, DeAngelis & Angelaki, 2009; Butler, Smith, Campos & Bülthoff, 2010; Prsa, Gale & Blanke, 2012; Rosas, Wagemans, Ernst & Wichmann, 2005; Bentvelzen, Leung & Alais, 2009). We therefore emphasize great caution when claiming that MLE is a fundamental or generic mechanism for multisensory integration. At best it provides an incomplete description of human multisensory perception.

In agreement with recent recommendations (Rahnev and Denison, 2018), we believe that it is not only important to state *whether* behavioral data is in agreement with statistically optimal models, but also to investigate *why* participants deviate from the models' predictions. The post-testing questionnaire has proven to be a useful tool to

direct such exploratory analyses. Participants' responses suggested that Bayesian causal inference is a possible explanation for the current data and previously published psychophysics and neuroimaging work supports its validity in describing human multisensory perception (Körding, Beierholm, Ma et al., 2007; Rohe & Noppeney, 2015a, 2015b, 2016). However, future research will be necessary to discard/confirm alternative mechanisms such as discussed above. Other questions also remain regarding the differences between our study's results and those of Alais and Burr (2004). For example, does visual overweighting depend on whether participants are trained extensively on the auditory localization task? Furthermore, we should question why there was such strong visual overweighting in the pilot study (Appendix A) and in Battaglia et al., (2003). Might this depend on the type of visual stimulus that was used (i.e. a blurred blob, cloud of dots, random-dot stereogram)? These and other research questions will need to be addressed to better understand human deviations from statistically optimal reliability-weighted multisensory integration.

Appendix A: Pilot study

Pilot study: Method

Ten subjects participated in this pilot study. One participant was excluded because her A localisation accuracy was below 90% at 10° azimuth. Two other participants were excluded post-hoc, because they pressed random buttons in the latter half of the study's experimental blocks. As a result, only seven participants (6 female, age range 18-20, all right handed) were included in the final analysis.

The experimental paradigm was comparable to the proposed research, but differed in the following aspects: First, the visual stimulus that was used in the pilot study was a cloud of 20 dots (diameter: 0.43° visual angle) sampled pseudo-randomly from a bivariate Gaussian distribution (as in: Rohe and Noppeney, 2015a). Participants were told that the 20 dots were generated by one underlying source in the centre of the cloud. Second, the size of the visual cloud (i.e. spatial reliability) was not titrated per participant: horizontal standard deviation was $10°$, vertical standard deviation was 3°. Third, the order of standard and probe stimulus was randomised over trials. Participants reported whether the first or second stimulus was more to the left. Fourth, the following 13 fixed locations were used: 0°, ±0.5°, ±1°, ±2.5°, ±5°, ±7.5°, ±10°. Fifth, the audiovisual disparity was fixed at ±5°.

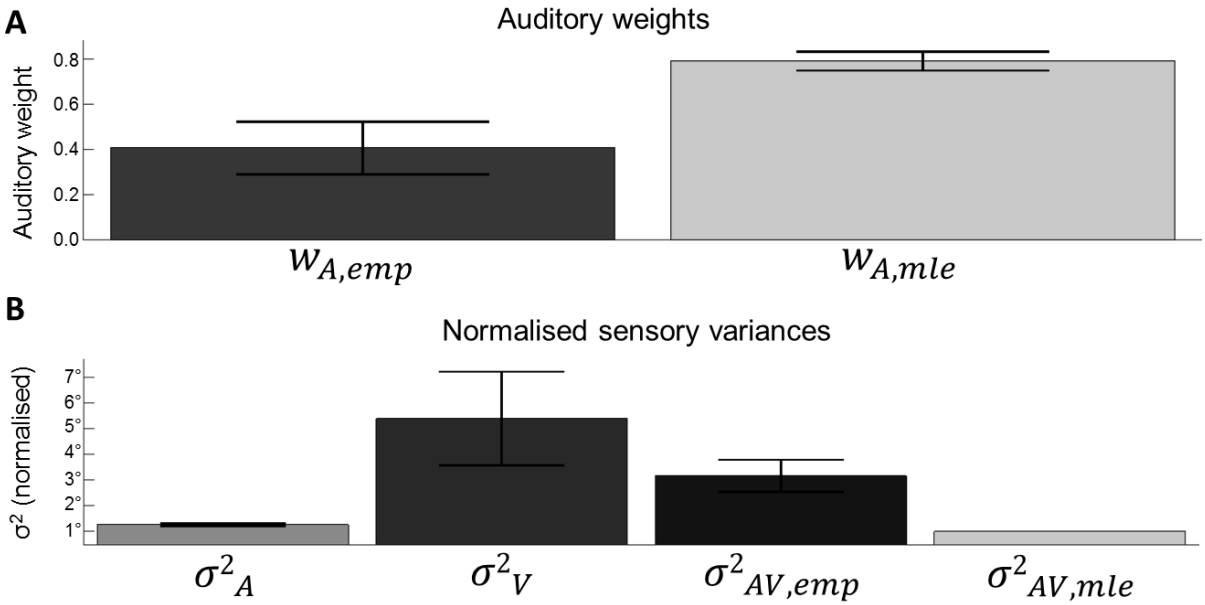


Fig A.1 - An overview of the most important pilot study results at the group level. The error bars depict the 95% confidence interval ($\pm 1.96 * SEM$). Fig A.1 a. Empirical and MLE-predicted auditory weights (w_A). Fig A.1 b. Empirical and MLE-predicted sensory

variances (σ^2). Before averaging across participants, the sensory variances of the individual participants were normalised with respect to $\sigma^2{}_{AV,mle}$ (for illustration purposes only).

Pilot study: Results

The most important results of the pilot study are shown in Fig A.1. Although the A reliability was greater than the V reliability in all participants (non-normalised group-level means: σ_V = 3.5°, σ_A = 1.77°), participants did not weight the more reliable A stimuli according to MLE predictions (Eq. 1, Section 2.1) in the AV context ($w_{A,emp}$ < $w_{A,mle}$; p < 0.05 for all participants). The group-level Wilcoxon signed rank test demonstrated a significant difference between the empirical and MLE-predicted auditory weights (p = 0.0078, d_z = 2.99). While the variance of the AV conditions decreased relative to the unisensory V condition, we observed no multisensory benefit relative to the more reliable unisensory A condition (σ_A < $\sigma_{AV,emp}$ for all participants). In line with this, the empirical variance for the AV conditions was significantly greater than predicted by the MLE model (Eq. 2, Section 2.1; p < 0.05 for all participants). The group-level Wilcoxon signed rank test demonstrated a significant difference between the empirical and MLE-predicted AV variances (p = 0.0078, d_z = 3.17).

Importantly, the goodness of fits were sufficient for all seven participants (p > 0.05) and no lapse rate was greater than 0.06. We conclude that although the empirical auditory weights are quite high for an audiovisual localisation task (mean $w_{A,emp}$ = 0.41), we did not observe the MLE-predicted and previously reported 'reverse ventriloquist effect', i.e.

a shift of the perceived AV location toward the auditory signal on cue conflict trials (Alais & Burr, 2004). Instead, during audiovisual integration, all participants overweighed their less reliable visual modality, similar to the results of Battaglia et al., (2003).

Appendix B: Selection of audiovisual disparity individually for each participant based on power analysis

Observers' sensory weights are estimated empirically by introducing a small conflict between the sensory cues– a procedure coined 'perturbation analysis' (Young, Landy & Maloney, 1993). This raises the question of how to select the conflict size (e.g. audiovisual disparity for spatial localization). On the one hand a greater conflict size is preferable, because it renders the perturbation analysis more sensitive for detecting deviations of observers' empirical weights from MLE predictions. On the other hand, greater conflict sizes may prevent participants from integrating sensory signals into one unified percept according to forced fusion assumptions. Instead, observers are then likely to compute a perceptual estimate that takes into account the uncertainty about the world's causal structure as accommodated by more complex models of causal inference (see Körding, Beierholm, Ma et al., 2007; Shams & Beierholm, 2010; Rohe and Noppeney, 2015a, 2015b, 2016). Moreover, previous research has shown that observers' sensitivity to intersensory conflicts depends on their perceptual reliability (Rohe and Noppeney, 2015a). Collectively, these considerations suggest that i. we need to determine the minimal conflict size (i.e. here: spatial disparity) that enables detections of deviations of the empirical sensory weights from MLE predictions with high statistical power (e.g. 1-β=0.95) and that ii. we need to adjust this minimal conflict

size individually for each participant based on their unisensory perceptual reliability as indexed by their auditory JND.

To determine the minimal conflict size (ΔAV in standardized JND units) that still enables detection of deviations of observers' empirical weights from MLE predictions with high statistical power, we performed the following simulations:

1. For each of the 36 participants, we initially sampled an auditory JND from a uniform distribution between 1.2 and 3.8° visual angle, i.e. a range of auditory JNDs that we usually observe for our audiovisual experimental set up and stimuli across participants. Exactly as in the current study, we set the visual JND equal to the auditory JND. The PSE of the A and V conditions were set to zero, i.e. we assumed no perceptual biases. The audiovisual JND and PSE were set to the MLE predicted values computed from the unisensory JNDs and PSEs according to Eq. 1 and Eq. 2 (Section 2.1); i.e. the PSE of the audiovisual condition was also set to zero. Hence, the 'true' empirical weights (to be computed from the AV PSEs) and the MLE-predicted weights (to be computed from the unisensory A and V JNDs) are by construction all equal to 0.5; reflecting equal perceptual reliabilities of the A and V signals. Likewise, by construction the difference between the 'true' empirical and MLE-predicted auditory weights is zero. To assess the variability (or uncertainty) of this difference between empirical and MLE-predicted auditory weight estimates, and how this variability will depend on the conflict size (i.e. spatial disparity ΔAV) we generate distributions of empirical and MLE-predicted weights as follows:

2. For each participant, we parametrically bootstrap (Palamedes toolbox 1.8.2, Kingdom & Prins, 2016) 1000 A and V data sets (with lapse rate parameter (λ) set to 0.02),

stimulus locations set according to the subject-specific auditory JND (see Section 2.6.1.2.) and 40 trials per location), and we subsequently fit psychometric functions to each simulated data set. From the fitted unisensory JNDs we compute the MLE-predicted auditory weights according to Eq. 1. This will generate a distribution of one thousand MLE-predicted auditory weights centred on 0.5, for each participant.

3. Likewise, we sample 2 x 1000 AV data sets and fit an AV psychometric function to each simulated data set. In order to evaluate the effect of spatial disparity on the precision of the estimated empirical auditory weights (and as a consequence also on the differences between empirical and MLE-predicted weights), we now arbitrarily assume that half of the AV data sets were generated by a positive conflict $AV_{\Delta=+X^\circ}$ and the other half by a negative conflict $AV_{\Delta=-X^\circ}$. Empirical auditory weights are then computed based on Eq. 6 (Section 2.9) for a range of spatial disparity sizes in participant's auditory JND units (i.e. $\Delta AV = factor * JND;$ with 'factor' logarithmically sampled in 50 steps from between 0.1 and 2). For each participant, this will generate a distribution of one thousand empirical auditory weights per spatial disparity ΔAV. By construction, the distributions of empirical weights are all centred on 0.5. Yet importantly, they vary in their spread: the larger the spatial disparity, the smaller the spread of the distribution of empirical weights (as follows directly from Eq. 6).

4. Finally, we shift the distributions of empirical auditory weights by subtracting a 'true' (i.e. to be detected) variable value (range 0 – 0.25); thereby creating empirical weights distributions that are no longer centred at 0.5 (i.e. we 'simulate' visual overweighting). Critically, while the difference in means between i. the distributions of the MLE-predicted weights and ii. the shifted distribution of empirical auditory weights is equal to

this specific value irrespective of spatial disparity, the variance of the empirical weight distribution and hence the overlap of the two distributions depends on the spatial disparity (ΔAV).

5. To enable a power analysis at the random effects group-level, we enter one bootstrapped pair of MLE-predicted and empirical weights (with specific subtracted 'true' difference and spatial disparity level ΔAV) for each participant into one-sided paired t-tests (or one-sided Wilcoxon signed-rank tests if Kolmogorov-Smirnov tests indicated non-normal distributions). These one-sided group-level paired t-tests for each level of spatial disparity and imposed 'true' difference between empirical and MLE-predicted weights are then repeated for each of the 1000 bootstraps. As a result of this procedure, we can compute the fraction of bootstraps (i.e. experiments) where the paired t-test successfully declares the empirical auditory weights for a particular 'true' difference as significantly smaller than the MLE-predicted auditory weights ($p < 0.05$). In other words, we compute the power of the statistical test separately for each combination of 'true' difference and spatial disparity.

Fig. B.1 shows the minimum 'true' difference between empirical and MLE-predicted auditory weights that can be detected with a power of ≥0.8, ≥0.9, ≥0.95, ≥0.99 as a function of spatial disparity (in JND-standardized units). The results from this power analysis suggest that for our study with 36 subjects, specific parameter choices and for a spatial disparity equal to an individual's auditory JND (i.e. in Fig B.1, ΔAV = 1 in standardized JND units) a difference in empirical and MLE-predicted auditory weights of 0.06 would be detected with a power of 0.95. Critically, while the minimal difference between empirical and MLE-predicted weights that can be detected with 0.95 power

rapidly increases for AV disparities smaller than one subject-specific auditory JND, AV disparities greater than one auditory JND have negligible impact on the power of perturbation analyses for our experimental choices.

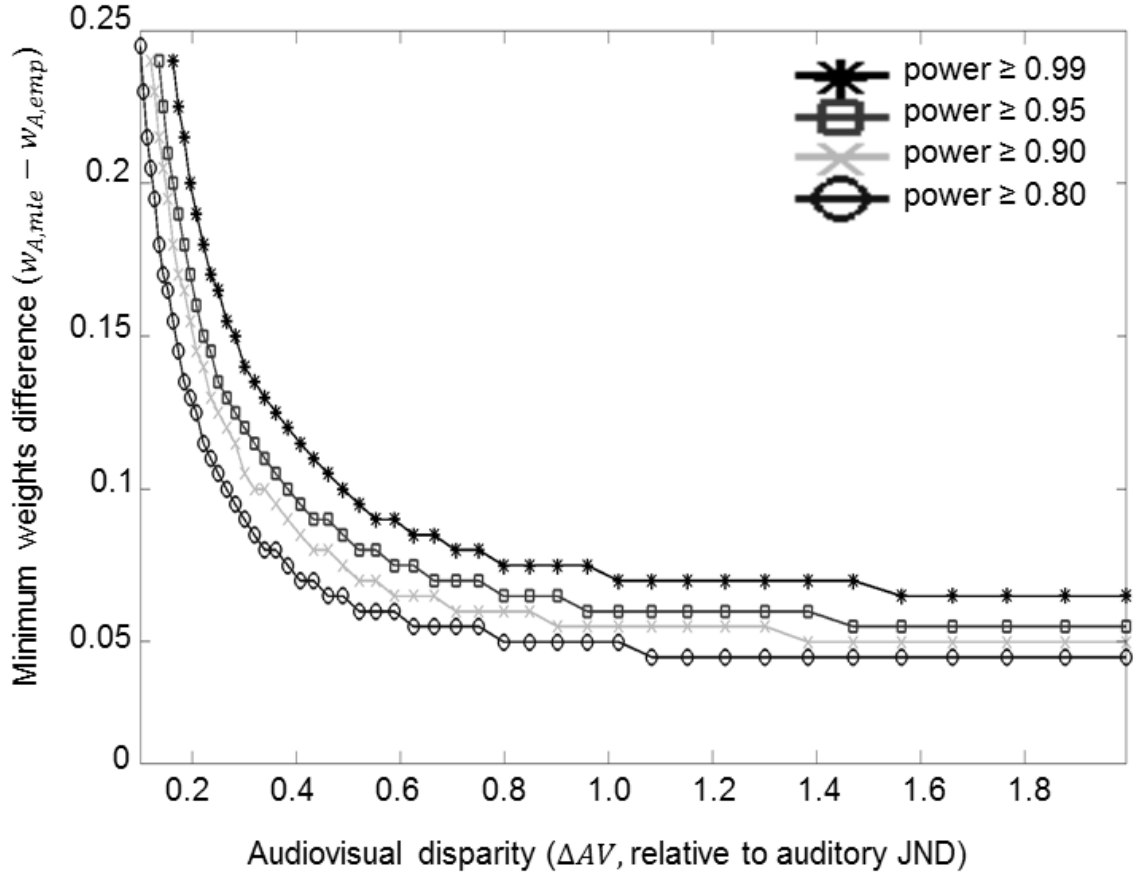


Fig B.1 – Results of power analysis simulations for selecting spatial disparity: Minimum deviations of empirical weights from MLE-predicted weights ($w_{A,mle} \approx 0.5$) that are detected with a power of ≥0.8 (circles), ≥0.9 (crosses), ≥0.95 (squares), ≥0.99 (asterisks) as a function of AV spatial disparity (in standardized subject-specific auditory JND units).

Appendix C: Rationale for use of the betabinomial model

Before any data was acquired in this study we had intended to use the binomial model for fitting of psychometric functions (with Eq. 4 for the likelihood). The main outcomes of the study (as determined by group-level one-sided t-tests), with analyses methods using the binomial model, are that we observe significant visual overweighting relative to MLE-predictions while the audiovisual SDs are not significantly different from MLE-predicted audiovisual SDs (i.e. very similar to the main outcomes of the betabinomial model as reported in the Results Section).

Methods as pre-registered (binomial model; N=36):

$\sigma_{AV,emp} > \sigma_{AV,mle}$: t(35) = 0.38, p = 0.35, BF_{10} = 0.246, BF_{01} = 4.07, d_z = 0.063

$w_{A,emp} < w_{A,mle}$: t(35) = 6.14, p < 0.0001, BF_{10} > 10000, BF_{01} < 0.0001, d_z = 1.02

However, we found that the goodness-of-fit according to the bootstrap analysis using the binomial model was significantly bad (p < 0.05) for 13 participants. We cannot technically trust the parameter estimates for 'bad' fits, because the bad fits suggest that the applied model (consisting of five psychometric functions) is incorrect for these datasets. There are two solutions: we either change the data or we change the model. We explore both options in the discussion below and we will argue for the latter solution: a minor change to the model.

We could have excluded and replaced all thirteen subjects that show a bad goodness-of-fit, as we had originally written in the pre-registered methods. However, when we wrote that we thought that the goodness-of-fit measure (using the binomial model) would allow us to detect outliers among all datasets that were exceptionally noisy. We did not

expect to find that over one third of our participants would show a 'bad' fit. This means that we cannot speak of exceptionally noisy outliers anymore. Instead, it seems that some amount of such noise is characteristic to many of the acquired datasets.

Some amount of measurement noise seems inherent to any psychophysical dataset, and it is expected to be larger for experiments that are spread over multiple sessions with thousands of trials and a total duration of several hours. Between sessions and blocks of trials attention and vigilance levels, as well as internalized criteria may fluctuate causing divergences in participants' performance, thus leading to noise in the data. The simulated datasets that we generate for the parametric bootstrap analysis (to determine goodness-of-fit using the binomial model) do not contain such noise (i.e. performance fluctuations over time are not modelled). The binomial model does produce some random noise in the data ('right' responses are drawn at random from binomial distributions), but the predicted amount of this noise is considerably reduced when the total number of trials per probe location is high (as is the case in the current study). So, the quantitative fit (by means of the likelihood ratio-based measure, see Section 2.10) is likely better for the simulated datasets than for the empirical datasets.

The above-raised concerns regarding the bootstrap test for goodness-of-fit using the binomial model are also described in the recent literature (e.g. Fründ, Haenel, & Wichmann, 2011). In fact, one of the authors that previously argued in favor of the bootstrapped goodness-of-fit test based on the binomial model (Wichmann & Hill, 2001a), has recently co-developed a novel betabinomial model for psychometric function fitting that takes into account noise due to performance fluctuations over time

(Schütt, Harmeling, Macke, & Wichmann, 2016)[13]. The betabinomial distribution has one additional parameter, η, that models the amount of noise present in the data in addition to the predicted variance from the binomial distribution (when η=0 there is no such noise, when η=1 there is only noise).

We have implemented the betabinomial model into our analysis scripts by (i) modifying the formula for computation of the likelihood per location (Eq. 4 becomes Eq. 5) and (ii) by drawing the probability of being correct per probe location from a beta distribution (with variance dependent on η and centred on the probability correct as given by the psychometric function) during parametric bootstrapping (i.e. when simulating datasets; see Section 2.10). To be clear, we only fit one additional parameter per participant: η is fit across the five psychometric functions simultaneously.

Most importantly, using the bootstrap-based goodness-of-fit measure with the betabinomial model we find that none of the 36 participants shows a 'bad' fit ($p > 0.05$ for all), thus suggesting that the betabinomial model is the correct model for all datasets. The fitted noise parameter η is higher for those thirteen participants that previously showed a bad fit ($0.04 < \eta < 0.13$) than for the other twenty-three participants ($0 < \eta < 0.07$), but all values of η are still modest (c.f. η=0.2 indicates a moderately overdispersed observer; Schütt, Harmeling, Macke, & Wichmann, 2016). The main group-level study outcomes using the betabinomial model (as reported in the Results Section) are very similar to the results that we found using the binomial model (see above), but the important difference is that we can now trust these fits to be of adequate quality.

[13] We were unaware of this development during stage-1 submission of the methods.

By applying the betabinomial model instead of the binomial model for psychometric functions we have effectively put all thirty-six datasets on a continuum of various noise levels in the data, instead of treating the thirteen datasets that resulted in a 'bad' goodness-of-fit as fundamentally different from the twenty-three other datasets. We will now provide evidence that justifies this choice by showing that the thirteen 'bad' datasets are not fundamentally different from the other twenty-three datasets.

First, it is important to mention that many of the 'bad' fits only barely reached significance in the bootstrap goodness-of-fit analysis. This is best supported by the finding that none of the 36 participants had a bad goodness-of-fit ($p<0.05$) if the p-value was alternatively computed by comparing the deviance of the dataset with the theoretical distribution of deviances for a binomial model (i.e. for large numbers of trials this asymptotically approximates the chi-squared distribution; Wichmann & Hill, 2001a; Kingdom & Prins 2016). The fact that none of the 36 datasets results in a significantly bad fit using that alternative, commonly-used method supports the notion that the amount of additional noise in the datasets is modest at most (in agreement with the relatively small values that we found for η, see above).

Second, when we perform the statistical tests for main outcome measures separately for the 'good' and 'bad' datasets, using the parameter values as fitted with the binomial model, the results for both groups showed very similar trends. This thus excludes the possibility that e.g. visual overweighting is due to bad model fits (the opposite seems true: visual overweighting is more pronounced in the 'good' fits).

Methods as pre-registered (binomial model): 'Good' goodness-of-fit subjects only (N=23):

$\sigma_{AV,emp} > \sigma_{AV,mle}$: t(22) = 0.93, p = 0.18, BF_{10} = 0.52, BF_{01} = 1.92, d_z = 0.19

$w_{A,emp} < w_{A,mle}$: t(22) = 8.75, p < 0.0001, BF_{10} > 10000, BF_{01} < 0.0001, d_z = 1.82

Methods as pre-registered (binomial model): 'Bad' goodness-of-fit subjects only (N=13):

$\sigma_{AV,emp} > \sigma_{AV,mle}$: t(12) = -0.57, p = 0.71, BF_{10} = 0.19, BF_{01} = 5.17, d_z = 0.16

$w_{A,emp} < w_{A,mle}$: t(12) = 1.55, p = 0.073, BF_{10} = 1.33, BF_{01} = 0.75, d_z = 0.43

Third, one may argue that the reason for the 'bad' fits is that we imposed constraints on the five psychometric functions: (i) they shared one lapse rate parameter and (ii) the three audiovisual conditions shared one slope parameter. If the underlying assumptions for these constraints are invalid (e.g. the lapse rate is different for auditory and visual conditions, or the variance is higher for spatially incongruent audiovisual conditions as opposed to the spatially congruent condition), then we should find that a model without these constraints fits the datasets much better. We tested this hypothesis by fitting a binomial model of five psychometric functions with five lapse rates and five slope parameters. We used parametric bootstrapping (applying the binomial model) to obtain a measure for goodness-of-fit for each dataset. For this unconstrained binomial model we again found thirteen significantly bad fits (ten of whom are the same datasets that showed 'bad' fits before). We further tested the assumption for equal audiovisual slopes by directly comparing the mean SD of the two incongruent audiovisual conditions with the SD of the congruent condition. We found no evidence that the assumption was violated (one-sided group-level t-test: $\sigma_{AV,incongr} > \sigma_{AV,congr}$, t(35) = 1.11, p=0.14, BF_{10} = 0.54, BF_{01} = 1.85, d_z = 0.18). We thus conclude that it is unlikely that the parameter

constraints on lapse rates and slopes were the reason for the significantly 'bad' goodness-of-fit test results. Instead, we argue that there is a modest amount of additional noise in all datasets that can be adequately taken into consideration by using the betabinomial model, leading to good model fits for all participants.

Appendix D: Two simulations to explain visual overweighting with apparent AV variance as predicted by maximum likelihood estimation.

The main outcome measures of this study suggest that participants put significantly more weight on the visual signals than is predicted by MLE ($w_{A,emp} < w_{A,mle}$), while at the same time managing to reduce their bisensory variance (relative to both unisensory variances) to the same extent as is predicted by MLE. This is an apparent contradiction because an observer should only be able to achieve such MLE-optimal variance reduction (Eq. 2) when the unisensory signals are weighted according to their respective reliabilities, exactly as the MLE model prescribes (Eq. 1). However, in two separate simulation studies we will here provide proof-of-principle evidence that it is possible to obtain a 'seemingly' MLE-optimal variance reduction while using a different weighting scheme. These simulations demonstrate that our study's design was likely more sensitive to detecting deviations of the weights than to detect deviations of the audiovisual variances.

Bayesian causal inference

The Bayesian causal inference (BCI) model describes how an ideal observer deals with the uncertainty that is associated with unisensory location estimates, similar to the MLE-

model (Section 2.1), while additionally taking into account the uncertainty about whether the unisensory signals come from common or separate sources. In other words, BCI prescribes whether or not unisensory signals should be integrated or segregated, and more specifically, to what extent they should be integrated. Without providing a full description of the BCI model (we kindly refer the reader to previously published work; Körding, Beierholm, Ma et al., 2007; Rohe & Noppeney, 2015a), we here simulated responses for a group of thirty-six BCI-ideal observers in the experimental setting of the current study for the two audiovisual incongruent conditions. We then fitted psychometric functions to these responses and show that, for certain parameter settings, the group-level statistical outcomes are similar to the behavioral outcomes as described in the Results Section.

1. First, we briefly describe how a single location estimate is simulated for an audiovisual stimulus using the principles of Bayesian causal inference. Given the true auditory and visual locations S_A and S_V, we randomly draw noise-disturbed internal estimates from their Gaussian sensory noise distributions $\hat{S}_A \sim N(S_A, {\sigma_A}^2)$ and $\hat{S}_V \sim N(S_V, {\sigma_V}^2)$. We then compute the posterior estimates under the two causal hypotheses: 1). the auditory and visual signals come from a common source (C=1) and need to be integrated; and 2). the two signals come from separate sources (C=2) and should not be integrated. N.b. in addition to the MLE-model, the Bayesian ideal observers simulated here multiply the likelihood with a normally distributed spatial prior $P(S) = N(\mu_P, {\sigma_P}^2)$, such that the best estimates for either causal structure can be computed according to:

$$\hat{S}_{AV,C=1} = \frac{\frac{\hat{S}_A}{{\sigma_A}^2} + \frac{\hat{S}_V}{{\sigma_V}^2} + \frac{\mu_P}{{\sigma_P}^2}}{\frac{1}{{\sigma_A}^2} + \frac{1}{{\sigma_V}^2} + \frac{1}{{\sigma_P}^2}} \quad , \quad \hat{S}_{A,C=2} = \frac{\frac{\hat{S}_A}{{\sigma_A}^2} + \frac{\mu_P}{{\sigma_P}^2}}{\frac{1}{{\sigma_A}^2} + \frac{1}{{\sigma_P}^2}}, \qquad \text{and} \qquad \hat{S}_{V,C=2} = \frac{\frac{\hat{S}_V}{{\sigma_V}^2} + \frac{\mu_P}{{\sigma_P}^2}}{\frac{1}{{\sigma_V}^2} + \frac{1}{{\sigma_P}^2}}$$

Applying the so-called 'model averaging' decision strategy, BCI ideal observers compute the final estimates for auditory and visual stimulus location as a weighted mean of the integration (C=1) and segregation (C=2) hypotheses, where the weights are determined by the posterior probability of each causal structure:

$$\hat{S}_{A,BCI} = P\left(C = 1 \middle| \hat{S}_A, \hat{S}_V\right) * \hat{S}_{AV,C=1} + \left(1 - P\left(C = 1 \middle| \hat{S}_A, \hat{S}_V\right)\right) * \hat{S}_{A,C=2}$$

$$\hat{S}_{V,BCI} = P\left(C = 1 \middle| \hat{S}_A, \hat{S}_V\right) * \hat{S}_{AV,C=1} + \left(1 - P\left(C = 1 \middle| \hat{S}_A, \hat{S}_V\right)\right) * \hat{S}_{V,C=2}$$

With (by Bayes' rule):

$$P\left(C = 1 \middle| \hat{S}_A, \hat{S}_V\right) = \frac{P(\hat{S}_A, \hat{S}_V | C=1) * P_{common}}{P(\hat{S}_A, \hat{S}_V | C=1) * P_{common} + P(\hat{S}_A, \hat{S}_V | C=2) * (1 - P_{common})}$$

P_{common} is the prior probability that the two signals originate from a common source independent of the actual stimulus locations (please see Körding, Beierholm, Ma et al., 2007 for full details).

2. For any one trial, we simulate location estimates for two audiovisual stimuli, standard and probe, and compute the difference between the estimates of each stimulus:

$\widehat{\Delta S}_{A,BCI} = \hat{S}_{A,BCI,probe} - \hat{S}_{A,BCI,standard}$ and $\widehat{\Delta S}_{V,BCI} = \hat{S}_{V,BCI,probe} - \hat{S}_{V,BCI,standard}$

We thus obtain two estimates for the difference between standard and probe: one for each sensory modality. We discretize these two difference estimates ($z = $ "right" $if\ \widehat{\Delta S}_{BCI} > 0°$ and $z = $ "left" $if\ \widehat{\Delta S}_{BCI} \leq 0°$) to simulate modality-specific responses z_A and z_V. However, since the experimental design of this study asks for the "location of the audiovisual probe relative to the audiovisual standard", the BCI ideal

observer is placed for an impossible choice when $z_A \neq z_V$: which sensory modality should (s)he respond to? Based on (unrequested) information provided by seven participants in the post-testing questionnaire (Section 2.6.2.3) we make the important assumption here that all simulated observers respond z_V when in doubt; as the seven participants had indeed self-reported to have done. This assumption results in visual overweighting (relative to MLE predictions) for any $P_{common} < 1$; i.e. when the forced-fusion assumption is violated.

3. For one experiment, we simulate 40 responses (z_V) for each of thirteen probe locations. We do this twice, once for each of the two incongruent conditions $AV_{\Delta=+X^\circ}$ and $AV_{\Delta=-X^\circ}$. The process is repeated for thirty-six BCI ideal observers. For each simulated observer, the individualized probe locations S_{pr}, audiovisual disparity ΔAV, and unisensory noise parameters σ_A, σ_V are one-to-one matched to the locations, disparity and unisensory noise parameters that were used/obtained for the thirty-six real participants. The unknown spatial prior parameters are set to $\mu_P = 0^\circ$ and $\sigma_P = 100^\circ$, thus assuming a near-uniform prior distribution for all simulated observers. P_{common} is constant across participants but its value varies for different simulation runs (see step 5 below).

4. For each BCI ideal observer two psychometric functions are fitted to the simulated responses using the binomial model (Eq. 4), one for each of the incongruent conditions. To obtain one estimate of $\sigma_{AV,emp}$ the two simulated empirical AV variances are averaged. Following the procedure as outlined in Section 2.8-9, group-level paired t-tests can then be performed to test for significant deviations from the MLE-predicted weights and audiovisual variances.

5. The above-described simulation analysis is repeated 1000 times for each of 11 different values of P_{common}, spread at regular intervals between 0.5 and 1. For each setting of P_{common} we compute the proportion of the 1000 statistical t-tests that resulted in significant deviations from the MLE-predictions, separately for weights and AV variances.

Summary results of these simulations are shown in Figure D.1. We conclude that if our thirty-six participants would have behaved like BCI ideal observers with relatively high but not 100% certain prior beliefs of the audiovisual signals coming from a common source (i.e. $0.75 < P_{common} \leq 0.95$), then it is likely to observe group-level test results similar to what we observed in the actual behavioral data: significant visual overweighting without a significant deviation from the MLE-predicted audiovisual variance.

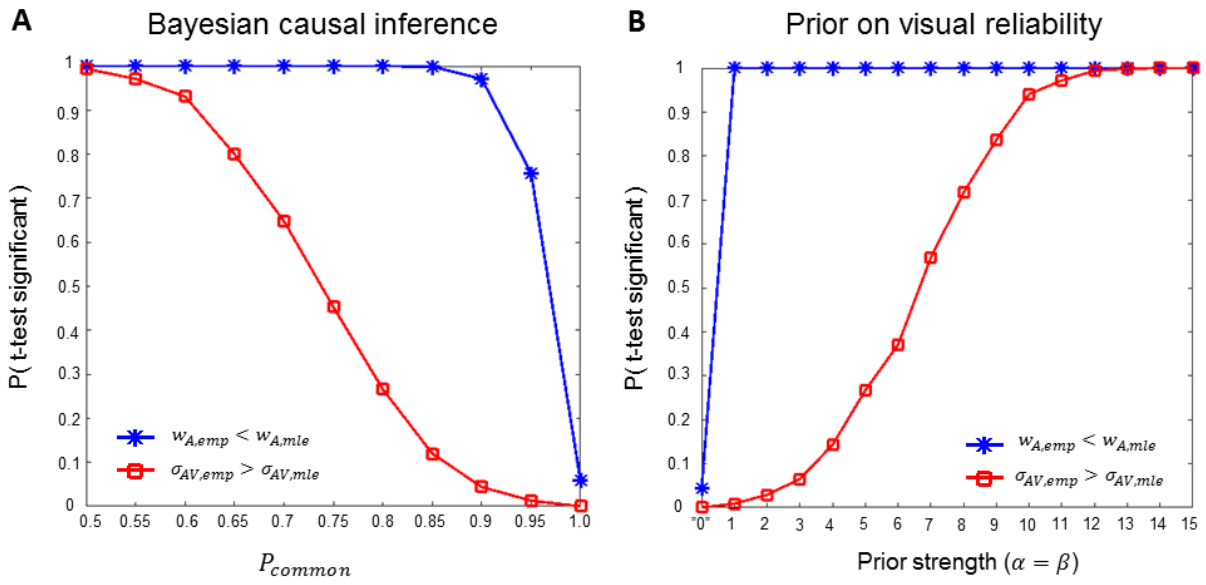


Fig D.1 – Results of the simulation analyses for ideal observers applying Bayesian causal inference with various strengths of their common source prior (panel A) and Bayesian observers utilizing a prior on their visual reliability estimates that favors smaller visual variance estimates (panel B). Displayed here are the proportions of simulated experiments (N=1000) that resulted in a significant difference between empirical and MLE-predicted parameters according to one-sided group-level t-tests. Fig. D.1 b. Please note that a prior strength of "0" indicates that a uniform prior was used for the simulations: i.e. the visual reliability estimates were unbiased.

Bayesian prior on modality-specific reliability estimates

Similar to what we observed in the current study, Battaglia et al. (2003) also observed visual overweighting in an audiovisual localization task. The authors proposed a Bayesian model that fitted their group-level data much better than the MLE-model. They hypothesized that a Bayesian observer would hold a prior belief that high visual spatial reliabilities are more probable than low visual spatial reliabilities, because under normal, non-laboratory, everyday-life conditions visual spatial reliability is most often reasonably high (at least more reliable than auditory spatial signals). When a Bayesian observer estimates the reliability of incoming visual sensory signals it makes use of a prior probability distribution to bias his visual reliability estimates towards higher values. Since the reliability *estimates* of the two unisensory modalities, r_A and r_V, are used to define the weights for audiovisual integration (Eq. 1), such a bias would inevitably result in visual overweighting. Taking a similar simulation approach as we have done above for BCI ideal observers, we here generate responses for thirty-six observers that use

Bayesian visual reliability estimates (BVRE) to perform reliability-weighted integration. Testing a variety of strengths for the prior over visual reliabilities we show that significant visual overweighting can be observed as a group-level statistical outcome without a concurrent deviation of the audiovisual variance from MLE-predictions.

1. We construct a set of 15 visual reliability priors using inverse-gamma probability distributions over the visual variance estimates ${\sigma_V}^2$ (Battaglia et al., 2003). By setting the scaling parameter equal to the shape parameter, $\alpha = \beta$, we ensured that all priors peaked at similarly high probabilities for low variances, i.e. ${\sigma_V}^2 \approx 1$, while their probabilities decrease at different rates for increasing variance values; see Figure D.2.

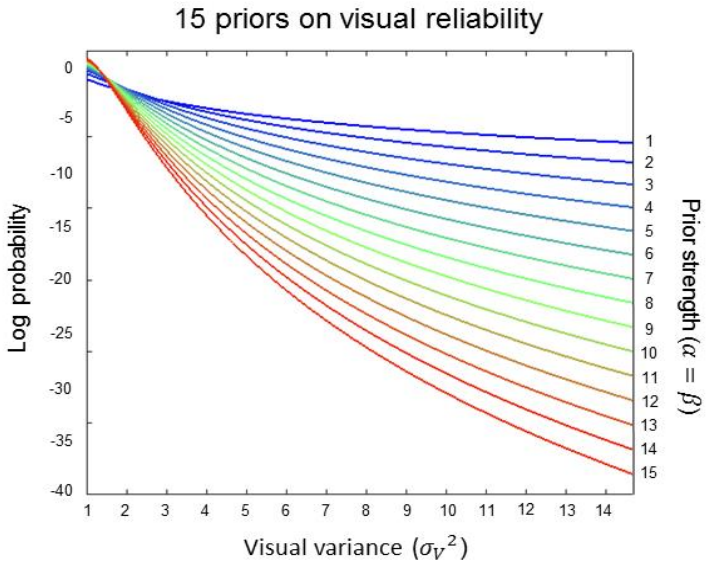


Fig D.2 – The fifteen prior probability distributions that were used in the simulations.

2. To create conditions that are similar to our behavioral experiment we again match the individualized probe locations S_{pr}, audiovisual disparity ΔAV, and (true) unisensory

noise parameters σ_A, σ_V of thirty-six simulated BVRE observers to those of the thirty-six actual participants. These parameters are used to simulate internal unisensory estimates by randomly drawing from Gaussian sensory noise distributions centered on the true auditory and visual locations: $\hat{S}_A \sim N(S_A, {\sigma_A}^2)$ and $\hat{S}_V \sim N(S_V, {\sigma_V}^2)$.

3. BVRE observers integrate these unisensory estimates (step 2) weighted according to their *estimated* reliabilities (as opposed to their *true* reliabilities as in the MLE-model). In order to simulate the visual reliability estimates of the BVRE observers we fit a new psychometric function to the behavioral responses of the actual participants for the unisensory visual condition. However, instead of optimizing the function's parameters by maximizing the likelihood we now maximize the posterior probability: i.e. the likelihood as obtained from Eq. 4 multiplied by the prior probability for these parameters. Notably, the prior probability over all parameters of the psychometric function is entirely determined by the prior probability of the visual variance estimate (as given by the prior probability distribution defined in step 1). The maximum posterior probability thus is biased towards lower visual variance estimates. These newly fitted visual variance estimates are used to determine the weights for audiovisual integration according to Eq. 1. N.b. The auditory variance estimates are assumed to be unbiased and set to equal the true auditory variance ${\sigma_A}^2$ (Battaglia et al., 2003).

4. For any one trial two integrated audiovisual location estimates: $\hat{S}_{AV,standard}$ and $\hat{S}_{AV,probe}$ are computed. If the difference between these two estimates, $\widehat{\Delta S}_{AV} = \hat{S}_{AV,probe} - \hat{S}_{AV,standard}$, is positive then the simulated response $z =$ "right". If not, $z =$ "left".

5. For each BVRE observer 40 responses per probe location (N = 13) are simulated for each of the two incongruent audiovisual conditions. Subsequently, two psychometric functions are fitted to the simulated responses (by maximizing the likelihood according to the standard experimental methods; Eq. 4). From these psychometric functions we obtain the empirical auditory weight (Eq. 6) and audiovisual noise parameters: $\sigma_{AV,emp} = \frac{1}{2}\sigma_{AV,VLAR} + \frac{1}{2}\sigma_{AV,ALVR}$.

6. The above-described simulation process is repeated for all thirty-six BVRE observers such that the simulated empirical weights and noise parameters can be compared with the MLE-predicted weights and noise parameters (n.b. MLE-predictions are computed using the *true* unisensory noise parameters σ_A, σ_V and equations 1 & 2). Group-level t-tests determine whether a significant difference exists. One thousand such experiments are simulated for each of the 15 visual reliability priors (step 1). Moreover, we additionally perform one thousand experiments for unbiased visual reliability estimates (i.e. simulating a uniform prior, which is equivalent to simply generating responses according to the MLE model).

Figure D.1b shows the summary results of these simulations. Significant visual overweighting is observed in all of the experiments that were modeled with a prior on the visual reliability; even when this prior was only moderate in strength. However, the prior strength needs to be reasonably strong in order to observe a significant deviation from the MLE-predictions in terms of audiovisual variance. This analysis thus confirms that the experimental design used in the current study is more sensitive for detecting weights differences and it provides a possible explanation for the two seemingly paradoxical main outcomes.

CHAPTER 4

NEITHER TRAINING NOR REWARD FORCES OBSERVERS TO FUSE SPATIALLY DISPARATE AUDIOVISUAL SIGNALS

David Meijer, Uta Noppeney

CONTRIBUTIONS

David Meijer and Uta Noppeney designed the experiment and wrote the introduction. David Meijer wrote the methods, results and discussion sections. David Meijer prepared the MATLAB scripts for stimulus presentation. David Meijer performed data analyses.

ACKNOWLEDGMENTS

We thank Bethany Robinson, Samina Begum and Wing Lam Cham (a.k.a. Chloe) for acquiring the behavioral data. We thank Sam Jones for helpful discussions.

This research was funded by ERC-2012-StG_20111109 multsens

ABSTRACT

It is widely believed that multisensory integration is consistent with the statistically optimal rules of maximum likelihood estimation (MLE): Best-precision bisensory estimates are obtained by computing a reliability-weighted average of the two unisensory signals. While human bisensory perceptual behavior has been shown to be in accordance with MLE-predictions for various combinations of sensory modalities, there also exist numerous reports of clear divergence from MLE-optimal multisensory integration. The current study aimed to investigate which factors affect statistical (sub-) optimality of human perception in a spatial audiovisual localization task. Our approach was two-fold: 1) We employed a 2x2 factorial between-subjects experimental design wherein we manipulated prior training and reward-based motivation. 2) We used Bayesian model comparison to arbitrate between previously-proposed explanations of apparent suboptimal multisensory perception. Results replicated previous findings of visual overweighting relative to MLE predictions. While prior training improved auditory localization performance significantly, we found no evidence that visual overweighting was smaller after training. Moreover, we found no effect of reward-based motivational factors on audiovisual integration. Model comparison analysis demonstrated that Bayesian causal inference with reliance on visual location estimates provided the most probable explanation for deviations from maximum likelihood estimation. This suggests that human observers who are required to make judgments about multisensory signals while facing uncertainty about the common origin of their unisensory components default to rely on cross-modally biased - but not fully integrated - estimates in the sensory modality that is normally most trustworthy (i.e. vision for localization).

1. Introduction

Multisensory integration is the combining of information from multiple sensory organs in order to obtain a coherent and robust understanding of the environment (Ernst & Bülthoff, 2004). For example, audiovisual speech integration may help us to infer who is saying what in a busy bar. If sensory information is merged probabilistically, i.e. taking into account the uncertainty associated with each unisensory stream, this will have the benefit of reducing one's perceptual variance. Maximum likelihood estimation (MLE) describes a probabilistic information integration strategy in which unisensory estimates are weighed by their respective reliabilities when computing a multisensory average (Ernst & Banks, 2002). It is statistically optimal in the sense that it leads to the most precise multisensory estimates given the sensory noise.

Many studies of human behavior have reported (near-) optimal multisensory integration for a variety of object features and sensory combinations, including visual-tactile size and shape estimates (Ernst & Banks, 2002; Helbig & Ernst, 2007), audiovisual and visual-proprioceptive location estimates (Alais & Burr, 2004; van Beers, Sittig & Denier van der Gon, 1996), audiovisual duration estimates (Hartcher-O'Brien, Di Luca & Ernst, 2014) and audiovisual, audiotactile, and visual-tactile numerosity judgments (Bresciani et al., 2005; Bresciani, Dammeier & Ernst, 2006; Shams, Ma & Beierholm, 2005; Wozny, Beierholm & Shams, 2008). However, over the past decade accumulating research has also revealed examples of suboptimal multisensory integration where the variance reduction for the multisensory percept was lower than predicted by statistically optimal models and/or reliability weighting, although present, did not follow MLE predictions (Rahnev & Denison, 2018).

This sort of suboptimal multisensory integration may occur if assumptions of the MLE model are violated. For instance, the standard MLE model does not account for dependencies in the sensory noise across the two modalities (Gepshtein & Banks, 2003, Oruç, Maloney & Landy, 2003), additional supra-modal sources of noise (Burr, Banks & Morrone, 2009; Battaglia, Kersten & Schrater, 2011), differences in the modality-specific sensory noise between unisensory and bisensory stimulus presentations (Bejjanki, Clayards, Knill & Aslin 2011) or non-normal noise distributions (Burr et al., 2009). Likewise, the MLE model does not account for short-term perceptual adaptations (Triesch, Ballard & Jacobs, 2002; Wozny & Shams, 2011) or long-term priors (Odegaard, Wozny & Shams, 2015). For example, it has been argued that human observers may impose a prior on the sensory variances based on their everyday experiences leading to overweighting of a sensory signal in multisensory perception (e.g. the visual signal in spatial localization, Battaglia, Jacobs and Aslin, 2003; for other modalities see Butler, Smith, Campos & Bülthoff, 2010; Fetsch, Turner, DeAngelis & Angelaki, 2009; Maiworm & Röder, 2011; Prsa, Gale & Blanke, 2012). Furthermore, it has been suggested that participants may not be able to estimate their sensory reliability accurately when the task is unfamiliar (Rosas, Wagemans, Ernst & Wichmann, 2005) or complex (Beck, Ma, Pitkow, Latham & Pouget, 2012). Finally, the MLE model is limited to the special forced fusion case. It ignores the so-called causal inference problem, i.e. whether two signals come from a common source and should be integrated, and mandatorily binds signals into one unified percept. By contrast, hierarchical Bayesian causal inference accommodates the observer's uncertainty about the world's causal structure by explicitly modelling the potential causal structures (i.e. one or two sources) that may have generated the sensory signals. This enables a graceful transition between sensory

integration and segregation depending on the probabilities of the world's causal structure (i.e. common vs. independent causes; Körding, Beierholm, Ma et al., 2007; Rohe & Noppeney, 2015a). Hence, human integration that has been considered putatively suboptimal from the MLE perspective may turn out to be optimal according to Bayesian Causal Inference.

Given this plethora of potential reasons for MLE suboptimal human behaviour, this study aimed to identify the conditions that enable observers to integrate information across the senses in line with the predictions of the MLE model. For this, we focused on the integration of spatial signals from vision and audition where previous research has found both MLE-optimality (Alais & Burr, 2004) and MLE-suboptimality (Battaglia et al., 2003; Meijer & Noppeney, 2018) in human observers. In particular, we investigated task experience and motivation as potential factors that may influence whether human observers integrate sensory signals optimally in line with MLE predictions. Furthermore, we applied Bayesian model comparison to test the likelihood of various hypotheses for human multisensory perception diverging from maximum likelihood estimation.

2. Method

2.1 Experimental overview

We investigated whether potential reward ('motivation') or prior training ('task experience') influence whether participants integrate visual and auditory signals for spatial localization optimally as predicted by maximum likelihood estimation. To address this, we manipulated whether observers were i. trained on the audiovisual localization

task for two days and/or ii. provided with a reward depending on their localization performance relative to a prior baseline measurement. Participants were grouped according to a 2x2 factorial between-subjects design: Group 1: no training, no reward; Group 2: no training, reward; Group 3: training, no reward; Group 4: training, reward. Fig. 1a provides an overview of the four versions of this experiment. Participants in groups 1 and 2 completed two experimental sessions: one introductory and one final test session during which the main experiment was performed and all data for analysis was obtained. Participants in groups 3 and 4 completed four experimental sessions: one introductory, two training and one final test session. All sessions lasted approximately 2.5 hours and were performed on separate days.

2.2 Participants

We set out to recruit forty participants in total: ten participants per group. Six participants were excluded during the first session for failure to meet the required auditory localization performance criterion (see Section 2.5.1.2). These participants were replaced such that forty participants completed the experiment successfully (35 females, mean age 19, $\pm$1.6 years SD). All participants were university students who reported normal hearing, (corrected to) normal vision and no history of neurological or psychiatric disorder. All provided informed consent at the start of session 1. Participants were compensated by means of course credits. At sign-up participants knew whether they would attend a two or four-day study (groups 1-2 or 3-4, respectively), but were unaware of the potential to receive an additional performance-dependent monetary reward (groups 2 and 4). Participant allocation to groups with a potential reward was performed pseudorandomly by the computer program at the beginning of the final test

session. The study was approved by the human research review committee of the University of Birmingham (approval number ERN_11-0470AP4).

2.3. Experimental setup and stimuli characteristics

Visual stimuli were low-contrast greyscale circular blobs (16.7 ms duration) that were brightest in the centre (1.2 cd/m^2) and gradually decreasing in brightness towards a black background (0.24 cd/m^2). Their size was defined by the standard deviation of a bivariate Gaussian amplitude envelope, σ_{blob}, which was adjusted for each individual observer to match the visual reliability to the auditory reliability (see Section 2.5.1.3). The blobs were back-projected (60Hz DLP projector, BenQ MW529) on a grey screen (opaque fine PVC fabric, 127.5 cm width x 170 cm height) that was placed at a distance of 78 cm from the participants.

Auditory stimuli were bursts of white noise (60 dB SPL, 16.7 ms including 5 ms on/off ramps) convolved with standardised head-related transfer functions (modulating phase and amplitude between left and right ear channels) to create the illusion of a spatial offset along the azimuth (Gardner & Martin, 1995; http://sound.media.mit.edu/resources/KEMAR.html). The sounds were presented by means of headphones (Sennheiser HD 280 Pro) with a playback frequency of 192 kHz. Audiovisual asynchronies were no larger than 3 ms. Stimulus presentation was controlled using Psychtoolbox 3.0.12 (Brainard, 1997; Kleiner, Brainard & Pelli, 2007; www.psychtoolbox.org) running on MATLAB R2016a (www.mathworks.com).

Participants sat in a dark room with a chinrest to provide support and stability. Participants were instructed to focus their gaze on a grey fixation cross (1.2 cd/m^2) prior to stimuli presentation and to remain central gaze focus during stimuli presentation. To

monitor their fixation we used a Tobii EyeX eye tracker (https://tobiigaming.com) that was calibrated before the start of each block of trials using the Matlab Toolbox EyeX (https://sourceforge.net/projects/matlabtoolboxeyex; Gibaldi, Vanegas, Bex & Maiello, 2016). Corrective feedback regarding participants' fixation was given after each block of trials.

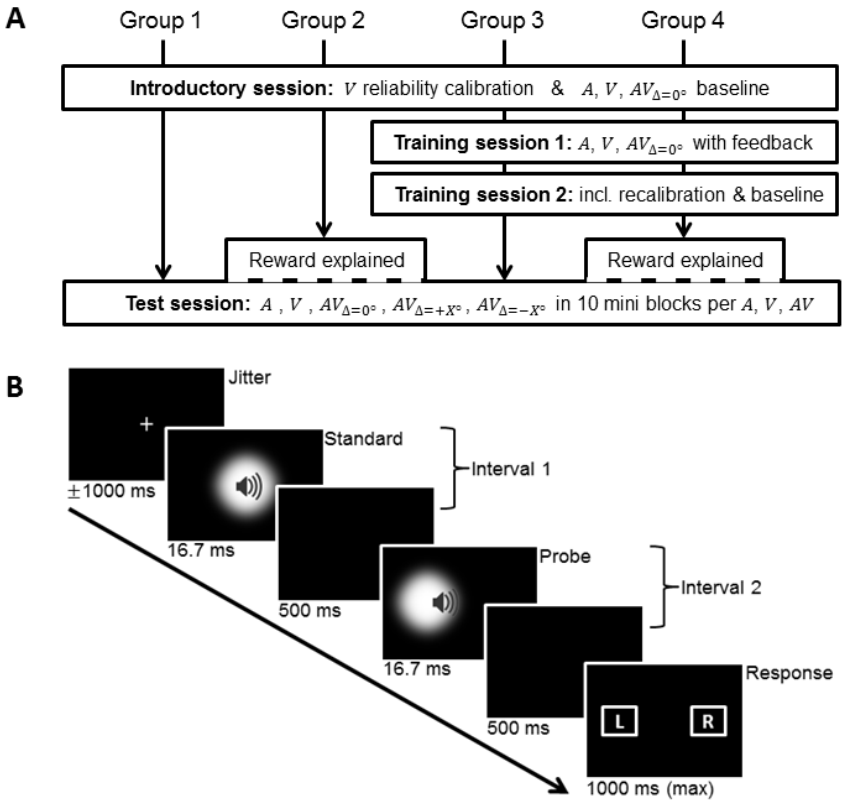


Fig 1 – a. Overview of the four versions of the experiment. Participants either attended two or four sessions (separate days), depending on whether they underwent two training sessions (groups 3 and 4). Participants that were allocated to groups 2 and 4 had the chance to win a performance-based reward. The potential for such a reward was explained on-screen at the beginning of the final test session. See Section 2.5 for details on the experimental procedure. Fig 1 b. Timeline of one AV-incongruent trial.

Participants were asked whether the second stimulus was to the left or right of the first stimulus (both A, V or AV). Audiovisual probe signals could contain a small spatial audiovisual conflict). Trial structure was identical for all experimental groups. See Section 2.4 for details on the two interval forced choice paradigm.

2.4. Two interval forced choice task

Similar to previous studies testing the validity of the MLE model for multisensory integration (e.g. Alais & Burr, 2004) we employed a two interval forced-choice (2IFC) paradigm wherein we presented a standard stimulus at $S_{st} = 0°$ and a probe stimulus at various locations along the azimuth (probe locations S_{pr} are individualized based on participants' auditory performances, see Section 2.5.1.2). However, contrary to previous studies we always presented the standard first and the probe second so that participants could use the standard to update their reference point on a trial-to-trial basis (Dyjas, Bausenhart & Ulrich, 2012) and thus avoid sequential order effects (e.g. Wozny & Shams, 2011). The interstimulus interval was 500 ms. Two rectangles appeared on screen 500 ms after probe offset to prompt participants to report, by means of a button-press using either one of their index fingers, whether the probe was on the left or on the right of the standard. Participants were instructed to focus on accuracy and not on speed. Maximum response time was 1 second. The time between response and the next trial's standard onset was jittered between 750-1250 ms. Fig. 1b provides a trial overview.

On each trial, probe and standard always had the same sensory modality: both auditory (A), visual (V), or audiovisual (AV). Participants were aware of the upcoming sensory modalities because they were presented in blocks and modalities were announced beforehand. Block order (A, V or AV) in the final test session was pseudo-randomized. Importantly, while AV standard stimuli were always spatially congruent (audiovisual disparity $\Delta AV = 0°$), audiovisual probes could be spatially congruent ($S_{pr,A} = S_{pr,V} = S_{pr}$) or spatially incongruent ($S_{pr,A} = S_{pr} - \frac{1}{2}\Delta AV$ and $S_{pr,V} = S_{pr} + \frac{1}{2}\Delta AV$, with $\Delta AV = \pm X°$ set individually per participant, see Section 2.5.1.2). N.b. AV incongruent stimuli were only presented in the final test session.

2.5. Experimental procedure

In the following we will briefly describe the introductory, training (groups 3 and 4) and test sessions:

2.5.1 Introductory session: The first session consisted of five experimental parts for all four groups:

2.5.1.1. In order to familiarize themselves with the stimuli participants completed a short series of V, A, and $AV_{\Delta=0°}$ trials (2 trials x 12 locations for each modality: S_{pr} was pseudo-randomly drawn from $\{\pm 15°, \pm 13°, \pm 10°, \pm 7°, \pm 4°, \pm 1°\}$). The visual blob size was jittered between relatively small/easy settings: $\sigma_{blob} \sim \{4°, 6°, 8°\}$. Immediate feedback was given after each response, i.e. a green (correct) or red (incorrect) circle was presented on screen for 200 ms. To ensure that participants understood the task these familiarization series were repeated if participants made errors in trials where $S_{pr} = \pm 15°$. Participants were told to consider the AV stimulus as a ball bouncing on

the back of the screen. Presumably, thinking of the AV stimulus as one audiovisual event increases an observer's likelihood to integrate the auditory and visual signals (Alais & Burr, 2004).

2.5.1.2. The first real task for participants was to complete a series of A trials (20 trials x 13 locations: $S_{pr} \sim \{\pm 10°, \pm 7°, \pm 5°, \pm 3°, \pm 2°, \pm 1°, 0°\}$). Participants were excluded (and replaced, see Section 2.2) from further participating in the study if they made more than four mistakes on those forty trials where $S_{pr,A} = \pm 10°$. A psychometric function was fit to the proportion of "right" responses (see Section 2.6) to obtain the perceived reliability of the auditory stimuli for each participant (n.b. reliability is here expressed in terms of the just noticeable difference: JND_A). This auditory reliability measurement served four important functions:

i. From here onwards, all probe locations (for A, V and AV trials) were set relative to the auditory JND as $S_{pr} = JND_A * \{\pm 3, \pm 2.5, \pm 2, \pm 1.5, \pm 1, \pm 0.5, 0\}$, rounded to 0.5° under the constraint of having thirteen unique probe locations for each participant. Using individualized probe locations relative to performance ensures that parameters of the psychometric functions can be adequately fitted because samples are obtained on all essential parts of the curve (Wichmann & Hill, 2001a).

ii. The blob size (σ_{blob}, see Section 2.3) was individually titrated such that the visual JND matched the auditory JND for each participant (see Section 2.5.1.3). Equating the unisensory reliabilities leads to the greatest audiovisual variance reduction according to MLE predictions, thus maximizing the sensitivity of this study to observe this characteristic signature of optimal multisensory integration.

iii. The audiovisual disparity for spatially incongruent stimuli was set to equal the auditory JND: $\Delta AV = \pm JND_A$ (rounded to the nearest half degree visual angle). The choice for a correct conflict size of the bisensory signals is a delicate one. On the one hand we would like to use large conflict sizes because this would maximize experimental sensitivity for measuring the sensory weights that are used for multisensory integration. On the other hand, by decreasing the audiovisual disparity we lower the risk that participants experience the two sensory signals as coming from different source locations and we thus increase the probability of the signals being integrated (Körding, Beierholm, Ma et al., 2007; Rohe & Noppeney, 2015a). By setting the disparity size equal to the unisensory reliability levels (assuming $JND_V \approx JND_A$) we balance these two objectives (Meijer & Noppeney, 2018; Rohde, van Dam & Ernst, 2016).

iv. The auditory performance measurement of this first session was used as a baseline reference for comparison with the auditory performance in the final test session for participants in group 2. In order to obtain a monetary reward these participants had to improve their performance (i.e. decrease JND_A) in the final test session relative to this baseline (and similarly improve V and $AV_{\Delta=0^\circ}$ performance; see Section 2.5.3).

2.5.1.3. Next, participants completed a series of V trials in which we adaptively controlled the Gaussian blobs' size by adjusting σ_{blob}. This task aimed to match the visual reliability to the auditory reliability for each participant. First we obtained the hypothetical probe locations ($\pm X^\circ$) that would have led to auditory accuracy levels of 68%, 79% and 87% (using the fitted auditory psychometric function; see Section 2.5.1.2 above). Then we used these three location pairs as visual probe locations at which we aimed to match the accuracy levels by increasing σ_{blob} if accuracy was too high and

decreasing σ_{blob} if accuracy was too low. For each location pair we made use of two adaptive weighted up/down staircases (Kingdom and Prins, 2016), one starting at an artificially high σ_{blob} and one starting at a very low value for σ_{blob}. All six staircases were presented interleaved. Each staircase was terminated after thirty reversals. The participant-specific σ_{blob} was computed as the average across all six staircases, where each staircase was averaged over the last twenty reversals.

2.5.1.4. To confirm that the unisensory reliabilities were successfully matched participants completed a series of V trials with the individualized blob sizes from the staircase procedure above (20 trials x 13 locations; using participant-specific probe locations, see Section 2.5.1.2). A psychometric function was then fitted to the responses in order to obtain an estimate of the visual reliability (Section 2.6). If the difference between the two unisensory reliabilities was too large, i.e. if ${JND_V}^2 < 0.5 * {JND_A}^2$ or ${JND_V}^2 > 2 * {JND_A}^2$, then we adjusted σ_{blob} slightly ($\pm 3°$ in desired direction) and asked the participant to complete another series of V trials with this new setting. After such an adjustment we were able to confirm for all participants that the unisensory reliabilities were sufficiently matched (i.e. JND_V was between limits mentioned above). The final participant-specific σ_{blob} was used in subsequent experimental parts. Moreover, the JND_V served as a baseline reference for participants in group 2 to determine whether a reward was earned (see Section 2.5.3).

2.5.1.5. The final task in the first session was a series with $AV_{\Delta=0°}$ trials (20 trials x 13 locations). Although the purpose of this task, to provide a baseline reference for the performance-based reward (see Section 2.5.3), was only relevant for participants in group 2, we required all participants to complete this task for consistency reasons across

groups. A psychometric function was fit to the responses (Section 2.6) to obtain $JND_{AV,\Delta=0^\circ}$.

2.5.2. Training sessions: Participants in groups 3 and 4 attended two training sessions that participants in groups 1 and 2 skipped.

During the first of two training sessions participants completed an equal number of A, V, and $AV_{\Delta=0^\circ}$ trials (60 trials x 12 individualized probe locations x 3 conditions), divided into 6 blocks per condition that were presented in pseudorandom order. Critically, after every response participants were given immediate corrective feedback, similar to the familiarization task of the introductory session (Section 2.5.1.1). In order to encourage participants' training progress, they were also shown whether their overall percentage of correct responses at the end of each block was better or worse (on a 6-point color-coded scale) than their score for the latest previous block of the same condition (A, V or $AV_{\Delta=0^\circ}$). To avoid participants comparing their performance across conditions absolute scores were not provided.

The second of the two training sessions started with a short repetition of the training (20 trials x 12 individualized probe locations x 3 conditions); in two blocks per condition (A, V, and $AV_{\Delta=0^\circ}$). Thereafter, participants repeated the calibration and baseline measurements of session 1 because the performance for each of the conditions may have changed as a result of the training. JND_A was acquired to reset S_{pr} and ΔAV (Section 2.5.1.2), σ_{blob} was recalibrated for the new JND_A (Section 2.5.1.3), JND_V was obtained to confirm that the new reliabilities were sufficiently matched (Section 2.5.1.4) and $JND_{AV,\Delta=0^\circ}$ baseline reference was renewed (Section 2.5.1.5). These new S_{pr}, ΔAV and σ_{blob} were used in the final test session.

2.5.3. Final test session: The main experiment for all four groups was performed entirely in the final session (second session for groups 1 and 2; fourth session for groups 3 and 4). Only the data that was acquired in this final session was used for the analyses on MLE parameter comparison (Section 2.8) and Bayesian model comparison (Section 2.9).

Main experiment: Participants completed 520 trials (40 trials x 13 individualized probe locations) for each of the five conditions (A, V, $AV_{\Delta=0^\circ}$, $AV_{\Delta=+X^\circ}$, $AV_{\Delta=-X^\circ}$), divided into 10 blocks per modality (A, V or AV) that were presented in pseudorandom order. The AV blocks were three times longer than the unisensory blocks because spatially congruent and incongruent trials were presented intermixed.

Performance dependent reward: Before starting the main experiment participants in groups 2 and 4 were informed that they were eligible to win a monetary reward (£10) if their overall percentage correct (across all locations) for the A, V and $AV_{\Delta=0^\circ}$ signal responses was greater in the final session than in their baseline (i.e. the introductory session for group 2, the second training session for group 4), consistently across all three conditions. Importantly, participants were not aware of any potential reward during the introductory/training sessions.

2.6 Fitting psychometric functions

Participants proportions of 'perceived right' across the thirteen probe locations for one stimulus condition (e.g. A, V or $AV_{\Delta=0^\circ}$, see Sections 2.5.1.2, 2.5.1.4, and 2.5.1.5) were summarized by means of fitting a cumulative Gaussian function:

(1) $$\psi(S_{pr}) = \lambda + (\frac{1}{2} - \lambda)\left[1 + \mathrm{erf}(\frac{S_{pr}-PSE}{JND\sqrt{2}})\right]$$

where PSE is the point of subjective equality (i.e. the mean of the Gaussian; $\psi(PSE) = 0.5$) and λ is the lapse rate (i.e. the probability that a participant makes a mistake independent of probe location S_{pr}, for example due to blinking, inattention, erroneous button press, etc.). The just noticeable difference, JND, represents the standard deviation of the noise across stimuli differences (i.e. probe relative to standard). Please note that, since the sensory noise of both stimuli contribute equally to the JND in 2IFC tasks, the JND equals $\sigma\sqrt{2}$; where σ is the standard deviation of the Gaussian-distributed sensory noise for the particular sensory signal (Kingdom & Prins, 2016). N.b. In Sections 2.8-9 and in the Results Section 3 we refer to $JND/\sqrt{2}$ as σ_{emp} (where emp stands for empirical).

In addition to the three parameters that define the psychometric function (JND, PSE, λ), we also fitted a fourth parameter, η, that represents a normalized value (between 0 and 1) for the amount of additional noise in the system, e.g. accounting for participants getting tired or otherwise changing their task performance over the course of the 2.5 hour session. In other words, we made use of the betabinomial model for fitting of psychometric functions (Schütt, Harmeling, Macke & Wichmann, 2016). The four parameter values were optimized (i.e. 'fit') by maximizing the likelihood (L) of observing participants' responses given the model:

$$(2)\ L_{condition} = \prod_{i=1}^{N} \frac{B(k_i + \eta'\bar{p}_i,\ n_i - k_i + \eta'(1-\bar{p}_i))}{B(\eta'\bar{p}_i,\ \eta'(1-\bar{p}_i))} \qquad \text{with } \eta' = \frac{1}{\eta^2} - 1$$

where $\bar{p}_i = \psi(S_{pr,i} | JND, PSE, \lambda)$ is the expected probability of observing a 'right' response given probe location $S_{pr,i}$ and parameter values for JND, PSE and λ (see Eq. 1); k_i is the empirical number of 'right' responses out of n_i trials, and N is the total

number of probe locations. Capital letter B represents the beta function. To find the maximum likelihood estimate (for values of JND, PSE, λ and η) we made use of the Nelder-Mead optimization algorithm as implemented in Palamedes toolbox 1.8.2 (www.palamedestoolbox.org).

2.7 Analysis Overview

In a previous study, with an experimental design that was very similar to the current methods, we found that behavioral responses of many participants deviated from MLE-predicted audiovisual spatial integration (Meijer & Noppeney, 2018). In this study our aims were two-fold: 1) To investigate whether we can influence participants' behavior with training or reward to behave more like statistically optimal 'ideal observers' that integrate multisensory signals according to MLE principles. 2) To assess which of a subset of previously proposed explanations for deviation from MLE predictions provides the best fit to participants' empirical behavior. To answer these two research questions we made use of two different analysis methods: 1) Parameter comparison according to classical statistical procedures in line with prior literature on audiovisual integration whereby we compared differences between empirical and MLE-predicted parameters (JNDs and sensory weights) across the four experimental groups in a 2x2 between-subjects factorial design, specifically looking at the main effects of training and reward (Section 2.8). 2) We applied Bayesian model comparison techniques to assess which of six alternative models best fitted the behavioral response data. The first of these models assumes that participants performed multisensory integration according to MLE, whereas the five other models introduce variations to the MLE computational rules that

may or may not be statistically optimal given certain circumstances (e.g. when assumptions of the MLE model are violated; Section 2.9).

2.8 Empirical vs. MLE-predicted parameter comparison

In order to obtain empirical parameter estimates we jointly fitted five psychometric functions to participants' responses for the five conditions of the main experiment (A, V, $AV_{\Delta=0^\circ}$, $AV_{\Delta=+X^\circ}$ and $AV_{\Delta=-X^\circ}$; Section 2.5.3). To avoid biases of the σ_{emp} estimates by different lapse rate estimates across conditions (see Wichmann & Hill, 2001a) we constrained the five lapse rates (λ, Section 2.6) to be equal. In other words, we assumed that the proportion of location-independent mistakes is comparable across conditions. Sharing one lapse rate across five conditions has the additional advantage that the estimate is more reliable because it is based on a greater number of trials (Kingdom and Prins, 2016). Similarly, we only fitted one betabinomial noise factor η. Furthermore, σ_{emp}s of both AV-incongruent conditions were constrained to be equal, because we expected no differences between them (n.b. in earlier studies a mean multisensory σ_{emp} was commonly computed, e.g. Alais & Burr, 2004). However, the AV-congruent σ_{emp} was allowed to differ from the AV-incongruent σ_{emp} in order to investigate whether the audiovisual disparity (albeit small) affects multisensory integration. Given those constraints we jointly optimized parameter values for five PSEs (one for each condition), four σ_{emp}s (A, V, $AV_{\Delta=0^\circ}$, $AV_{\Delta=\pm X^\circ}$), one lapse rate (λ) and one betabinomial noise factor (η) by maximizing the overall likelihood across all five conditions (i.e. the product of the condition-specific likelihoods: $L_{overall} = \prod_{c=1}^{5} L_{condition,c}$; see Eq. 2) making use of the Nelder-Mead optimization algorithm (as above).

The empirical auditory weight was computed from the difference between the two AV-incongruent PSEs relative to the individualized audiovisual disparity according to (Fetsch, Pouget, DeAngelis & Angelaki, 2011; Meijer & Noppeney, 2018):

(3) $w_{A,emp} = \frac{PSE_{\Delta AV=+X^\circ} - PSE_{\Delta AV=-X^\circ}}{2*|\Delta \mathrm{AV}|} + \frac{1}{2}$

Using empirically determined σ_A and σ_V we computed MLE-predictions for σ_{AV} and w_A (i.e. parameter values that are to be expected if participants integrated the audiovisual signals according to maximum likelihood estimation; Ernst & Banks, 2002):

(4) $\sigma_{AV,mle} = \sqrt{\frac{{\sigma_V}^2 * {\sigma_A}^2}{{\sigma_V}^2 + {\sigma_A}^2}}$ and $w_{A,mle} = \frac{{\sigma_V}^2}{{\sigma_V}^2 + {\sigma_A}^2}$

Critically, the MLE model assumes that audiovisual signals are fused into a single unified percept independent of whether the audiovisual disparity is present. In other words, it ignores the so-called 'causal inference problem' for the AV conflict trials (Shams & Beierholm, 2010) and assumes that σ_{AV} is identical for spatially congruent and incongruent stimuli (i.e. the forced fusion assumption).

We compared the empirical auditory weight ($w_{A,emp}$) and audiovisual sensory noise parameters (σ_{emp}) to the respective MLE-predicted parameters for each participant. To visualize the effect size of such differences at the individual level we parametrically bootstrapped 95% confidence intervals for each parameter estimate using the following procedure (Wichmann & Hill, 2001b; Kingdom & Prins, 2016; Meijer & Noppeney, 2018): Using each observer's empirically determined parameter values and assuming the betabinomial generative model (Schütt, Harmeling, Macke & Wichmann, 2016) we simulated 5000 datasets, i.e. proportions of 'right' responses across the individualized

probe locations, for all five conditions. We then jointly fitted five psychometric functions (as above) to each simulated dataset and computed the empirical auditory weights and MLE-predictions according to equations 3-4. The 95% confidence interval for each parameter was defined as the interval between the 2.5 and 97.5 percentiles across all 5000 bootstrapped parameter estimates.

The main interest of this parameter-based analysis was to investigate whether training and motivation (i.e. reward) affected the differences between empirical and MLE-predicted parameter values. To test for such main effects we performed group-level 2x2 ANOVAs on these differences for σ_{AV} and w_A, categorizing the parameter values according to the four experimental groups. Additionally, in order to quantify evidence against the main effects of training and/or reward (i.e. in favour of the null-hypotheses stating there are no such effects) we also computed Bayes factors between the full model (that included all factors) and reduced null models (where one factor was not taken into account) (Rouder, Morey, Speckman & Province, 2012; as implemented in BayesFactor Package 0.9.12-2, using Cauchy priors for standardized effect sizes and default settings; http://bayesfactorpcl.r-forge.r-project.org/; computations in R 3.4.1). $BF_{01} > 1$ denotes the Bayes Factor in favor of the null model (i.e. no effect), while $BF_{10} > 1$ denotes its inverse where model evidence suggests an effect of the particular factor.

2.9 Bayesian model comparison

Using Bayesian model comparison we contrasted the MLE model with five competitor models that embody different multisensory decisional or perceptual strategies. All six models jointly predict the behavioral response pattern for the five empirical conditions

in terms of the proportions of 'perceived right' ($\bar{p}_i$) per probe location ($S_{pr,i}$), enabling us to compute the overall likelihood similar to the above-described psychometric function-based approach (Eq. 2). The main difference with the method above is that these six models constrict predictions for AV responses almost entirely based on unisensory parameter values (PSE_A, PSE_V, σ_A and σ_V) and model-specific computational rules that prescribe how two unisensory signals are combined to reach 'left'/'right' decisions on AV trials. Specifically, four of the audiovisual parameters that were fitted in Section 2.8 ($PSE_{AV,\Delta=+X^\circ}$, $PSE_{AV,\Delta=\pm X^\circ}$, $\sigma_{AV,\Delta=0^\circ}$ and $\sigma_{AV,\Delta=\pm X^\circ}$) were not included as free parameters in any of these six models. Nevertheless, all models contain one free parameter for the audiovisual bias: i.e. PSE_{AVc} (where the 'c' stands for congruent, i.e. $\Delta AV = 0^\circ$, but we note that this audiovisual left/right bias is also applied in the AV-incongruent conditions). Furthermore, all models contain one lapse rate (λ) and one betabinomial noise factor (η) as free parameters that are shared across the five conditions. Finally, four models include an additional eighth parameter that is specific to those models (see Table 1 for an overview).

For all models, the unisensory (A, V) expected proportions of 'right' responses ($\bar{p}_i$ for $S_{pr,i}$, Eq. 2) were computed directly using the formula for cumulative Gaussians (Eq. 1). However, since no such analytical solution for response probabilities $\bar{p}$ is known for the Bayesian causal inference model (Section 2.9.5; but see Körding, Beierholm, Ma et al., 2007) they were approximated for AV conditions by means of simulating 1000 trials (Monte Carlo sampling) per individualized probe location in all six models (for reasons of consistency).

We now first describe the computational details for each of the six models, before we elaborate on the methods that were used to compare the models' predictive performances (Section 2.9.10).

2.9.1. Maximum likelihood estimation (MLE)

For any audiovisual trial, noise-disturbed auditory and visual sensory signals, x_A and x_V, were independently sampled from normal distributions centered on the true (individualized) stimulus location, separately for standard and probe (with $S_{pr,A}$ and $S_{pr,V}$ not necessarily equal):

(5) $x_{st,A} \sim N(S_{st}, {\sigma_A}^2)$, $x_{st,V} \sim N(S_{st}, {\sigma_V}^2)$, $x_{pr,A} \sim N(S_{pr,A}, {\sigma_A}^2)$, $x_{pr,V} \sim N(S_{pr,V}, {\sigma_V}^2)$

For both standard and probe, these unisensory signals were then integrated according to MLE-prescribed reliability-weighted summation (with w_A computed via Eq. 4; $w_V = 1 - w_A$):

(6) $\hat{S}_{st,AV} = w_A x_{st,A} + w_V x_{st,V}$ and $\hat{S}_{pr,AV} = w_A x_{pr,A} + w_V x_{pr,V}$

In order to account for left/right biases we subtracted PSE_{AVc} from the difference of the two spatial estimates: $\widehat{\Delta S}_{AV} = \hat{S}_{pr,AV} - \hat{S}_{st,AV} - PSE_{AVc}$. A 'right' response was assigned when the estimated difference $\widehat{\Delta S}_{AV}$ was greater than zero.

This procedure was repeated for each of 1000 simulations, after which the probability of a 'right' response could be computed. Finally, the expected proportion of 'right' responses was obtained by adjusting for the lapse rate (λ):

(7) $\bar{p} = \lambda + (1 - 2\lambda) p('right')$

For each AV condition, $\bar{p}$ was separately computed for all thirteen individualized probe locations.

2.9.2. Visual reliability increase

One of the key assumptions that we and other researchers have made when validating MLE as a mechanism for multisensory integration via psychometric function fitting (e.g. Section 2.8) is that the sensory reliabilities do not change when the signals are presented as bisensory stimuli. However, this assumption might be violated. For example, low-level auditory-to-visual cross-modal modulation may boost the visual signals' salience prior to spatial integration leading to an increased reliability of the visual signals (Meijer & Noppeney, 2018; see also Bejjanki et al., 2011 for an alternative reliability change of the visual signal in audiovisual speech integration).

In the 'V reliability increase' model we introduced a cross-modal modulation factor, $c_{A\rightarrow V}$, that decreases visual sensory noise under AV conditions: $\sigma_{V,AV} = c_{A\rightarrow V}\,\sigma_V$ (where $c_{A\rightarrow V} \in [0,1]$). We assumed that observers are aware of such a change in reliability and use $\sigma_{V,AV}$ to compute the sensory weights (Eq. 4) for reliability-weighted integration (Eq. 6). All other model computations are similar to the MLE model (Section 2.9.1).

The effect of an increased visual reliability under AV conditions, which observers have correctly taken into account but researchers were unaware of, is that the empirical auditory weight is smaller than the MLE-predicted auditory weight (i.e. apparent visual overweighting). However, contrary to other variations of the MLE-model that are discussed below, the empirical multisensory variance in such cases should be lower than predicted by MLE: $\sigma_{AV,emp} < \sigma_{AV,mle}$.

2.9.3. Correlated noise

While the parameter-based method that we described in Section 2.8 takes into account some additional sources of measurement noise (e.g. by including lapse rate λ and betabinomial noise factor η) the empirically determined JND is assumed to solely depend on the amount of sensory noise ($JND = \sigma\sqrt{2}$). However, it is possible for additional supra-modal sources of noise (e.g. an unstable amodal spatial reference frame) to increase the variance of the spatial estimates and so contribute to the JND (Burr et al., 2009; Mueller & Weidemann, 2008). Likewise, the final spatial estimates, $\hat{S}_{st}$ and $\hat{S}_{pr}$ could be based on a sample of the posterior rather than the full distribution (Battaglia et al., 2011). Such post-integration perturbations can be viewed as a special case of correlated noise and violates the MLE assumption for independent noise in each modality. A statistically optimal observer would estimate the amount of correlated noise and take it into account when computing reliability-based sensory weights (Oruç, Maloney & Landy, 2003):

(8) $w'_A = \frac{{\sigma_V}^2 - \rho\ \sigma_A\ \sigma_V}{{\sigma_V}^2 + {\sigma_A}^2 - 2\ \rho\ \sigma_A\ \sigma_V}$ and $w'_V = 1 - w'_A$

In the 'correlated noise' model we jointly sampled x_A and x_V from a bivariate Gaussian distribution with covariance $\rho\sigma_A\sigma_V$; separately for standard and probe (n.b. correlation ρ is a free parameter). We assumed that observers were able to accurately estimate ρ and performed reliability-weighted integration utilizing the corrected weights, w'_A and w'_V, to obtain $\hat{S}_{st,AV}$ and $\hat{S}_{pr,AV}$ (Eq. 6). Further computations to obtain the expected proportions of 'right' responses $\bar{p}$ were identical to the MLE model.

If an observer optimally applied the corrected weighting scheme for multisensory signals with correlated noise, but a researcher ignorantly compared empirical with MLE-predicted parameter values, then (s)he would find that the more reliable modality is overweighed: i.e. $w'_A < w_{A,mle}$ if $\sigma_V < \sigma_A$, and vice versa. Moreover, (s)he would find that: $\sigma_{AV,emp} > \sigma_{AV,mle}$ (Oruç, Maloney & Landy, 2003).

2.9.4. Visual reliability prior

If observers are unable to accurately estimate the reliability of each sensory stimulus, then their behavior is likely to deviate from MLE predictions. As an explanation for visual overweighting (relative to MLE) Battaglia et al. (2003) proposed that observers made use of a prior probability distribution to estimate the visual reliability. This prior favors smaller estimates of σ_V because the visual modality is normally highly reliable for spatial localization. The biasing effect of such a prior would be greater for visual stimuli whose reliability is difficult to estimate; for example because of unfamiliarity or complex designs (Beck et al., 2012).

We here adopted the suggestion for such Bayesian estimation of the visual sensory reliability by allowing the estimate to be smaller than the true reliability: $\hat{\sigma}_V \leq \sigma_V$. Implementation was performed by addition of free parameter $\pi_{\hat{\sigma}V}$, constrained to the interval $[0,1]$; where $\hat{\sigma}_V = \pi_{\hat{\sigma}V}\, \sigma_V$. Trial simulations used the true σ_V to randomly sample sensory noise (Eq. 5), but the biased estimate $\hat{\sigma}_V$ was used to compute the sensory weights (Eq. 4) for audiovisual integration (Eq. 6). Other computations were similar to the MLE model.

Contrary to the correlated noise model, the 'V reliability prior' model can only inflict visual overweighting (relative to MLE; i.e. not auditory overweighting). As for any deviation from MLE-weights (Eq. 4), the bisensory variance reduction is statistically suboptimal: $\sigma_{AV,emp} > \sigma_{AV,mle}$.

2.9.5. Bayesian causal inference (BCI)

On many, if not most, occasions in daily life two sensory signals should not be integrated, even when they are received simultaneously. Yet, the MLE model assumes that all multisensory signals are integrated (i.e. forced fusion). Bayesian causal inference (BCI; Körding, Beierholm, Ma et al., 2007; Rohe & Noppeney, 2015a) allows observers to assess whether to integrate or segregate multiple sensory signals on each occurrence by computing the probability that the sensory signals originate from a common source:

$$(9)\ P(C=1|x_A,x_V) = \frac{P(x_A,x_V|C=1)*P_{common}}{P(x_A,x_V|C=1)*P_{common}+P(x_A,x_V|C=2)*(1-P_{common})}$$

P_{common} represents a participant's prior belief that two sensory signals are caused by the same event. In this study we have attempted to create conditions where $P_{common} \approx 1$, but the BCI model allows this parameter to be fit freely for each participant. The terms $P(x_A,x_V|C=1)$ and $P(x_A,x_V|C=2)$ are the likelihoods for observing x_A,x_V given either one or two sources; they depend on the sensory noise parameters, σ_A and σ_V, as well as on a spatial prior distribution, $P(S) = N(\mu_p,{\sigma_p}^2)$. We set $\mu_p = 0°$ and $\sigma_p = 100°$ in our simulations; thus essentially assuming an uninformative spatial prior. For computational details of the BCI model we refer the interested reader to Körding, Beierholm, Ma et al. (2007).

On each AV trial simulation we sampled x_A and x_V as before (Eq. 5) and computed spatial estimates for both causal hypotheses (i.e. one or two sources): $\hat{S}_{AV,C=1}$ is the integrated spatial estimate using MLE-optimal weights (Eqs. 4 & 6), while $\hat{S}_{A,C=2} = x_A$ and $\hat{S}_{V,C=2} = x_V$ are the segregated spatial estimates (n.b. negligible effects of the near-uniform spatial prior were not ignored in the actual computations but are not mentioned here for sake of brevity). Simulated BCI observers then utilized the so-called 'model-averaging' decision function (Rohe & Noppeney, 2015a) to obtain statistically optimal spatial estimates (i.e. most accurate across many occurrences) by computing the probability-weighted mean of the integrated and segregated estimates:

$$\hat{S}_{V,BCI} = P(C = 1|x_A, x_V)\hat{S}_{AV,C=1} + (1 - P(C = 1|x_A, x_V))\hat{S}_{V,C=2} \quad (10)$$

and likewise for auditory estimates $\hat{S}_{A,BCI}$ (Körding, Beierholm, Ma et al., 2007). Critically, BCI observers thus ended up with two distinct location estimates for one audiovisual stimulus: $\hat{S}_{V,BCI}$ and $\hat{S}_{A,BCI}$. Similarly, there were two modality-specific estimates for the difference between probe and standard: $\widehat{\Delta S}_{V,BCI} = \hat{S}_{pr,V,BCI} - \hat{S}_{st,V,BCI} - PSE_{AVc}$ and $\widehat{\Delta S}_{A,BCI}$ (likewise). Occasionally, this resulted in BCI observers being forced to choose between a 'left' or 'right' response based on either auditory or visual estimate $\widehat{\Delta S}$. In the trial simulations here we follow the proposal by Meijer & Noppeney (2018) who argued that observers may have consistently provided responses based on visual estimates, since that could potentially explain visual overweighting in their behavioral data.

Finally, the probability of a 'right' response is computed across 1000 simulations for each probe location and the lapse rate is taken into account according to Eq. 7. While

the distribution of spatial estimates $\hat{S}_{pr,V,BCI}$ across trials is bimodal for $\Delta AV > 0°$ and $P_{common} < 1$, we note that the distribution of differences $\widehat{\Delta S}_{V,BCI}$ across trials for relatively small ΔAV is approximately normally distributed. Divergence from normality might plausibly be mistaken for measurement noise when fitting a cumulative Gaussian (n.b. similarly true for the cue-switching model, Sections 2.9.6). We further note that the JND of such a fitted psychometric function for spatially incongruent AV conditions should be greater than predicted by the MLE model if $P_{common} < 1$.

2.9.6. Cue switching

Finally, the cue switching model embodies a decisional strategy that does not integrate the unisensory signals, but instead selects one of the signals (A or V) as their final estimate on every trial in proportion to relative sensory reliabilities. This statistically suboptimal strategy may explain empirical findings that show sensory weighting according to their relative reliabilities identical to MLE, but without any multisensory benefit in terms of variance reduction: $\sigma_{AV} \geq min(\sigma_A, \sigma_V)$ (Ernst & Bülthoff, 2004).

To implement cue switching in AV trial simulations we computed two modality-specific estimates for the difference between probe and standard: $\widehat{\Delta S}_A = x_{pr,A} - x_{st,A} - PSE_{AVc}$ and $\widehat{\Delta S}_V = x_{pr,V} - x_{st,V} - PSE_{AVc}$, where noise-perturbed signals x_A, x_V were randomly sampled according to Eq. 5. MLE-optimal sensory weights (Eq. 4) then determined the probability with which either estimate was selected: $\widehat{\Delta S}_A$ or $\widehat{\Delta S}_V$. The expected proportion of 'right' responses across simulations was computed following Eq. 7.

Model Name	**Overweighting**	σ_{AV} **Prediction**	**8th Parameter**
MLE	No	$\sigma_{AV,emp} = \sigma_{AV,mle}$	No
V reliability change	Vision	$\sigma_{AV,emp} \leq \sigma_{AV,mle}$	$c_{A \to V} \in [0,1]$
Correlated noise	Most reliable modality	$\sigma_{AV,emp} \geq \sigma_{AV,mle}$	$\rho \in [0,1]$
V reliability prior	Vision	$\sigma_{AV,emp} \geq \sigma_{AV,mle}$	$\pi_{\hat{\sigma}V} \in [0,1]$
BCI	Vision	$\sigma_{AV,emp} \geq \sigma_{AV,mle}$	$P_{common} \in [0,1]$
Cue switching	No	$\sigma_{AV} \geq min(\sigma_A, \sigma_V)$	No

Table 1 - An overview of the six models for Bayesian model comparison with regards to expected sensory overweighting (relative to MLE) and the predicted amount of audiovisual sensory noise. All six models contain the following seven free parameters: σ_A, σ_V, PSE_A, PSE_V, PSE_{AVc}, λ, η. Four models contain an additional eighth free parameter.

2.9.10. Model comparison:

Simply put, the aim of this model comparison analysis was to select the model with the highest likelihood of generating behavioral responses such as the ones that were acquired in this study. To control for model overfitting on this particular data set we computed an approximation of the cross-validated likelihood for each model using the following procedure:

Parameter values for each of the six models (per participant) were first optimized using maximum likelihood estimation as implemented in the Bayesian Adaptive Direct Search algorithm (BADS v1.0.5; Acerbi & Ma, 2017; https://github.com/lacerbi/bads). Thereafter, we performed Markov Chain Monte Carlo (MCMC) simulations by which we obtained an estimate of the entire posterior probability distribution across all parameters. The here employed Ensemble Inversion Slice Sampling algorithm (Acerbi, Dokka, Angelaki & Ma, 2018; https://github.com/lacerbi/eissample) is an affine-invariant MCMC sampler that requires relatively little tuning. We used a default number of chains per ensemble ($2p + 2$, with p the number of free parameters of the model), where each chain was initialized to the maximum likelihood estimate plus some random jitter. Prior distribution functions were flat but bounded within a reasonably wide interval for each parameter. We acquired 10,000 samples for each posterior distribution, after discarding twice that number for warm-up (i.e. burn-in) and further discarding 9/10 samples to reduce autocorrelation of the chains (i.e. thinning). Convergence of the chains was checked (for both mixing and stationarity; Gelman et al., 2013) by splitting each chain in half (thus obtaining $m = 4p + 4$ half-sequences), and comparing the across-chains estimate of the variance of the posterior distribution (for each parameter) with the mean of the variances for each half-sequence. Specifically, we computed the potential scale reduction factor: $\hat{R} = \sqrt{\widehat{var}_{pos} / \left(\frac{1}{m}\sum_{j=1}^{m} \widehat{var}_j\right)}$. Moreover, we computed the effective sample size $\hat{n}_{eff}$ for each parameter by correcting for its autocorrelation. If either $\max(\hat{R}) > 1.1$ or $\min\left(\hat{n}_{eff}\right) < 5m$ (Gelman et al., 2013), then we reran the MCMC simulation with increased warm-up period (10^5 samples) and more thinning (29/30); repeating if necessary until both criteria were met. While these tests do not

guarantee convergence, visual inspection of the marginal distributions did not indicate reasons for concern. Finally, having obtained an estimate of the multidimensional posterior we approximated the cross-validated leave-one-out log-likelihood (LL_{cv}) by means of Pareto-smoothed importance sampling (Vehtari, Gelman & Gabry, 2017; https://github.com/avehtari/PSIS).

A fixed-effects model comparison at the group level was performed by bootstrapping the sum across subjects of the differences between LL_{cv} of one model versus the LL_{cv} of a reference model (Adler & Ma, 2018a). Specifically, for each particular model, we first computed the LL_{cv} differences with the MLE model per participant. We then randomly sampled, with replacement, N of these differences and computed the sum (where N is the number of participants). We repeated this sampling procedure 10,000 times to obtain a distribution of summed LL_{cv} differences. The bootstrapped distributions for each model were then summarized by their medians and 95% confidence intervals. The fixed-effects analysis implicitly assumes that all participants utilized the same model (i.e. the same multisensory integration strategy).

Alternatively, Bayesian model comparison at the random-effects group-level allows heterogeneity between participants and selects the best fitting model based on the probability across participants that a certain model outperforms all other models. We computed the protected exceedance probabilities (PXP) and Bayesian omnibus risk (BOR) (Rigoux, Stephan, Friston & Daunizeau, 2014) as implemented in SPM12 (www.fil.ion.ucl.ac.uk/spm). PXP quantifies the probability for each model that its likelihood exceeds that of other models (corrected for chance), whereas BOR is a normalized Bayes factor that indicates whether there is evidence to belief that the

observed PXP differences are due to chance (e.g. $BOR < 0.25$ corresponds to $BF_{10} > 3$ and means that differences are probably not caused by chance).

To visualize the ways in which the various models modulated their parameters to match participants' behavioral data we summarized the expected response probabilities $\bar{p}(S_{pr})$ that were generated during the MCMC simulations by means of psychometric functions. Specifically, for each model and participant we randomly sampled 1000 parameter sets from the posterior distribution. For each of the parameter sets we i. generated response probabilities $\bar{p}$ according to the particular model's rules, ii. jointly fitted five psychometric functions (according to the methodology described in Section 2.6 and 2.8), and iii. computed the empirical auditory weight according to Eq. 3. We thus obtained 1000 x N estimates for each summary parameter (σ_A, σ_V, $\sigma_{AV,\Delta=0°}$, $\sigma_{AV,\Delta=\pm X°}$, $w_{A,emp}$). Next, we computed the group-level means for each of the 1000 samples per parameter. In a final step we then summarized the 1000 synthesized group means by their mean and standard deviation (i.e. mean and SEM) for each parameter (Acerbi et al., 2018; Adler & Ma, 2018a).

3. Results

Forty participants completed the main experiment in their final session of the study. Auditory and visual reliabilities were approximately matched for all participants in a separate session prior to the main task by modulating the size of the visual blob (i.e. larger blobs appear blurred and are more difficult to localize). Additionally, twenty participants underwent two training sessions in which they received feedback on nearly one thousand trials per modality: A, V, AV. Their reliabilities were re-matched after the

training. Finally, before starting the first trial of the main experiment half of the trained and half of the untrained participants (selected at random) were told that they could win a monetary reward if they improved their localization performance for all three modalities relative to the previous session (i.e. baseline measurements). Participants were thus effectively subdivided into four groups of N=10 whereby we manipulated training and reward (i.e. 'motivation') in a 2x2 between-subjects factorial design. (Please see Method Sections 2.3-5 and Figure 1 for details on the experimental procedure.)

We here first present results of the groups-based analysis for effects of training and reward on audiovisual spatial integration (see Section 2.8), before we evaluate participants' behavioral results by means of Bayesian model comparison (see Section 2.9).

3.1 Effects of training and reward

Efficacy of training was assessed by means of a one-sided paired t-test on the auditory JNDs that were obtained from the baseline measurements before and after training (i.e. in session 1 and 3; see Section 2.5.1.2 and 2.5.2). See Figure 2a. The auditory JNDs were significantly smaller after training, $t(19) = 2.5$, $p = 0.01$, thus indicating that participants had improved their auditory localization performance through training. We note that our experimental design did not allow us to test the efficacy of training in V and AV modalities because visual blob sizes were individually re-adjusted after the training. While it was important to provide equal training in all three modalities to avoid participants becoming disproportionately confident for auditory stimuli and therefore possibly changing their audiovisual response strategies (e.g. actively ignoring the visual stimuli), we also argue that the auditory training was most important for our purposes

because auditory spatial localization tasks may feel somewhat unnatural to first-time participants and extensive training may thus help participants to better judge their own auditory spatial localization abilities (and presumably be able to set the weights for audiovisual integration accordingly).

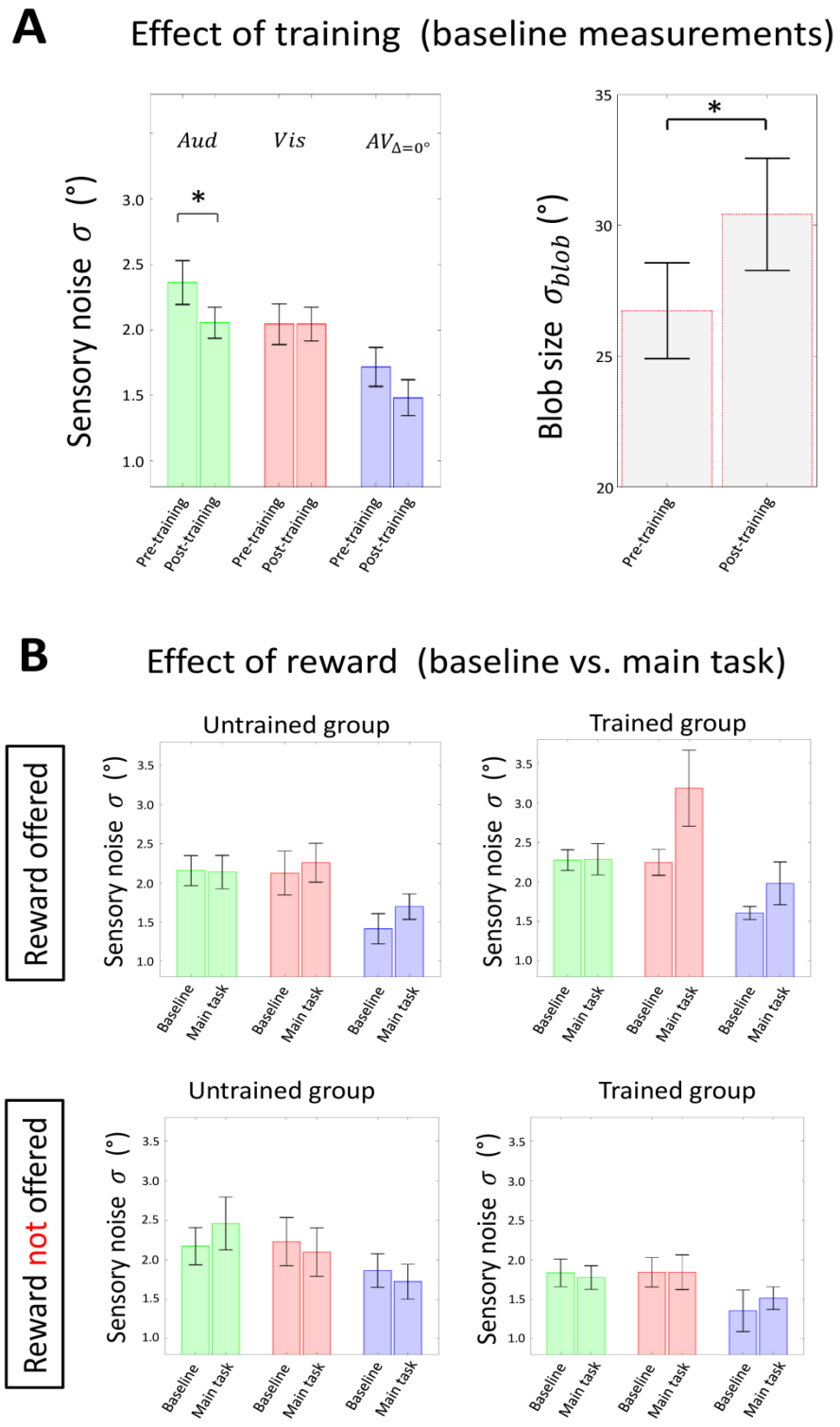


Fig 2 – Effects of training and reward on sensory noise parameters. Fig. 2a left-hand side: Mean (±SEM) fitted sensory noise parameters (i.e. behavioral variance; $\sigma = JND/\sqrt{2}$) across all twenty participants in the training groups, before and after training. Auditory sensory noise decreased significantly after training indicating the expected performance increase. Although visual and audiovisual sensory noise parameters have not changed significantly with training, the size of the blob had to be increased in order to equate visual performance to the auditory modality once again (see Section 2.5; i.e. more spatial blurring was applied, indicating that training had also led to a performance boost in the visual modality), see panel on right-hand side. Fig. 2b effect of reward: Fitted auditory (green), visual (red) and audiovisual (spatially congruent; blue) sensory noise parameters of baseline and main task measurements, for the four experimental groups. Performance did not increase for participants that were offered a conditional reward.

Efficacy of the offered reward, that should have motivated participants to aim for maximum performance in the main experiment, was assessed by comparison with A, V, and AV baseline performance measurements (acquired in the experimental session prior to the main experiment). See Figure 2b. Only four out of twenty participants managed to (numerically) improve performance in all three modalities and thus received the monetary reward. For a more detailed analysis we performed a 2 (training/no training) x 2 (reward/no reward) x 3 ($A/V/AV$) mixed ANOVA on the relative JND differences, $\frac{JND_{main} - JND_{baseline}}{JND_{baseline}}$, with data from all participants. We found no significant main effect for 'reward', F(1) = 3.34, p = 0.076. In fact upon closer inspection this 'nearly

significant' effect was caused by *decreased* performance in the visual and audiovisual modalities for participants who were offered the reward. We conclude that we were unable to demonstrate a modulatory effect of the motivational rewards on participants' performance (i.e. JNDs). Another way of assessing the effect of reward on participants' behavior was by comparison of the fitted lapse rates (λ) and betabinomial noise factors (η) in the main experiment (Sections 2.6 and 2.8) between the twenty participants who were/were not offered a potential reward: we would expect that motivated participants have smaller values for both. Indeed, we found that the group means for λ and η were numerically smaller in the reward-offered group, though neither difference was significant: $t(38) = 1.22$, $p = 0.11$ for λ and $t(38) = 0.26$, $p = 0.40$ for η (two-sample t-tests). However, we note that fitted values for λ and η approached zero for many participants in both groups (with and without reward), so the reason for not finding a significant difference might be a ceiling effect.

Having established a significant performance improvement for training, but not for reward, we then focused on our main question: whether any of the two interventions had affected multisensory integration in the main experiment. To obtain summary parameters of participants' behavior that we could compare to MLE-predicted parameter values, five psychometric functions were fitted to participants' empirical proportions of 'right' responses (40 trials x 13 locations for each of five conditions: A, V, $AV_{\Delta=0^\circ}$, $AV_{\Delta=+X^\circ}$, $AV_{\Delta=-X^\circ}$, where X° is the individualized audiovisual disparity for spatially incongruent trials, see Section 2.5.1.2). The results for each of the four experimental groups are shown in Figure 3.

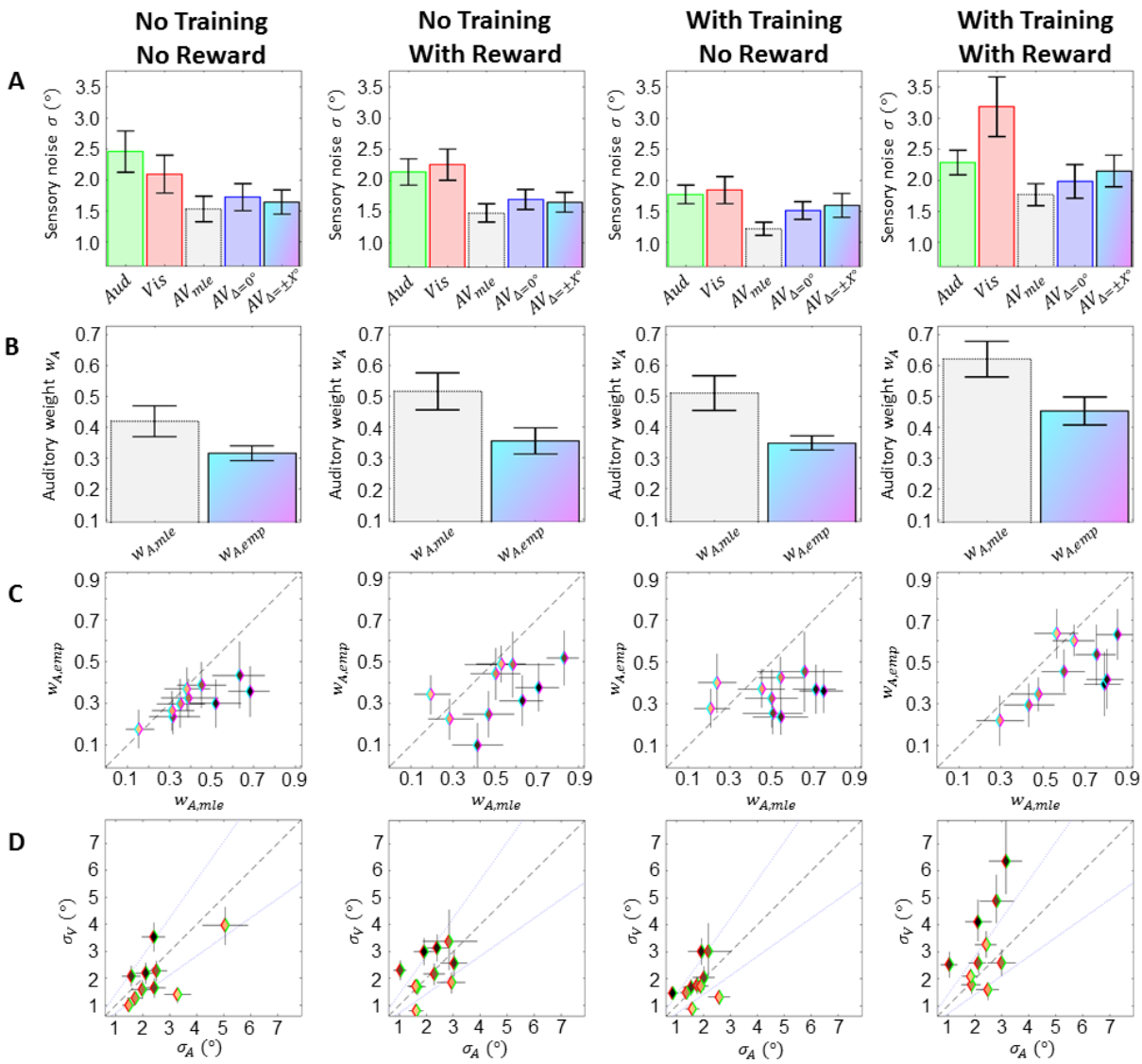


Fig 3 – Overview of the parameter-based results for the four experimental groups. Fig. 3 a-b. Mean fitted sensory noise parameters (i.e. behavioral variance; $\sigma = JND/\sqrt{2}$) and auditory weights per group ($\pm$SEM). MLE-predicted AV sensory noise and auditory weights were computed using σ_A, σ_V per participant (according to Eq. 4). Fig. 3 c-d. auditory weights (empirical vs. MLE-predicted) and unisensory noise parameters (σ_V vs. σ_A) for individual participants. Black solid lines depict bootstrapped 95% confidence intervals for each parameter estimate. Dashed black lines indicate equality (i.e. x=y). Participants were color coded by ranking them per group for the amount of

visual overweighting: i.e. dark shaded diamonds indicate participants with large auditory weight differences $w_{A,mle} - w_{A,emp}$. Color coding in panel D was consistent with that of panel C. While unisensory reliabilities (σ_A, σ_V) of all participants were matched within acceptable limits prior to the main experiment (see Section 2.5.1.4), for some participants they were no longer matched during the main experiment. These participants can be recognized in Fig 3d by diamonds that lie outside the acceptable range as depicted by the blue dotted lines.

Overall we observed a variance reduction for audiovisual relative to unisensory conditions across all four groups (Fig 3a), but the audiovisual variances are larger than predicted by MLE ($\sigma_{AV,mle} < \sigma_{AV,\Delta=\pm X^\circ}$) as confirmed by independent one-sided paired t-tests for each group: p < 0.05 for all. The deviation from MLE predictions is even stronger for the empirical sensory weights (Fig 3b): visual overweighting is observed in all four groups ($w_{A,mle} < \sigma_{AV,mle}$; one-sided paired t-tests for each group: t(9) > 2.9, p < 0.01). To test whether the factors 'training' or 'reward' affected the amount of suboptimal variance reduction and/or visual overweighting we performed two independent two-way ANOVAs on the differences between empirical and MLE-predicted parameters ($\sigma_{AV,mle} - \sigma_{AV,\Delta=\pm X^\circ}$, and $w_{A,mle} - w_{A,emp}$). We found no effect of reward for either parameter comparison (F(1) < 1, BF_{01} = 3.1 and F(1) < 1, BF_{01} = 2.7, respectively) and no significant interactions between reward and training. However, there was a significant main effect of training on the variance differences (F(1) = 5.6, p = 0.02, BF_{01} = 0.34); but not on the weights differences (F(1) < 1, BF_{01} = 2.7). Contrary to the hypothesis that training would help participants to behave more like MLE-optimal

ideal observers, we observed that trained participants exhibited greater suboptimality in terms of their audiovisual variance (at least for the spatially incongruent conditions; we found no such significant effect of training for the congruent condition: $\sigma_{AV,mle} - \sigma_{AV,\Delta=0°}$, F(1) < 1, BF_{01} = 2.9).

Rather than suggesting that prior training causes observers to integrate audiovisual signals less efficiently, we argue that the above-described finding is confounded by the fact that trained participants were more likely to have larger σ_V than σ_A (as demonstrated by an ANOVA on $\sigma_V - \sigma_A$; there was a significant difference between trained and untrained participants: F(1) = 4.3, p = 0.04; driven mainly by four observers in group 4: 'With Training, With Reward', see Fig. 3b). MLE predicts high auditory weights for $\sigma_V \gg \sigma_A$, but the data showed that all participants had relied on the visual signals substantially ($max(w_{A,emp}) = 0.64$, see Fig 3c). Color-coding participants by the amount of visual overweighting (ranking participants within each experimental group according to $w_{A,mle} - w_{A,emp}$) reveals that participants whose weights deviated most from MLE-predictions were participants whose visual sensory noise estimates were relatively high (throughout all experimental groups; see Fig 3d). Indeed, we found significant correlations (across all participants regardless of their experimental group) between the relative difference in unisensory noise ($\frac{\sigma_V - \sigma_A}{\sigma_A}$) and the extent to which participants deviated from MLE predictions in terms of audiovisual variance and visual overweighting (Pearson's r = 0.53 and 0.72, respectively; see Figure 4).

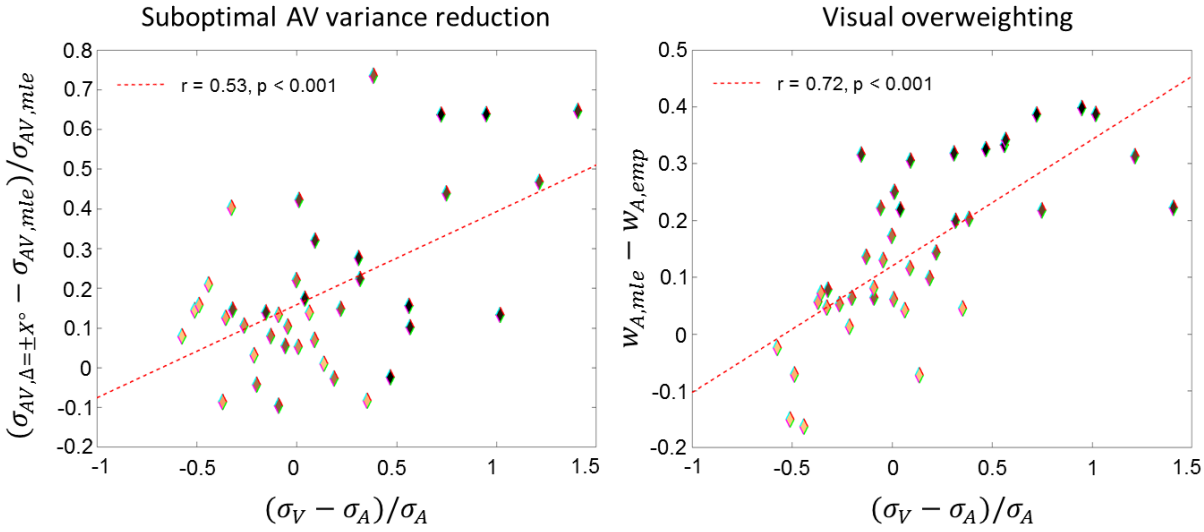


Fig 4 – MLE predictions deviate progressively from human audiovisual integration when auditory spatial signals are more reliable than visual spatial signals. Diamonds indicate individual participants (across all four experimental groups) with color coding identical to Fig. 3c-d: i.e. darker shading for participants who exhibit stronger visual overweighting (as ranked within each experimental group). Dashed red lines illustrate results of a linear regression. Pearson's correlation coefficients and their respective p-values are shown in the legend (top-left corners).

In conclusion, these behavioral data do not support the hypotheses that prior training or extra motivation leads to audiovisual spatial integration in line with maximum likelihood estimation. Instead, we observed evidence to suggest that divergence from MLE predictions was larger when the auditory reliability happened to be higher than the visual reliability.

3.2 Bayesian model comparison

Since the parameter-based analysis did not demonstrate meaningful modulatory effects of training or reward (see above), we pooled across all participants for the model-based analysis. The aim of this analysis was to compare several previously proposed explanations for findings that multisensory perceptual behavior diverged from MLE predictions (e.g. Battaglia et al., 2003; Burr et al., 2009; Meijer & Noppeney, 2018). Based on such proposals we designed six models (one of which implemented MLE) to generate synthetic 'left'/'right' responses for audiovisual trials under experimental conditions that were identical to those used for our individual participants (Section 2.9). This approach essentially enabled us to compare how well the model-specific artificially generated response patterns matched response patterns of human observers. The fit was quantitatively expressed for each participant by computing an approximation of the cross-validated log-likelihood (LL_{cv}, see Section 2.9.10) across all five conditions: A, V, $AV_{\Delta=0^\circ}$, $AV_{\Delta=+X^\circ}$, $AV_{\Delta=-X^\circ}$.

To visualize the extent to which the models were able to simulate human observers we summarized their generated data in the same way: by fitting psychometric functions and plotting the relevant parameter estimates. Fig. 5a (top rows) shows the results with regards to the sensory noise and auditory weights: The colored bars represent participants' group-level means $\pm$SEMs for reference; the black error bars on top show the models' simulated group-level means $\pm$SEMs (see Section 2.9.10 for methodological details). Four of the models contained an additional model-specific parameter; those are presented in the bottom row of Fig. 5a. We shortly discuss each model's fit to gain insight into each of the proposed explanations for human multisensory perception:

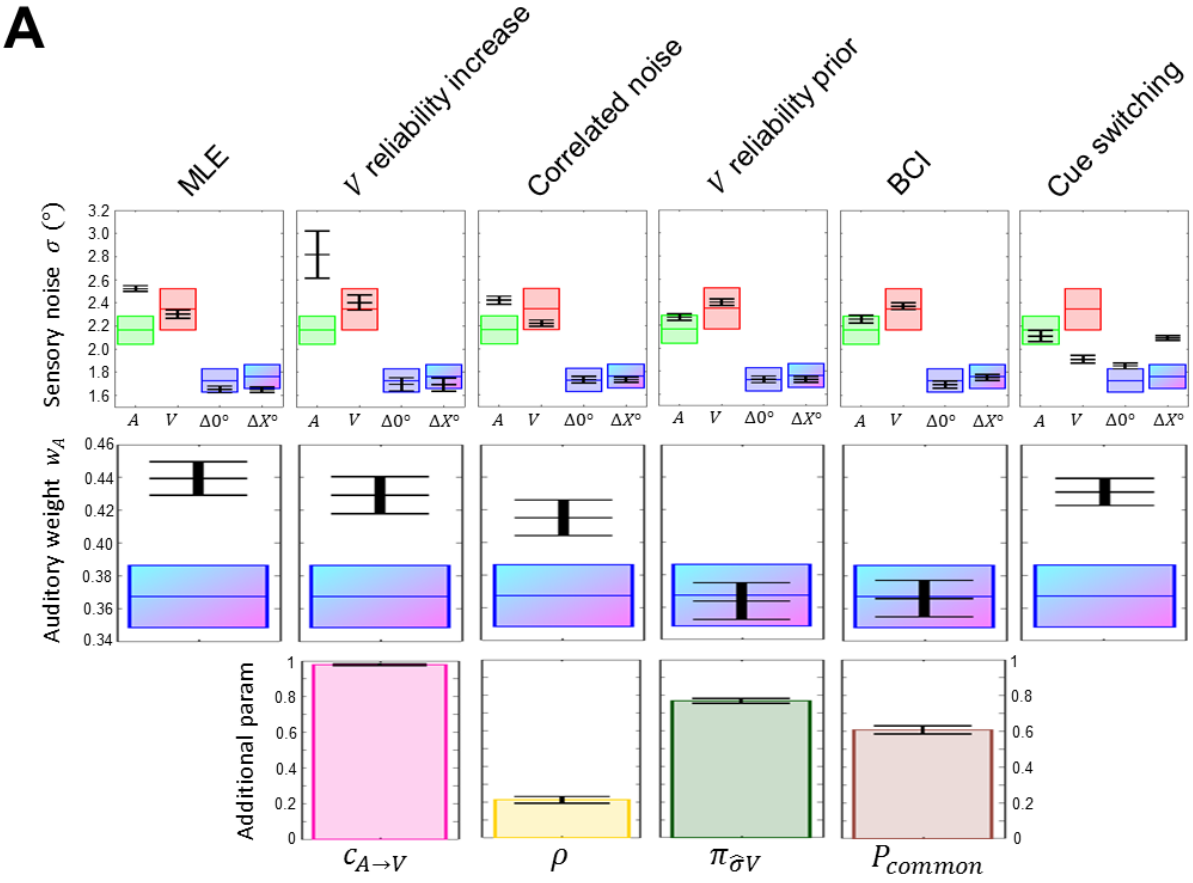
1. The MLE model predicts AV responses based on reliability-weighted averages of the unisensory estimates under forced fusion assumptions. In order to approximately fit human observer's visual overweighting pattern (i.e. to match the PSEs in the AV-incongruent conditions), it was forced to modify its unisensory noise parameters σ_A and σ_V. The model balanced this trade-off and ended up predicting neither of them perfectly.

2. The second model allowed visual sensory noise under AV conditions to be smaller than the unisensory visual noise: $\sigma_{V,AV} = c_{A\to V}\,\sigma_V$ (where $c_{A\to V} \in [0,1]$). By setting $c_{A\to V} < 1$ the model would be able to predict visual overweighting. However, if it did so, it would also predict σ_{AV} to be smaller than predicted by MLE. Since the latter would negatively affect the fit of all three AV conditions this was not preferred and $c_{A\to V}$ was instead set close to one, thereby making the model essentially similar to the MLE model. In an attempt to better account for visual overweighting, the second model further compromised its fit of the auditory variance.

3. Correlated noise leads to visual overweighting if the visual reliability is higher than the auditory reliability. By allowing some amount of correlated noise, $\rho > 0$, the third model was not forced to compromise on the unisensory noise levels as much as the two models discussed above.

4. The fourth model implements the suggestion by Battaglia et al. (2003) that visual overweighting can be explained by a Bayesian prior on the visual reliability estimates that favors high values (i.e. low $\hat{\sigma}_V$). We modelled this prior by introduction of $\pi_{\hat{\sigma}V}$, a factor between 0 and 1 that modifies the visual reliability estimate according to

$\hat{\sigma}_V = \pi_{\hat{\sigma}V}\,\sigma_V$. By setting $\pi_{\hat{\sigma}V} < 1$ the model was effectively capable of simulating visual overweighting without having to compromise fits of the unisensory noise estimates.



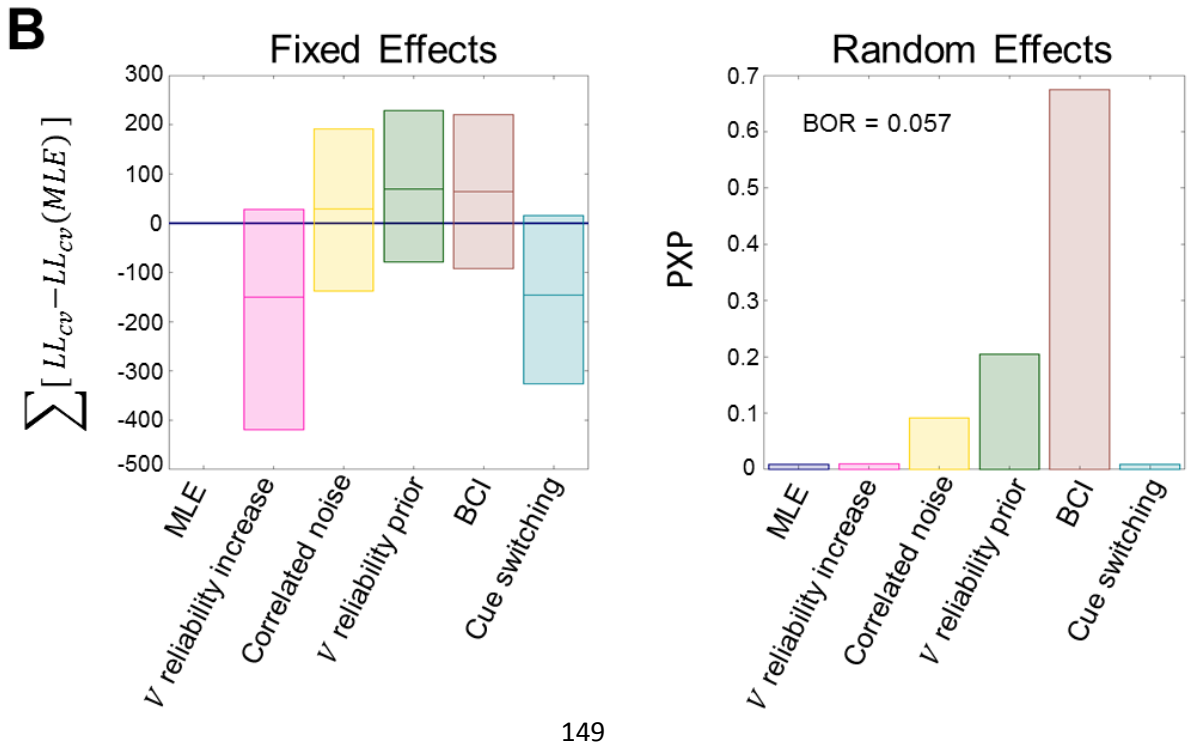


Fig 5 – Qualitative and quantitative model comparison results. Fig. 5 a. Psychometric functions were fitted to the model-generated proportions of 'right' responses to compare the empirical parameter estimates of our participants (colored bars; group mean ±SEM, same data as Fig 3 a-b but pooled across groups) to the fitted models' parameter estimates (black error bars). The third row depicts the model-specific additional parameter estimates (group-level means as colored bars, ±SEM as black error bars). Fig. 5 b. Quantitative model comparison results of fixed-effects (left panel) and Bayesian random-effects (right panel) analyses. See main text for details (Section 3.2).

5. Bayesian causal inference (e.g. Körding, Beierholm, Ma et al., 2007) is different from all of the above models because it does not assume forced fusion. Instead it modulates the prior probability of a common source by adjusting the free parameter P_{common} per participant (i.e. $P_{common} = 1$ leads to forced fusion, $P_{common} = 0$ leads to no fusion at all). Sensory signals on any one trial are only integrated to the extent that a common source can be inferred for them: i.e. they are partially integrated under what has become known as the 'model-averaging' decision function (e.g. Rohe & Noppeney, 2015a). Critically, in our implementation of the BCI model here we let the final 'left'/'right' response depend on the (partially-integrated) location estimates for the visual modality (as previously proposed in Meijer & Noppeney, 2018). This choice lets the BCI model simulate visual overweighting by setting common-source prior $P_{common} < 1$. The main difference with the 'visual reliability prior model' (Nr. 4) is that the BCI model predicts that the fitted JND of spatially incongruent conditions (i.e. $\sigma_{AV,\Delta=\pm X^{\circ}}$) is higher than the JND of the spatially congruent condition.

6. The sixth model implements cue switching: a strategy in which auditory and visual signals are not integrated. Instead, the model probabilistically selects either auditory or visual estimates in proportion to their reliabilities (Ernst & Bülthoff, 2004). Similar to the MLE model, cue switching will assign higher visual weights when $\sigma_V < \sigma_A$. Since the signals are not integrated, cue switching does not predict a multisensory variance reduction. In an attempt to approximately fit the low empirical audiovisual sensory noise and visual overweighting it set σ_V artificially low. Despite major deviations for the visual noise parameter the model is still incapable of substantially lowering σ_{AV} estimates in the spatially incongruent conditions.

The above-described visualizations of the model fits suggest that the 'V reliability prior' and BCI models performed best, on average across participants. A fixed effects quantitative model comparison confirms that suggestion (see Fig 5b, left panel). The group-level analysis was performed on the sum of the participant-specific differences between LL_{cv} and the LL_{cv} of the MLE model (nr. 1). To account for variance across participants we performed a bootstrap analysis of the summed differences and depict the median and 95% confidence intervals (Adler & Ma, 2018a). With summed LL_{cv} differences of 66 and 68, respectively, it is clear that the 'V reliability prior' and BCI models performed much better than MLE model. However, the 95% confidence interval sizes are relatively large, suggesting substantial variations of LL_{cv} across participants and models. This makes conclusive inference from the fixed-effects analysis risky, because few participants with large LL_{cv} differences might be driving group-level effects.

To better account for the variance across participants we performed a Bayesian random-effects analysis that is based on participant-specific probabilities that a certain model-evidence (LL_{cv}) exceeds that of all other models (Rigoux et al., 2014). The group-level protected exceedance probabilities (PXP) for each model are shown in Fig 5b (right panel). This analysis leads to a clear winner: the BCI model has the highest corrected-for-chance probability of being the best fitting model across participants: PXP = 0.675. Moreover, the Bayesian omnibus risk is low (BOR = 0.057), thus suggesting that the observed differences between PXP are very unlikely to be a coincidence.

4. Discussion

The objectives for this study were two-fold: we first examined the impact of prior training and motivational rewards on audiovisual spatial integration by comparing human behavioral results against MLE predictions in a 2x2 between-subjects factorial design. We found that neither of these factors influenced participants to behave more like MLE-optimal ideal observers: audiovisual integration diverged from MLE predictions in all four experimental groups independent of the manipulations. We then set out to compare various explanations for such divergence by means of Bayesian model comparison. The results from this second analysis were unambiguous: Bayesian causal inference provided the best fit to the data. It thus seems that the most likely reason participants did not integrate audiovisual stimuli according to MLE was that they did not automatically infer a common causal origin for the auditory and visual signals (i.e. the forced fusion assumption was violated).

Bayesian causal inference is a powerful explanatory model for multisensory integration that has successfully matched human perception across various combinations of sensory modalities and tasks, including audiovisual localization (Körding, Beierholm, Ma et al., 2007; Rohe & Noppeney, 2015a), audiovisual temporal judgments (Magnotti, Ma & Beauchamp, 2013; McGovern, Roudaia, Newell & Roach, 2016), audiovisual speech processing (Magnotti & Beauchamp, 2017), and recently even visual-vestibular heading estimation (for which it was previously thought that the brain performs forced fusion under all conflict sizes; Acerbi et al., 2018; de Winkel, Katliar & Bülthoff, 2017). Furthermore, there exists substantial evidence from neuroimaging studies that supports the notion that Bayesian causal inference is implemented in a hierarchical fashion by the human cortex (Rohe & Noppeney, 2015b, 2016). Given the abundance of studies that endorse BCI as a perceptual mechanism it may not be a surprise that this model performed best in the current study. However, the novelty here is that it fitted best in experimental conditions that were previously assumed to impose forced fusion (Alais & Burr, 2004). The finding that BCI explains human perceptual behavior better than MLE even under individually optimized 'forced-fusion conditions' provides a clear word of caution for future experiments that still wish to assume forced fusion, and may help to explain previously reported findings where human multisensory integration deviated from MLE predictions (see further below for examples and Rahnev & Denison, 2018 for a review).

In the current study, we made one critical assumption in the BCI model: that observers base their 'left'/'right' responses for audiovisual trials on partially-integrated location estimates in the visual sensory modality. On audiovisual trials where a common source

cannot be inferred with full certainty, disparate modality-specific signals are 'corrected' for the probability of a common source by applying partial reliability-weighted integration according to BCI, but an observer will be left with two spatial estimates that do not necessarily agree with each other: one auditory and one visual. What to do if only one 'audiovisual' response is required? Further combining the estimates is not appropriate, because that would imply certainty of a common origin, which was ruled out earlier. We argue that participants make a cognitive decision based on life-long experience to let their responses depend on the partially-integrated visual estimates because vision is generally most informative for localization.

Supporting evidence for such an experience-based decision strategy comes from a study by Jacobs and Fine (1999). Their participants performed a 2IFC depth discrimination task based on texture and motion cues (both in vision). By introducing various conflict sizes between the two cues reliability-weighted integration could be examined (similar to the current study). Participants performed two test sessions, each of which was preceded by a training session with auditory feedback. Crucially, the training trials differed from the test trials in that only one cue was informative with regards to the difference between probe and standard; the other cue was identical for both stimuli. In a cross-over design, the two training sessions alternated which of the cues was informative: texture or motion. Results of the two test sessions showed that equally-reliable cues were weighted heavier if that cue was informative in the preceding training session (as opposed to when it was uninformative in the preceding training session). We suggest that these empirical weight differences can be explained by partial integration under BCI

rules where participants selected the depth estimate of the cue that was most accurate in the preceding training session.

Direct support for the notion that observers are able to cognitively select the most relevant of the partially-integrated sensory signals comes from a study by Ernst and Banks (2000; but see Ernst, 2006). Their participants performed two versions of a 2IFC visual-tactile size discrimination task. In one version participants compared a bisensory visual-tactile stimulus (VT) in one interval to a unisensory visual stimulus (V) in the other interval. In the second version they compared VT against a unisensory tactile stimulus (T). Results indicated that the visual weights were higher in the VT vs. V task version and, vice versa, tactile weights were higher for VT vs. T. It is noteworthy that visual and tactile weights would be identical in both task versions under assumptions of MLE in forced fusion. Instead, partial integration according to BCI can explain the discrepancy if participants selected the most relevant size estimate for VT depending on the sensory modality against which it had to be compared.

Selecting one modality's partially-integrated estimates (over the other's) leads to overweighting of that modality relative to MLE predictions. In the current study, responding 'left' or 'right' based on visual estimates because that sensory modality is deemed most informative for spatial localization leads to visual overweighting when compared to MLE. This was a crucial assumption for the BCI model to fit our behavioral data well. However, it seems plausible that reliance on partially-integrated estimates of the sensory modality that is normally most informative is not limited to audiovisual spatial integration. The sensory modality that is reportedly overweighted in multisensory integration experiments is oftentimes the modality that is generally

dominant for the task at hand. For example, vestibular overweighting for visual-vestibular heading (Butler et al., 2010; Fetsch et al., 2009) but visual overweighting in visual-vestibular self-rotation tasks (Prsa et al., 2012), haptic overweighting in visual-haptic slant discrimination (Rosas et al., 2005) and auditory overweighting for audiovisual temporal discrimination tasks (Burr et al., 2009; Maiworm & Röder, 2011). Future research, applying similar hypothesis-driven (model) comparison methods, will be necessary to confirm whether the suggestion for such a general mechanism holds true across various combinations of sensory modalities.

Returning to the other findings that were presented in this study, we first note that the correlation between unisensory reliability differences and deviations from MLE predictions (see Figure 4) is a natural consequence of Bayesian causal inference: as the unisensory reliability difference increases the BCI model predicts fewer trials on which a common source is inferred (Rohe & Noppeney, 2015a). Interestingly, this correlation can also be observed in data that was previously presented as evidence to support the MLE model. For example, one of the figures in the seminal paper by Ernst and Banks (2002) shows that the visual modality (normally dominant in size estimation) was overweighted during visual-haptic integration in the condition where the visual noise was highest and where the difference between unisensory reliabilities was greatest ($w_{V,emp} \approx 0.3$, whereas $w_{V,mle} \approx 0.15$; their figure 3c).

The reason why larger unisensory reliability differences were more common for our trained participants is not immediately clear. We speculate that some of these participants were unable to preserve the performance improvement that they had gained in the training sessions, specifically for the visual sensory modality (Fig 2b shows

a visual performance decrease between baseline after training and main task). This may have happened because the maximum number of seven days that we allowed for between two consecutive sessions was quite a long period. Similarly, the extensive time gaps between consecutive sessions are a valid concern that could potentially explain why we did not observe an improvement of multisensory integration efficiency with training. Alais and Burr (2004) reported that their participants, who integrated audiovisual spatial signals according to MLE, were well trained on the localization task. So why did our trained participants, under similar experimental conditions, deviate so much from MLE? There might be other reasons for the observed behavioral discrepancy besides effective training. For example, one may argue that our participants integrated audiovisual signals to a lesser extent because the auditory signals were here presented through headphones, whereas Alais and Burr (2004) had used speakers that were placed at the edge of the screen. Our finding of a relatively low P_{common} (group mean $\approx$ 0.6, see Fig. 5a) would support such a hypothesis.

The fact that we did not observe a significant improvement of performance or lapse rate (i.e. measures that are independent of AV-integration strategies) as a consequence of the motivational reward is another limitation of the current study. The amount of the monetary reward (£10, i.e. 13 USD) may not have been high enough to motivate our participants to stay highly attentive for the 2.5 hour duration of the final test session. Alternatively, one may argue that participants in the non-reward groups were already motivated without the need for a monetary reward. Finally, we acknowledge that subtle changes that were possibly induced by our manipulations of training and reward may have failed to appear in analyses because our 2x2 between-subjects design with ten

participants in each group would have been underpowered to detect them with certainty. Nonetheless, we note that deviations from MLE predictions were independently significant for all four experimental groups.

In conclusion, our data have not shown evidence that prior training or motivational reward aid observers in perceptually behaving more like MLE-optimal observers, although we have discussed some limitations of our experimental design. We also evaluated various reasons for why the MLE model might not be the correct 'ideal observer' to compare behavioral data against. Comparison with our participants' responses suggested that Bayesian causal inference provides a better model for statistically optimal multisensory integration, because it is also valid when observers cannot be entirely sure that multisensory signals are caused by common events: i.e. in most daily life situations and in many laboratory experiments, even under so-called forced-fusion conditions.

CHAPTER 5

MULTISENSORY INTEGRATION BOOSTS CONFIDENCE AND DOES NOT INCREASE METACOGNITIVE NOISE

David Meijer, Jake Rhodes, Uta Noppeney

CONTRIBUTIONS

David Meijer and Uta Noppeney designed the experiment. David Meijer prepared the MATLAB scripts for stimulus presentation. Jake Rhodes acquired behavioral data. David Meijer performed the data analyses. David Meijer wrote the manuscript.

ACKNOWLEDGMENTS

We thank Steffen Bürgers for helpful discussions

This research was funded by ERC-2012-StG_20111109 multsens

ABSTRACT

One advantage of multisensory integration is that it reduces the amount of sensory uncertainty that is associated with perceptual estimates. While this statement is generally considered to be true because of the precision improvement that is observed across many trials, direct evidence on a per-trial basis is not readily available. Here we tested whether sensory uncertainty decreases after audiovisual spatial integration in a two-interval forced choice task by asking participants to report their confidence on each trial. Moreover, we studied whether participants' ability to introspect on stimulus uncertainty was different for multisensory and unisensory stimuli. To be able to answer both research questions while avoiding the risks for confounds we extended an existing model-based approach for estimation of metacognitive sensitivity (meta-d') such that it could be applied to experimental designs with psychometric functions. Results showed that participants' type-2 criteria were lower for more reliable stimuli, across three visual (and audiovisual) reliability levels. Importantly, criteria were also lower for audiovisual versus unisensory conditions thus demonstrating that multisensory integration boosted participants' confidence. After correcting metacognitive sensitivity levels for type-1 task performance, we concluded that participants' introspective abilities were unaffected by the sensory modality of the stimuli. Specifically, we found that the amount of metacognitive noise was equal for unisensory and multisensory conditions, and independent of stimulus reliability. These results support a hierarchical view of perceptual decision making in which multisensory integration occurs at an early stage and supra-modal metacognitive processes are able to access sensory uncertainty

similarly for unisensory and multisensory estimates to compute decision confidence at a later stage in the hierarchy.

1. Introduction

Sensory signals vary from very clear to clearly unreliable. Think of the difference between instantly recognizing your friend's face in a classroom and trying to figure out whether the light that shines through the morning fog is from an oncoming vehicle or an immobile street light. Such sensory (un)certainty plays a large role in our every-day decisions: Do we greet the person in class? Should we step aside for the light? To avoid costly mistakes by fully committing to potentially erroneous inferences about sensory signals it is essential that the human brain is able to assess sensory reliability and use it to attach appropriate levels of confidence to their perceptual decisions (Pouget, Drugowitsch & Kepecs, 2016). In simple perceptual experiments (where task difficulty is determined by sensory characteristics such as the signal-to-noise ratio) an increase of sensory uncertainty should lead to lower confidence, and vice versa. Indeed, it seems that humans are able to access sensory uncertainty through metacognitive mechanisms and use this knowledge to predict the validity of their perceptual decisions (Yeung & Summerfield, 2012). The extent to which observers are able to differentiate correct from incorrect responses in perceptual tasks, by means of high versus low confidence judgments, is called metacognitive sensitivity (Fleming & Lau, 2014).

Multisensory integration is one of the prime examples for showing how the perceptual system deals with sensory uncertainty through probabilistic inference (Pouget, Beck, Ma & Latham, 2013). The brain combines redundant sources of sensory information from

multiple modalities in a (nearly) statistically-optimal fashion by weighing each sensory signal by its reliability according to maximum likelihood estimation (MLE; Ernst & Banks, 2002). The benefit of such integration is that multisensory estimates are more precise, as can be measured empirically by a reduction in the behavioral variance across many experimental trials (Rohde, van Dam & Ernst, 2016). Importantly, MLE-like probabilistic integration not only predicts a behavioral benefit across trials, precision should also increase for each single-trial estimate. Presumably, multisensory integration thus reduces sensory uncertainty and should lead to increased levels of confidence. However, according to our current knowledge, empirical research that conclusively confirms this hypothesis has not yet been performed (Deroy, Spence & Noppeney, 2016). The current study addresses this fundamental question by investigating whether confidence is higher for audiovisual versus unisensory auditory and visual spatial localization responses.

If multisensory integration does indeed lead to a boost of confidence because of a reduction in sensory uncertainty, this would suggest that observers are able to estimate the sensory reliability of their integrated percepts. The logical second research question is whether the degree to which observers can use such metacognitive insight to their benefit by making appropriate confidence judgments is equal for unisensory and multisensorily integrated percepts (Deroy et al., 2016). In other words, we would wish to quantitatively compare metacognitive sensitivity for identical stimuli that are presented in a unisensory or a multisensory context.

Two methodological concerns need to be addressed when confronting these two research questions. 1) Mean confidence levels are confounded by perceptual accuracy (so-called type-1 task performance) and by metacognitive sensitivity (type-2

performance). The explanation for that statement is simple: mean confidence is generally higher for correct (as opposed to incorrect) responses, and this difference is (by definition) greater with higher metacognitive sensitivity. 2) Metacognitive sensitivity has a theoretical upper limit that is defined by type-1 performance. While the full explanation for this relationship is more complex (see Maniscalco & Lau, 2012, 2014) the intuitive understanding is that one cannot expect an observer that is relatively bad at performing the type-1 perceptual task to simultaneously be able to differentiate well between his/her correct and incorrect responses by means of high/low confidence judgments. While the second point of concern can potentially be avoided by ensuring that type-1 performance is identical for unisensory and multisensory conditions (e.g. by means of staircase procedures), the first point implies that mean confidence level is not a safe measure to use for assessing whether confidence increases with multisensory integration (unless one can ascertain that metacognitive sensitivity and type-1 accuracy are both equal across conditions). It would be better to take a different approach.

Maniscalco & Lau (2012) have introduced the concept of meta-d': a measure for expressing metacognitive sensitivity in terms of the type-1 performance (d') that a hypothetical type-2 ideal observer (who makes statistically-optimal confidence judgments based on signal detection theory) would need to have in order to best match a (true) observer's type-2 performance. In other words: they proposed to fit an ideal observer model to the confidence responses of a participant (conditional on type-1 responses, but independent from the participant's type-1 performance) with the ideal observer's perceptual performance as a free parameter termed meta-d' (where the prefix indicates that it is based on the participant's metacognitive sensitivity). While

meta-d' thus represents the observed metacognitive sensitivity of the participant, d' can be interpreted as the statistically optimal metacognitive sensitivity (given type-1 performance). The ratio between these two measures, $M_{ratio} = \frac{meta-d'}{d'}$, has become a common measure to quantify metacognitive efficiency (Fleming & Lau, 2014; Bang, Shekhar & Rahnev, 2018). The ratio will be exactly 1 for ideal observers, and lower for most human observers. By computing metacognitive efficiency index M_{ratio} one effectively overcomes the confounding effect of type-1 performance on metacognitive sensitivity by normalization (which has become possible since the two measures are expressed in common signal-to-noise units). From this methodological evaluation we conclude that metacognitive efficiency can be used to investigate potential differences between unisensory and multisensory metacognitive insight.

The meta-d' ideal observer model additionally fits as free parameters so-called type-2 criteria that determine the level of evidence that needs to be exceeded in order to make a higher confidence judgment. These type-2 criteria effectively model a participant's metacognitive bias (general under- or overconfidence): i.e. the fitted criteria will be greater (/smaller) if a participant requires more (/less) sensory evidence to make high confidence judgments. Such type-2 criteria thus hold potential for being used to answer our first research question: We expect type-2 criteria to be lower for multisensory as opposed to unisensory trials, because the sensory difference (e.g. between two stimuli in a two-interval forced-choice task) that is required to make a highly confident decision is smaller under conditions of reduced sensory noise. However, one major limitation of the meta-d' model fitting approach is that its parameters are normalized to signal-to-noise units of the hypothetical ideal-observer's type-1 performance. In other words, the

fitted type-2 criteria values depend on the observer's metacognitive sensitivity (by which they are normalized). Therefore, the interpretation of any type-2 criteria difference between conditions with unequal type-2 performance is troublesome.

Another limitation of the meta-d' approach is that the ideal observer can only be fit to one stimulus level. This requires the experimenter to carefully titrate type-1 performance for each participant to optimize chances of avoiding type-2 ceiling/floor effects: e.g. when the task is too easy (/hard) an observer will only respond with high (/low) confidence. It has been proposed that one could fit meta-d' to multiple stimulus levels separately, e.g. using an experimental design that is commonly used for fitting type-1 psychometric functions (Fleming & Lau, 2014), after which the multiple meta-d' estimates could be summarized by a linear regression analysis (e.g. see Klein, 2001). However, this may not be a practical approach as the range of stimulus levels on which metacognitive sensitivity can be assessed without meeting ceiling/floor complications is limited. Moreover, the proposed approach is likely theoretically invalid in asymmetrical experimental designs or when participants exhibit type-1 biases/priors (Drugowitsch, Moreno-Bote & Pouget, 2014).

To overcome these limitations we have developed a new approach that applies the rationale behind meta-d' to psychometric function-based experimental designs. Instead of using signal-detection theory to model an ideal observer, we used Bayesian probability theory to derive the statistically optimal observer's type-2 response probabilities. In analogy to meta-d', we express metacognitive sensitivity as the type-1 performance measure that the ideal observer would have needed to best match a participant's confidence responses across multiple stimulus levels: meta-JND (where JND

stands for the just-noticeable difference; Kingdom & Prins, 2016). By comparing the fitted meta-JND to an observer's type-1 JND, we can express metacognitive efficiency using a similar M_{ratio}. Moreover, meta-JND does not normalize its parameters to signal-to-noise units, such that they can be meaningfully interpreted and compared across different conditions with various metacognitive sensitivities. We can thus use meta-JND and its associated type-2 criteria to investigate whether multisensory integration affects confidence and/or metacognitive efficiency.

2. Method

2.1 Experiment

We employed an audiovisual spatial localization task for which we have previously established that a significant variance reduction can be observed for multisensory relative to unisensory stimulus conditions (Meijer & Noppeney, 2018; see also Alais & Burr, 2004). The experimental design was very similar to our previous study. We here present a summary that includes the most important modifications (for a detailed description of the experimental procedures see additionally Appendix A).

In a two-interval forced-choice (2IFC) task participants were asked to indicate whether they perceived the second of two consecutively presented stimuli to the left or to the right of the first stimulus (guessing if unsure). In addition to this type-1 location response, participants concurrently selected a confidence level for their location decision from a 4-point scale that was indicated as going from "guess" to "certain". They were actively encouraged to introspect, emphasizing accuracy over speed (maximum

response time was 3 seconds), and to use all four confidence levels. By means of a mouse response that allowed participants to simultaneously select a location response and a confidence level (see Figure 1) we have attempted to avoid discrepancies between type-1 performance and metacognitive sensitivity (n.b. previous studies that used sequential response designs have occasionally reported super-optimal type-2 performance, suggesting that the accumulation of evidence continues after type-1 responses have been made such that subsequent confidence judgments can be based on enhanced sensory information; Fleming, 2016; Murphy, Robertson, Harty & O'Connell, 2016; Siedlecka, Paulewicz & Wierzchoń, 2016; van den Berg et al., 2016; see also Kiani, Corthell & Shadlen, 2014).

The first stimulus (i.e. standard) was always presented straight ahead of the participant (at 0° visual angle) and served as a reference point. After an interstimulus interval of 500 ms the second stimulus (probe) was presented pseudo-randomly at one of thirteen locations along the azimuth. After another 500 ms the butterfly-shaped response prompt appeared (see Figure 1). Both stimuli were either auditory (A), visual (V), or audiovisual (AV). For reasons of investigating reliability-weighted integration (that fall outside the scope of the current report), the audiovisual probe stimulus was presented with a small spatial disparity, $\Delta AV = \pm X°$, in two out of three audiovisual conditions (the third AV condition being spatially congruent: $\Delta AV = 0°$). Probe locations and audiovisual disparity sizes were adjusted for each individual to ensure adequate sampling of the psychometric functions (Wichmann & Hill, 2001a) and to minimize the risk of participants noticing the audiovisual conflict (i.e. to avoid violations of the forced-fusion assumption; Meijer & Noppeney, 2018).

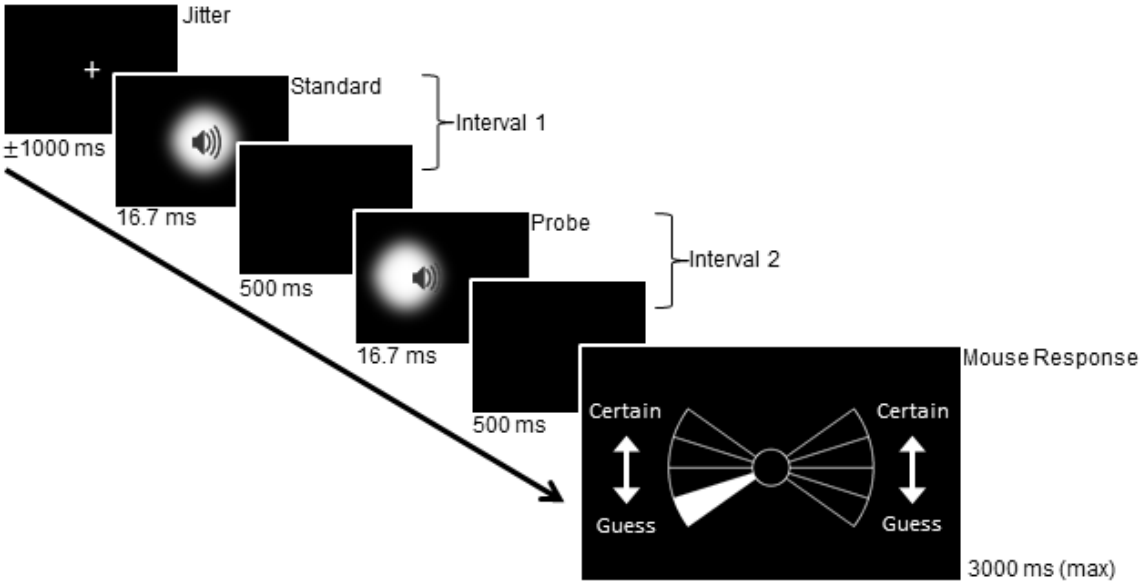


Fig 1 – Illustration of the 2IFC spatial localization task and response mechanism. Participants indicated whether the probe was perceived as left or right of the standard by moving their mouse cursor to either side of the 'butterfly' shape (prior to every trial the cursor was centered to avoid sequential trial-to-trial biases). One of four confidence levels was automatically highlighted (white) when the cursor was above it. By moving the mouse up or down a different confidence level could be selected. It was possible to change the selected the type-1 location and confidence level until the response was confirmed by a left mouse-button click.

The total experiment was spread over four sessions that took place on separate days. In the first session we measured the subjective spatial reliability of the auditory stimuli (short bursts of white noise) and consecutively modified the spatial reliability of the visual stimuli (greyscale blurred Gaussian blobs) to match the auditory reliability (by means of adjusting the blob's size through an adaptive staircase procedure).

Approximate matching of the unisensory reliabilities is important to optimize conditions for observing a multisensory variance reduction (Meijer & Noppeney, 2018; Rohde et al., 2016). The main experiment, wherein we acquired the behavioral data that is presented in Section 3, took place in sessions 2-4. In the main experiment we introduced two additional visual reliability levels: for the so-called high visual reliability level we set the blob size 25% smaller relative to the size for which unisensory reliabilities were individually matched (i.e. medium visual reliability level), and for the low visual reliability level the blobs were 25% larger. All three visual reliability levels were also used for the audiovisual spatially congruent and incongruent conditions, thus making a total of thirteen conditions: 1xA + 3xV + 3x$AV_{\Delta=0^\circ}$ + 3x$AV_{\Delta=-X^\circ}$ + 3x$AV_{\Delta=+X^\circ}$. For each of the twelve V or AV conditions participants completed 468 trials (36 x 13 locations). Twice more A trials were completed (72 x 13 locations) to compensate for the fact that there was only one A condition.

2.2 Analysis

For each observer, and separately for each of the thirteen conditions, we fitted psychometric functions to the proportions of 'perceived right' location responses (i.e. we fitted cumulative Gaussians with a correction for the lapse rate making use of the Palamedes toolbox 1.8.2 for Matlab; www.palamedestoolbox.org). We thus obtained a type-1 JND and PSE (point of subjective equality; Kingdom & Prins, 2016) for each condition.

Building on and extending previous work by Acuna, Berniker, Fernandes & Körding (2015) and Maniscalco & Lau (2012) we have derived (using Bayesian probability theory, see Appendix B) a pair of equations to compute the probability with which a statistically

optimal ideal observer would select a certain confidence level k (out of N possible confidence levels) conditional on location response z for standard and probe stimulus locations S_{st} and S_{pr}, and given certain parameter values for type-1 JND and PSE, and a set of N+1 type-2 criteria $\boldsymbol{C_{2\Delta x}}$ (where the subscript Δx indicates that these criteria represent differences between internal estimates x of probe and standard; i.e. in the current spatial localization experiment the criteria would be in units of degrees visual angle, just like JND, PSE and the stimulus location difference $\Delta S = S_{pr} - S_{st}$):

$$P("ConfLevel = k"|z = "right", S_{st}, S_{pr}) = \frac{\mathrm{erf}\left(\frac{C_{2\Delta x,k+1} - \Delta S + PSE}{\sqrt{2}JND}\right) - \mathrm{erf}\left(\frac{C_{2\Delta x,k} - \Delta S + PSE}{\sqrt{2}JND}\right)}{1 + \mathrm{erf}\left(\frac{\Delta S - PSE}{\sqrt{2}JND}\right)}$$

$$P("ConfLevel = k"|z = "left", S_{st}, S_{pr}) = \frac{\mathrm{erf}\left(\frac{-C_{2\Delta x,k} - \Delta S + PSE}{\sqrt{2}JND}\right) - \mathrm{erf}\left(\frac{-C_{2\Delta x,k+1} - \Delta S + PSE}{\sqrt{2}JND}\right)}{\mathrm{erfc}\left(\frac{\Delta S - PSE}{\sqrt{2}JND}\right)}$$

We use this pair of equations to compute the likelihood for each confidence judgment that was made by the participant, given a set of subject- and condition-specific model parameter values. It thus enables us to optimize the parameter values by means of maximum likelihood estimation (across all trials and locations within one condition after making a correction for a type-2 lapse rate that is fitted as an additional free parameter; see Appendix B Section B.6 for details of the fitting procedure). We note that the type-1 PSE that was obtained from the psychometric function (see above) is a fixed parameter and is not fitted. We further fix the values for $C_{2\Delta x,1} = 0°$ and $C_{2\Delta x,N+1} = \infty°$. We thus obtain maximum likelihood estimates for the JND, N-1 type-2 criteria and one type-2 lapse rate. Critically, this JND is fitted to the metacognitive judgments of the participant, so we will henceforth refer to it as meta-JND.

By fitting the ideal observer model to the confidence responses of a participant (for a particular condition in the current experiment) we summarize the selected proportions of all four confidence levels, conditional on both left and right responses, for each of thirteen probe locations (i.e. up to 4 x 2 x 13 = 104 empirical proportions of confidence judgments) with the following five parameters: meta-JND, three type-2 criteria and one type-2 lapse rate. This method begs the question of how well the ideal observer model is capable of matching the empirical proportions of confidence responses. Therefore, we quantify an absolute goodness-of-fit for each model fit by computing the ratio between the information that would be gained (in terms of relative entropy or Kullback-Leibler divergence) by fitting a chance model (wherein all confidence levels are equally probable independent of stimulus locations and type-1 response) versus the information that was gained by fitting the ideal observer model (Acerbi, Dokka, Angelaki & Ma, 2018; Shen & Ma, 2016; https://github.com/lacerbi/gofit). A goodness-of-fit of 0% means that the model performs equally well as the chance model, whereas 100% means that all empirical information (i.e. entropy) is effectively explained by the model. This method for computing the goodness-of-fit is similar to computing the coefficient of determination (R^2) but it applies to discrete response distributions. While we acknowledge that the use of cross-validated (log-)likelihoods is advocated for computation of the information gained by the model, we instead used an approximation for computational ease and speed: $-0.5 * AIC_c$ (where AIC_c is the Akaike Information Criterion corrected for small sample sizes; Burnham & Anderson, 2002).

Although, confusingly, meta-JND is a measure for the observed metacognitive sensitivity of the participant parameterized as the type-1 sensitivity of a hypothetical ideal

observer, it is important to note that fitting meta-JNDs does not assume that observers compute confidence judgments in a statistically optimal way. The ability to also match suboptimal metacognitive behavior is paramount given the increasing amount of evidence that disputes the hypothesis for metacognitive optimality of human observers; Adler & Ma, 2018a; Aitchison, Bang, Bahrami & Latham, 2015; Denison, Adler, Carrasco & Ma, 2018; Maniscalco, Peters & Lau, 2016; but also see Fetsch, Kiani, Newsome & Shadlen, 2014; Sanders, Hangya & Kepecs, 2016). The model allows for various sources of divergence from optimality. First, we have incorporated a type-2 lapse rate parameter. Second, the free mapping of confidence as an optimally defined posterior probability of being correct (Drugowitsch et al., 2014, Hangya, Sanders & Kepecs, 2016; Pouget et al., 2016) onto an N-point confidence scale by means of individually adjusting the type-2 criteria allows for a generous amount of flexibility. Third, and possibly most important, a certain level of metacognitive sensitivity can only meaningfully be interpreted as statistically optimal when it is conditional on the type-1 performance level (Maniscalco & Lau, 2012). However, it is fundamental to the meta-JND (and meta-d') approach that the ideal observer model is fit without making it conditional on a participant's type-1 performance.

In fact, fitting of meta-JND (and meta-d') allows researchers to compare directly the empirical metacognitive sensitivity of the participant against the theoretical metacognitive sensitivity of a statistically optimal observer with the same type-1 performance. The comparison is commonly known for meta-d' as the metacognitive efficiency ratio (Fleming & Lau, 2014; Bang, Shekhar & Rahnev, 2018). Here, we define a

similar ratio for meta-JND: $M_{ratio} = \frac{JND}{meta-JND}$. Since we expect meta-JND to be larger for suboptimal observers, this ratio would normally range between zero and one.

Computing the M_{ratio} is only one of many ways in which we could compare empirical versus statistically optimal metacognitive sensitivity. From the perspective of the type-1 JND as a measure of sensory noise (Kingdom & Prins, 2016) it is natural to express the difference between type-1 JND and meta-JND in terms of added (metacognitive) Gaussian noise with standard deviation $\sigma_{meta} = \sqrt{metaJND^2 - JND^2}$ (see also Bang et al., 2018). This method of quantifying metacognitive suboptimality is appealing because of its intuitive interpretation. Since we expect $metaJND \geq JND$, we also expect $\sigma_{meta} \geq 0°$. Metacognitive noise σ_{meta} is zero when $metaJND = JND$ (i.e. for an ideal observer). Please note that there is no upper bound for $metaJND$ or σ_{meta}: they will tend to go to infinity when participants provide random confidence responses. We have computed σ_{meta} alongside M_{ratio} for all conditions of each participant. For practical reasons, we have set $\sigma_{meta} = 0°$ whenever $(metaJND^2 - JND^2) < 0°$ (e.g. this may occur due to measurement noise).

To determine whether differences between the various measures (JND, meta-JND, M_{ratio}, σ_{meta}) could be explained by experimental conditions (e.g. visual reliability level) we performed repeated-measures ANOVAs. To be able to quantify evidence against the existence of any such effect we also computed default Bayes factors for ANOVA designs (making use of BayesFactor Package 0.9.12-2 with default parameter settings, applying Cauchy priors to standardized effect sizes and setting 'subject-ID' as random factor; Rouder, Morey, Speckman & Province, 2012; http://bayesfactorpcl.r-forge.r-project.org/). A Bayes factor $BF_{01} > 1$ indicates that there is more evidence that

supports the null hypothesis that there is no effect of the tested factor. Both ANOVAs and Bayes Factor ANOVAs were performed in R 3.4.1.

1. Results

Data from twenty-two participants was included in the analyses (see Appendix A.1 for details). We fitted psychometric functions to their type-1 localization responses and Bayesian ideal observer models to their type-2 confidence judgments, separately for each of thirteen experimental conditions: one A condition and three visual reliability conditions for V, $AV_{\Delta=0^\circ}$ (i.e. spatially incongruent), $AV_{\Delta=-X^\circ}$ and $AV_{\Delta=+X^\circ}$ (i.e. spatially incongruent with a small, so-called unnoticeable spatial disparity ΔAV). Before we investigate group-level results to address the main research questions on metacognitive efficiency and absolute confidence for multisensory versus unisensory conditions, we will first evaluate the performance of the here developed meta-JND model.

3.1 Fitting meta-JND

Figure 2 illustrates fitting behavior of the meta-JND model for three representative fits (from three different conditions of three different participants). We have summarized the empirical and model-predicted behavior in terms of mean confidence levels per probe location (following suggestions by Hangya et al., 2016; Sanders et al., 2016; see also Adler & Ma, 2018a, 2018b). The depicted data shows three very different metacognitive biases: Panel A shows an observer with a tendency to report high confidence whereas the observer in panel B prefers low confidence levels. The observer in panel C, on the other hand, seems to avoid either of the extremes and rather reports intermediate confidence levels 2 and 3. The model efficiently fits such biases by

adjusting the type-2 criteria $\boldsymbol{C2_{\Delta x}}$: fitting smaller criteria for the observer in panel A, larger ones for the observer in panel B, and adjusting $C2_{\Delta x,2}$ and $C2_{\Delta x,4}$ in opposite directions for the observer in panel C. Furthermore, the observer in panel C generally needed larger spatial differences between probe and standard to distinctively differentiate correct from incorrect responses by means of high versus low confidence judgments. Such lower metacognitive sensitivity is fitted by increasing the meta-JND parameter value. The actual mean confidence level difference between correct and incorrect responses (i.e. yellow shaded areas in Figure 2) is the result of a complex interplay between all model parameters: type-1 PSE, meta-JND, type-2 criteria and type-2 lapse rate.

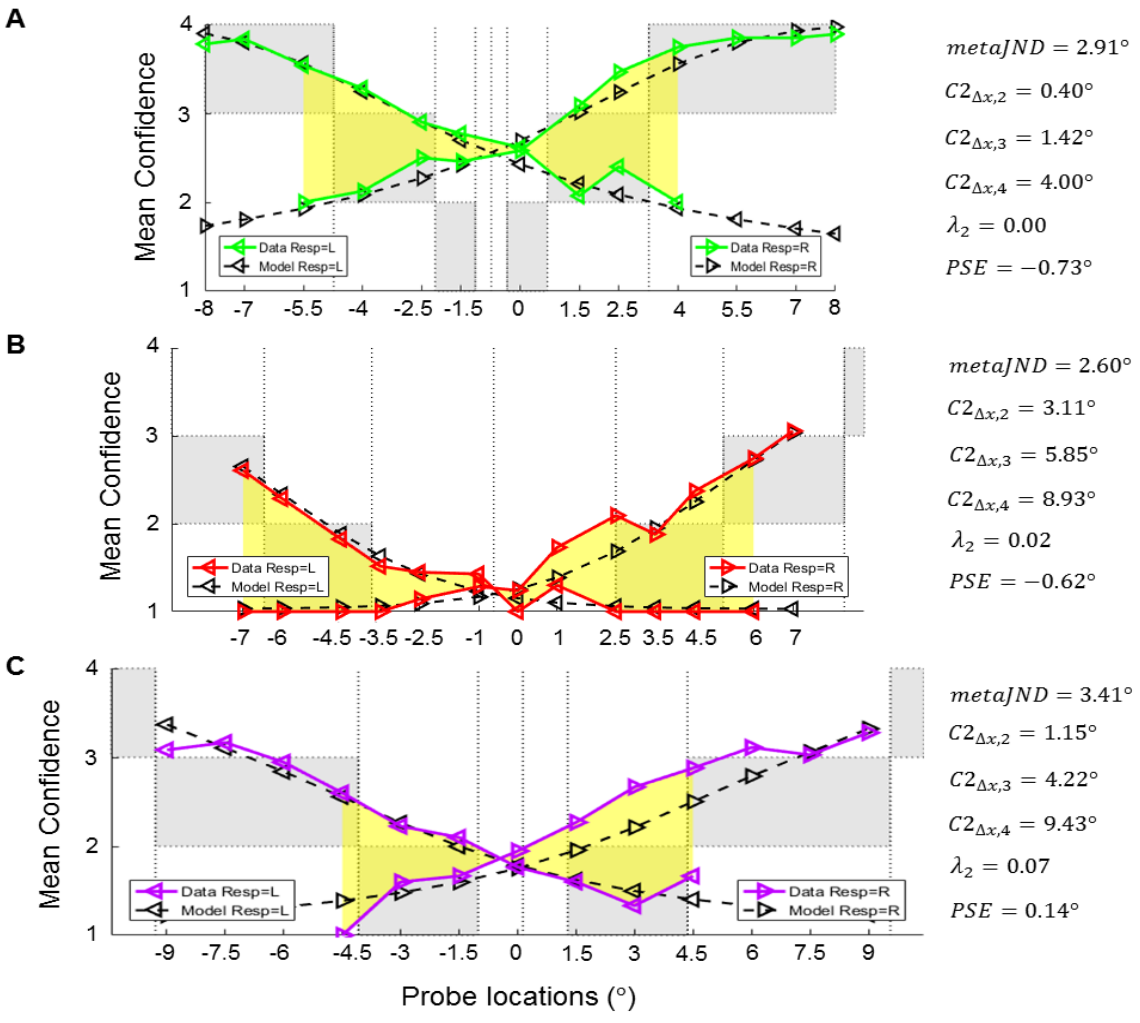


Fig 2 – Illustration of the meta-JND model fits for three conditions of three different participants (with individualized probe locations on x-axis, see Appendix A.3). Fig. 2 a. Auditory condition. Fig. 2 b. Visual condition with medium reliability. Fig. 2 c. Audiovisual incongruent condition ($AV_{\Delta=-X^{\circ}}$) with low visual reliability. Participant's type-2 responses (confidence levels 1-4) were summarized per probe location by computing the mean confidence level, conditional on their type-1 location responses (solid colored lines; triangular markers indicate the type-1 location responses: Left or Right). The expected mean confidence levels for the model (with parameter values that were fitted to participant's data listed on the right) were computed as a weighted average, using the predicted confidence probabilities as weights (dashed black lines). Yellow shaded areas illustrate the difference between mean confidence levels for correct and incorrect type-1 responses (but note that the yellow area is not a good quantitative indicator of metacognitive sensitivity in terms of meta-JND, because it is also affected by the other model parameters). The vertical dotted lines illustrate the type-2 criteria adjusted for the PSE (i.e. left-right bias) in units of visual angle (note that these are internal brain estimates, unlike the true spatial locations of the probes). The upper bounds of the grey shaded areas indicate the confidence levels that would have been chosen if the internal spatial estimate ($\Delta x = x_{pr} - x_{st}$) falls between two particular type-2 criteria (confidence level 1 is given for Δx near zero but is not shown). N.B. some empirical data points are 'missing' because the participants made no incorrect responses for the more eccentric probe locations.

To test whether the model is capable of disentangling the effects of each parameter we performed a model parameter recovery analysis for a multidimensional grid of realistic parameter values for JND, $\boldsymbol{C2}_{\Delta x}$, λ_2 and PSE. For each combination of parameters we simulated one thousand confidence responses for each of thirteen probe locations using Monte Carlo sampling and the generative model as described in Appendix B. We then fitted the model to each of the generated datasets. Visual inspection of the fitted parameters against the generating parameters, separately for each of the three relevant parameters, revealed no systematic biases of the model fits and only very small inaccuracies. We thus concluded that, given a sufficient amount of data, the model is capable of retrieving the true parameter values despite their complex interplay.

Although parameter recovery thus suggests that the model by itself is valid, the more important question is whether the model is also capable of accurately fitting human behavior. To quantify the model's ability to do so, we have computed the absolute goodness-of-fit (Shen & Ma, 2016; Acerbi, et al., 2018). This measure expresses the goodness-of-fit as a percentage between 0% (not better than a chance model) and 100% (a perfect fit in terms of the variance in the data that was explained by the model). The mean goodness-of-fit across all 286 (= 22 participants x 13 conditions) fits is 92%. The minimum goodness-of-fit was 66%. The minimum mean goodness-of-fit for any participant (average across conditions) was 80%. We conclude that the model fits were of reasonable quality. We also checked whether a bad fit automatically leads to high values for meta-JND, but we found no correlation between absolute goodness-of-fit and meta-JND (Pearson's r = -0.10, p = 0.09), thus suggesting that the model does not exhibit such an undesired bias.

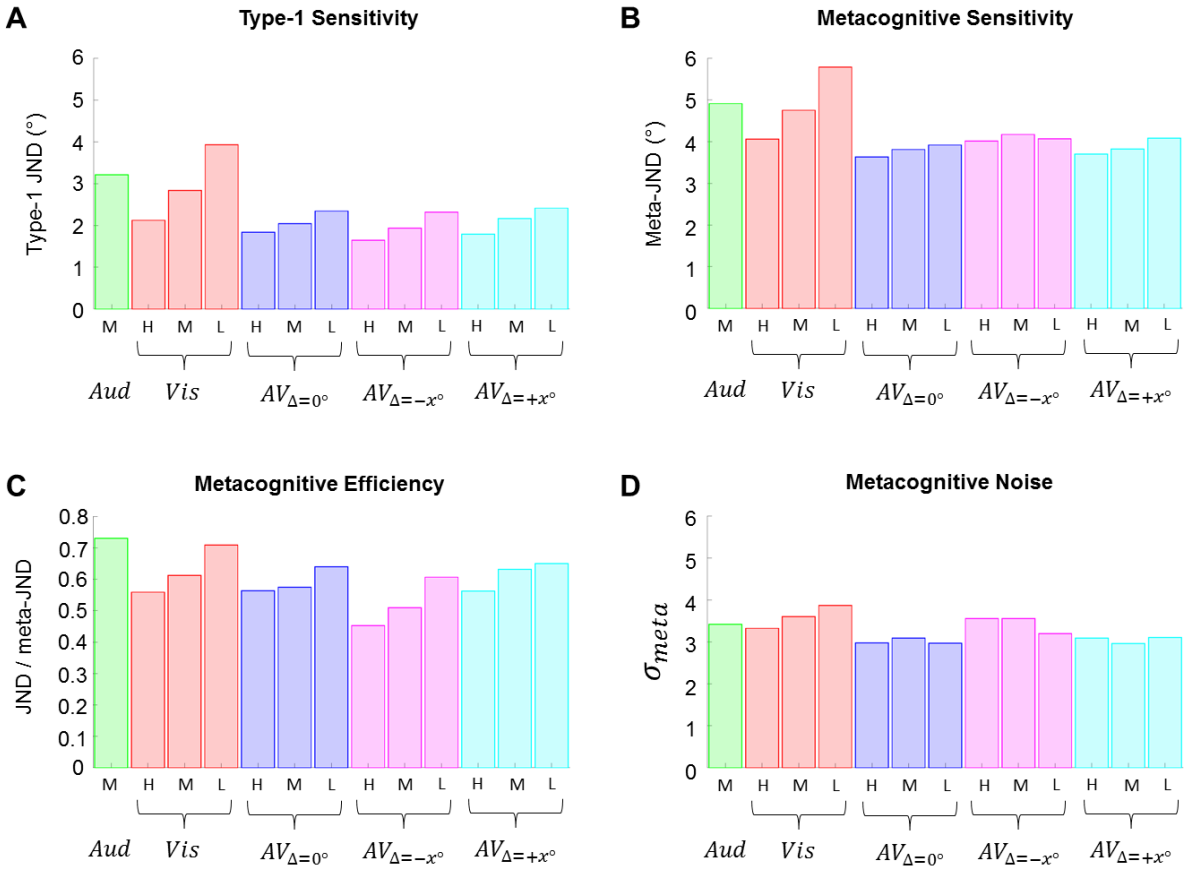


Fig 3 – Mean group-level results for type-1 JND (panel A), meta-JND (panel B), M_{ratio} (panel C) and σ_{meta} (panel D). The bars are colored by sensory modality (and audiovisual spatial congruency). There are three visual reliability levels for the visual and audiovisual conditions, indicated as H = high, M = medium, L = low. The auditory condition is also labelled as 'medium' because the visual medium reliability level was matched to the auditory reliability individually for each participant (see Appendix A.3).

3.2 Type-1 JND, meta-JND, M_{ratio} and σ_{meta}

Group-level averages for type-1 sensitivity (JND), metacognitive sensitivity (meta-JND), metacognitive efficiency (M_{ratio}) and metacognitive noise (σ_{meta}) are shown in Figure 3 for each of the thirteen experimental conditions. To investigate the effect of visual

reliability (H = high, M = medium, L = low) on each of these measures, we performed eight independent repeated-measures ANOVAs. For the unisensory visual condition, 'reliability' was the only within-subjects factor with three levels. For the audiovisual conditions we included 'AV-congruency' as an additional within-subjects factor with two levels: congruent vs. incongruent (computing the mean across both incongruent conditions per participant). Bayes factors BF_{01} express evidence in favour of the null-hypothesis (i.e. no effect). The results of these statistical tests are summarized in Table 1.

	Effect of Reliability :: Vis	**Effect of Reliability :: AV**
JND	F(2) = 40.8, p < 0.001, BF_{01} < 0.001	F(2) = 54.1, p < 0.001, BF_{01} < 0.001
meta-JND	F(2) = 21.7, p < 0.001, BF_{01} < 0.001	**F(2) = 2.25, p = 0.11, $\boldsymbol{BF_{01}}$ = 2.09**
$\boldsymbol{M_{ratio}}$	F(2) = 5.75, p = 0.006, BF_{01} = 0.066	F(2) = 12.6, p < 0.001, BF_{01} < 0.001
$\boldsymbol{\sigma_{meta}}$	**F(2) = 1.06, p = 0.36, $\boldsymbol{BF_{01}}$ = 2.47**	**F(2) < 1, $\boldsymbol{BF_{01}}$ = 11.1**

Table 1 – Repeated-measures ANOVAs and corresponding Bayes Factor analyses. Shown are the results for a main effect of the visual reliability separately for unisensory visual conditions (left) and audiovisual conditions (right). The tests that resulted in no significant effect of 'reliability' are printed in bold. There were no significant effects of 'AV-congruency' in any of the tests: F(1) < 3.83, p > 0.05, $\boldsymbol{BF_{01}}$ > 0.94 for all tests.

As intended by design, we found that the size of the visual blobs (that defined the three visual reliability levels for visual and audiovisual conditions) had significantly affected

localization performance: i.e. more spatial blurring of the visual stimuli led to reduced levels of type-1 sensitivity (JND) in both unisensory and audiovisual conditions. Furthermore, decreasing the visual reliability also negatively affected type-2 sensitivity (meta-JND). These effects were to be expected based on the relationship that exists between type-1 and type-2 performance (Maniscalco & Lau, 2012). However, the influence of visual reliability on meta-JNDs is less pronounced, and it is not significant for the audiovisual conditions.

Meta-JNDs were generally greater than type-1 JNDs resulting in metacognitive efficiency ratios below one. (We found a few exceptions where M_{ratio} > 1, with a maximum of 1.27, but we contribute these type-2 super-optimal findings to measurement noise.) Interestingly, we found a significant effect of visual reliability level on M_{ratio}, for both visual and audiovisual conditions. The fact that we observed this difference for the audiovisual efficiency ratios and type-1 JNDs, but not for meta-JNDs, suggests that the reliability effect on metacognitive efficiency ratios is driven by the type-1 JNDs rather than the meta-JNDs. This suggestion is further corroborated by looking at the metacognitive noise estimates σ_{meta}. The statistical tests revealed that metacognitive noise levels were unaffected by the amount of type-1 sensory noise. We conclude that, while metacognitive sensitivity (meta-JND) is influenced by type-1 sensory performance because the latter provides a theoretical maximum for the former (Maniscalco & Lau, 2012), the amount of noise that is added during metacognitive processing is not affected by type-1 task performance.

To focus more specifically on the current study's interest in multisensory integration, we next assessed the difference in metacognitive noise between auditory, visual and

audiovisual conditions. Since we found no effect of 'AV-congruency' (see caption of Table 1) we averaged the audiovisual σ_{meta} for each participant. To test whether metacognitive noise levels were different in any of the sensory modalities (A, V, AV) we performed three independent repeated-measurement ANOVAs: one for each visual reliability of V and AV, with (the only set of) auditory σ_{meta} included in all three tests. We found no effect of sensory modality for any of the reliability levels: F(2) < 1, BF_{01} = 9.44 (High), F(2) < 1, BF_{01} = 4.62 (Medium), F(2) = 1.11, p = 0.34, BF_{01} = 2.28 (Low).

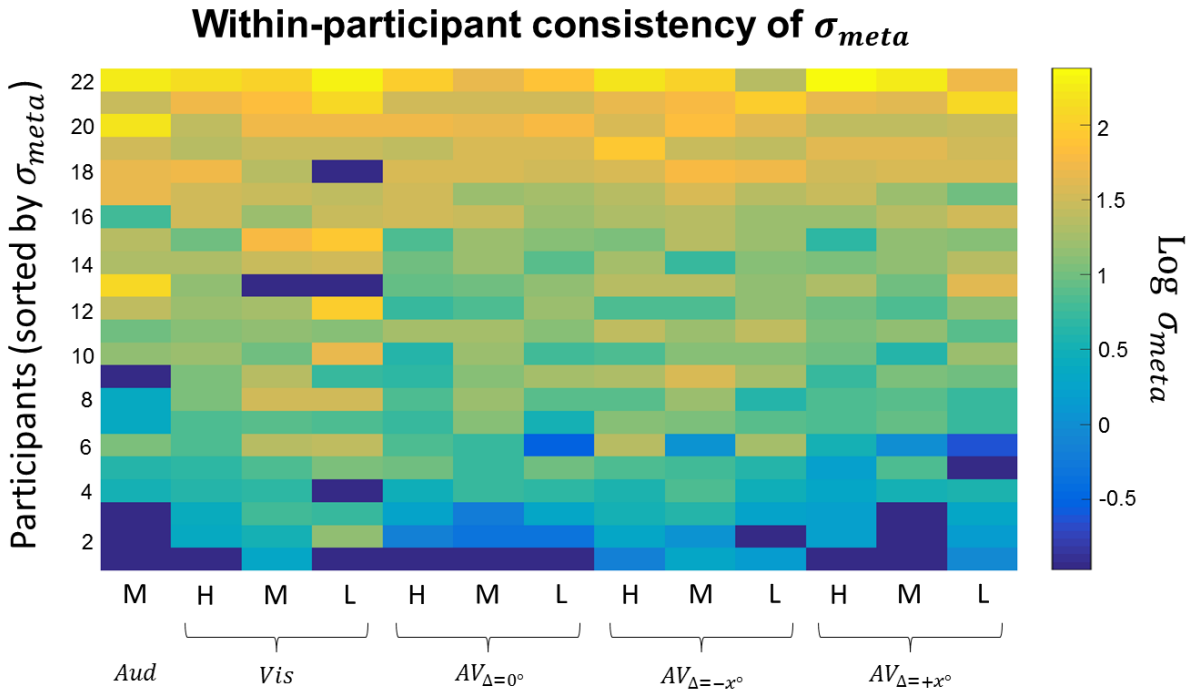


Fig 4 – Metacognitive noise for all 286 fits (22 participants x 13 conditions). To illustrate the consistency of σ_{meta} for each participant across conditions we sorted the participants by their mean σ_{meta}. Note that the color-coding depicts the natural logarithm of each σ_{meta}.

The above-described results suggest that metacognitive noise is stable across modalities and independent of visual reliability levels. To further investigate this we compared the intra-participants variance (i.e. the SD of σ_{meta} across conditions for each participant) with the inter-participants variance (the SD of σ_{meta} across participants for each condition). Figure 4 illustrates the difference between both measures (i.e. compare horizontal versus vertical colors). While there are large differences in the amount of metacognitive noise between participants, the amount of variance across various sensory modalities and reliability levels for each participant is relatively small. A two-sample t-test (for 22 against 13 SDs, respectively) confirmed that the intra-participant variance of σ_{meta} (mean SD = 1.02) is significantly smaller than the inter-participant variance (mean SD = 1.94): $t(33) = -6.01$, $p < 0.001$.

3.3 Type-2 criteria

The experimental design allowed participants to select one of four confidence levels on each 2IFC trial. The meta-JND model assumes that such a confidence judgment depends on the internal estimate of the spatial difference between probe and standard (Δx): a higher confidence level is selected when Δx exceeds a certain type-2 criterion. There are three criteria to separate four confidence levels. The group-level means of these criteria are shown in Figure 5.

A lower type-2 criterion generally results in more high confidence ratings. We hypothesized that the placement of such criteria depends on the reliability of the stimuli (since we expected higher confidence judgments for more reliable stimuli). To investigate this hypothesis we performed two repeated-measures ANOVAs, one for V and one for AV conditions (as before). Besides 'reliability', we added 'C2-level' as a

second within-subjects factor with three levels: $C2_{\Delta x,2}$, $C2_{\Delta x,3}$ and $C2_{\Delta x,4}$. The ANOVA for the AV conditions also contained a third within-subjects factor: 'AV-condition' with three levels ($\Delta AV = 0°$, $\Delta AV = -X°$ and $\Delta AV = +X°$). Both ANOVAs confirmed the hypothesis that fitted type-2 criteria were smaller for more reliable stimuli: F(2) = 23.8, p < 0.001, BF_{01} < 0.001 (V) and F(2) = 4.55, p = 0.011, BF_{01} < 0.60 (AV). Noteworthy is that there was no difference between AV conditions: F(2) < 1, BF_{01} = 49.2.

The question of whether multisensory integration leads to higher confidence can be viewed as a special case of the above-discussed $C2_{\Delta x}$ reliability dependency. If we assume that multisensory integration does not only lead to a variance reduction across trials (as was found empirically; e.g. Ernst & Banks, 2002; Alais & Burr, 2004), but also results in a sensory noise reduction for any one multisensory stimulus (as is predicted by maximum likelihood estimation) then we should also observe a confidence boost for multisensory as opposed to unisensory stimuli. Specifically, we should find that audiovisual type-2 criteria are lower than the minimum of either auditory or visual type-2 criteria. To assess this hypothesis we computed the mean audiovisual $\boldsymbol{C2_{\Delta x}}$ across the three (in)congruent conditions (per participant, reliability level and $C2_{\Delta x}$ level) and compared those against the lowest unisensory type-2 criteria (either A or V, per participant, reliability level and $C2_{\Delta x}$ level) by means of a repeated-measures ANOVA with three within-subject factors: 'reliability' (3 levels), 'C2-level' (3 levels) and 'Uni-vs-AV' (2 levels). The analysis revealed a significant difference between unisensory and multisensory type-2 criteria, F(1) = 14.6, p < 0.001, BF_{01} < 0.0074, indicating that participants adjusted their criteria for audiovisual relative to unisensory trials. Simply put, this criterion change is a clear signature of the confidence boost for multisensory

trials. Neither the interaction between 'Uni-vs-AV' and 'reliability' nor the interaction between 'Uni-vs-AV' and 'C2-level' were significant (F(2) = 0.72, p = 0.49 and F(2) = 1.39, p = 0.25, respectively) thus suggesting that the difference between the minimum unisensory and bisensory type-2 criteria does not depend on the particular reliability level or type-2 criterion.

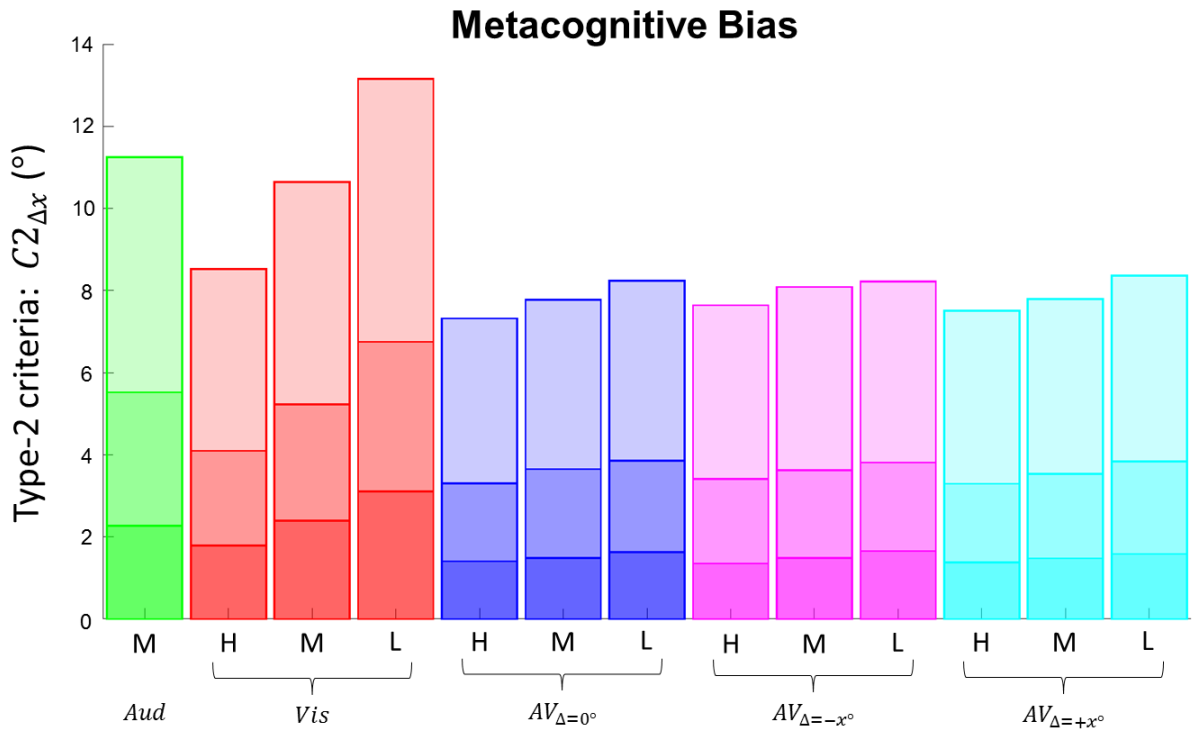


Fig 5 – Group-level mean metacognitive biases. We computed the mean across participants for each of the three fitted criteria per condition. These criteria separate each bar in the plot into three confidence levels: darkest = 1 (low confidence), lightest = 3. The highest confidence level = 4 corresponds to the area above the bar (top of the bar is the largest type-2 criterion).

3.4 Predicting Type-1 JNDs and type-2 criteria based on maximum likelihood estimation

To see whether participants had indeed integrated the multisensory signals, we tested for an audiovisual variance reduction of the type-1 localization responses. Similar to the above-described approach for type-2 criteria, we now computed the mean audiovisual type-1 JNDs and compared these to the smallest unisensory JNDs (per participant and reliability level). A repeated-measures ANOVA (two within-subject factors: 'reliability' and 'Uni-vs-AV') confirmed significance of the audiovisual variance reduction: F(1) = 48.2, p < 0.001, BF_{01} < 0.001 (see Fig. 3A). A more stringent test for multisensory variance reduction is to compare the empirical mean audiovisual JNDs to the predicted bisensory JNDs based on the unisensory JNDs and integration according to maximum likelihood estimation (MLE; Ernst & Banks, 2002; Alais & Burr, 2004): $JND_{AV,MLE} = \sqrt{\left(JND_A{}^2 * JND_V{}^2\right) / \left(JND_A{}^2 + JND_V{}^2\right)}$. The repeated-measures ANOVA (with the same factors as before) suggests that participants did indeed integrate the sensory signals in line with MLE: F(1) < 1, BF_{01} = 5.44.

Since we were able to predict the degree to which the audiovisual JNDs decreased relative to the unisensory JNDs (using MLE predictions, see above), we set out to see whether we could also predict the degree to which the audiovisual type-2 criteria decrease relative to the unisensory type-2 criteria. In order to do so, we make an important assumption: each criterion that separates two confidence levels ($C2_{\Delta x}$), relates to a particular Bayesian posterior probability of being correct which is independent of the sensory modality (see also Appendix B.3). In other words, if for example $C2_{\Delta x,3}$ for the auditory modality relates to a posterior probability of 0.8, then we assume that $C2_{\Delta x,3}$ for the visual modality relates to that same posterior probability.

If we further assume a flat spatial prior, then the posterior probability of being correct can be computed using the formula for a cumulative Gaussian: $\frac{1}{2}\left[1 + \mathrm{erf}(\frac{C2_{\Delta x}}{\sqrt{2}JND})\right]$ (c.f. Equation B.4). By inserting the unisensory (type-1) JNDs and type-2 criteria we computed the posterior probability of being correct for each of the four unisensory conditions (1 x A and 3 x V) per participant and criterion level. We then averaged the posterior probabilities across A and V to obtain three reliability-specific posterior probabilities (per participant and criterion level). Following the assumption that these posterior probabilities for each criterion level are independent of the sensory modality, we then predicted (i.e. computed) the audiovisual type-2 criteria by inserting the MLE-predicted audiovisual JNDs ($JND_{AV,MLE}$) into the inverse of the cumulative Gaussian (i.e. 'norminv' in Matlab; μ = 0°). Note that the predicted audiovisual type-2 criteria are thus entirely based on unisensory data: auditory and visual type-1 JNDs and type-2 criteria. We then compared the predicted audiovisual criteria with the actually fitted audiovisual criteria (mean across the three AV conditions: $AV_{\Delta=0°}$, $AV_{\Delta=-x°}$ and $AV_{\Delta=+x°}$).

Figure 6 depicts the results of the above described analysis. The scatter plot shows an excellent correlation between predicted and fitted type-2 criteria (coefficient of determination R^2 = 0.85). Importantly, the relationship between predicted and fitted criteria is also unbiased. (In contrast, a similar scatter plot, not shown, for audiovisual versus unisensory type-2 criteria shows a clear bias with ~75% of the markers indicating that the audiovisual criteria are smaller than the unisensory criteria; c.f. last analyses in Section 3.3). A repeated measures ANOVA with three within-subject factors: 'reliability' (3 levels), 'C2-level' (3 levels) and 'Predicted-vs-Fitted' (2 levels) revealed no significant difference between predicted and fitted audiovisual type-2 criteria: F(1) = 2.57, p = 0.11,

BF_{01} = 2.60 (n.b. the data were log-transformed for this analysis). This result (i) confirms the fixed mapping of a particular criterion level ($C2_{\Delta x,i}$) to the Bayesian posterior probability of being correct independent of the sensory modality, and (ii) it demonstrates that observers are able to set their type-2 criteria based on the sensory uncertainty of the stimuli, equally for unisensory and multisensory stimuli.

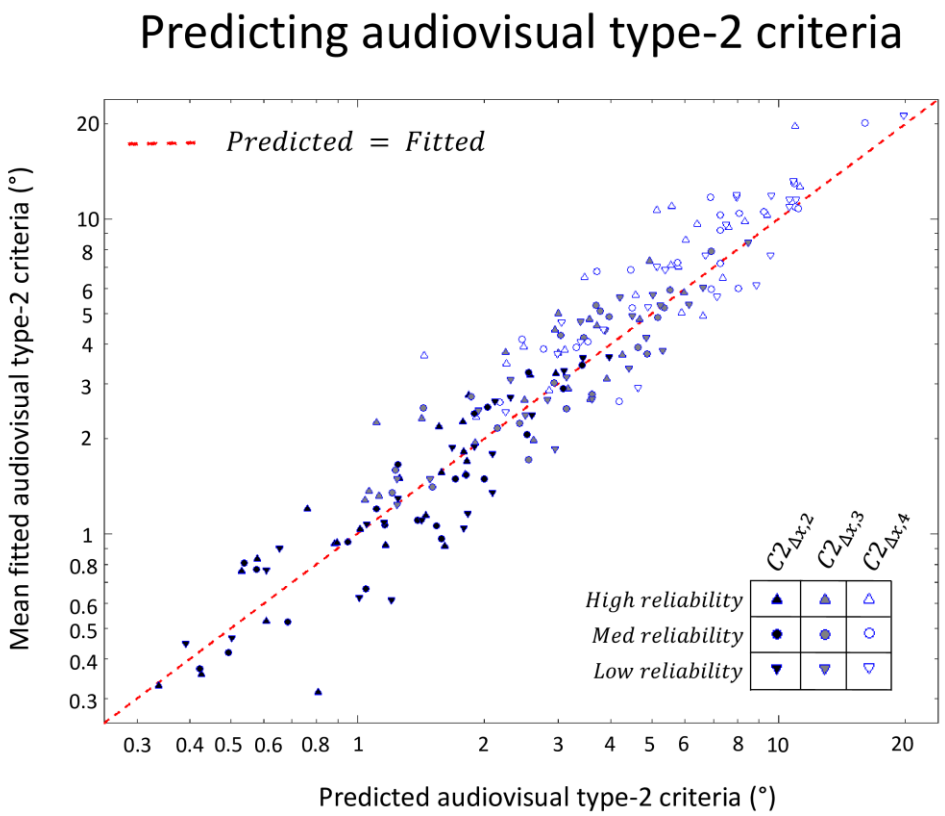


Fig 6 – Predicting audiovisual type-2 criteria. The figure shows a scatterplot for comparison of the mean (across congruent and incongruent AV conditions) fitted audiovisual type-2 criteria (y-axis) and the predicted audiovisual type-2 criteria based on unisensory JNDs and unisensory type-2 criteria only (x-axis). The diagonal dashed red line indicates equality (x=y). Different colors and symbols indicate the reliability level and type-2 criterion (see legend). There is one marker of each kind per participant (N=22).

2. Discussion

The current audiovisual spatial integration study empirically answers two fundamental questions concerning the interaction between multisensory integration and metacognition (Deroy et al., 2016). First, we have shown that metacognitive processes for making confidence judgments on perceptual decisions are equally effective for unisensory and multisensory stimuli. Specifically, we found that metacognitive noise was not larger for audiovisual relative to unisensory signals, thus implying that multisensory integration is not associated with a loss of introspective ability. Second, we demonstrated that multisensory integration leads to increased levels of confidence through a type-2 criteria shift that corresponds to the degree of integration-induced audiovisual variance reduction. This finding provides strong support for the hypothesis that multisensory integration reduces sensory uncertainty on a per-stimulus basis, above and beyond the established behavioral precision improvement that is found across many stimulus presentations (e.g. Ernst & Banks, 2002; Alais & Burr, 2004).

To be able to answer these questions adequately we have extended the now popular meta-d' approach that expresses metacognitive sensitivity in terms of the type-1 performance of an ideal observer model (Maniscalco & Lau, 2012). Using Bayesian probability theory, we have derived a statistically-optimal observer model for confidence judgments that can be applied to psychometric function-based experimental designs and expresses metacognitive sensitivity as 'meta-JND'. The extension yielded some noteworthy benefits over the original signal-detection based approach. 1. The use of multiple stimulus levels avoids the need to carefully titrate experimental conditions to prevent ceiling/floor effects. 2. Estimates of metacognitive sensitivity and metacognitive

bias based on a range of stimulus levels are likely to be more accurate and potentially generalize better across various experimental tasks. 3. The measurement units for both meta-JND and type-2 criteria are intuitive to comprehend and free of interdependencies. In particular, the fitted values for type-2 criteria are unaffected by an observer's metacognitive sensitivity, which allows for straightforward comparisons of metacognitive biases across conditions and sensory modalities.

Nevertheless, a model-based parameterization of human metacognitive sensitivity and metacognitive bias would only be useful if the model is sufficiently able to describe human behavior. Based on the model's absolute goodness-of-fit (Acerbi et al., 2018; Shen & Ma, 2016) we considered the fits to be suffienciently good to make reliable inferences with the parameter estimates that were obtained. Yet, we believe it worthwhile for future research to formally test the current model, that is loosely based on a Bayesian ideal observer with flexibility to be able to make suboptimal inferences, against other observer models that instead apply heuristics to predict human confidence judgments (Adler & Ma, 2018a).

Our finding that the commonly-used metacognitive efficiency ratio (Fleming & Lau, 2014) negatively correlates with stimulus reliability was previously also recognized by Bang, Shekhar & Rahnev (2018). Similar to us, they reasoned that this was most likely a consequence of type-1 performance on the computed sensitivity ratio and does not indicate a genuine decrease of introspective ability for more reliable stimuli. Our novel approach allowed us to directly quantify the difference between type-1 JND and meta-JND in terms of the added metacognitive noise. By doing so we demonstrated that the amount of metacognitive noise was unaffected by type-1 sensory noise and relatively

stable for each participant across various conditions. Specifically, we found that is was not affected by the sensory modality of the stimuli (A, V, AV). Such intra-subject stability for metacognitive performance across modalities corresponds well with previous findings of metacognitive consistency across various tasks and experimental designs (Ais, Zylberberg, Barttfeld & Sigman, 2016; Song et al., 2011). Together with the finding of a fixed mapping relationship between type-2 criteria and the Bayesian posterior probability of being correct independent of sensory modality (Section 3.4), this strongly supports previous proposals for the supra-modality of metacognition (De Gardelle, Le Corre & Mamassian, 2016; Faivre, Filevich, Solovey, Kühn & Blanke, 2018).

The here presented results on supra-modal metacognitive noise contribute to the abundance of evidence that points towards a hierarchical model for metacognitive perceptual decision making (Bang et al., 2018; De Martino, Fleming, Garrett & Dolan, 2013; Fleming et al., 2015; Jang, Wallsten & Huber, 2012; Maniscalco & Lau, 2016; Mueller & Weidemann, 2008; Rahnev, Nee, Riddle, Larson & D'Esposito, 2016; Van den Berg, Yoo & Ma, 2017). According to such a hierarchical view, metacogniton follows ordinary perceptual processes in a second stage of processing. Neuroimaging evidence suggests that the prefrontal cortex plays an important role in this late-stage forming of confidence judgments about perceptual decisions (De Martino et al., 2013; Murphy et al., 2016; Rahnev et al., 2016; Wokke, Cleeremans & Ridderinkhof, 2017). Presumably, these frontal areas would base their confidence judgments on 'read outs' of sensory uncertainty that is encoded elsewhere in the brain (e.g. sensory and parietal areas). Given the amount of noise in the nervous system as a whole (Faisal, Selen & Wolpert, 2008) it is very plausible that metacognitive noise impedes observers from reaching

statistically optimal metacognitive sensitivity. Task complexity (Beck, Ma, Pitkow, Latham & Pouget, 2012) and unstable use of type-2 criteria over the course of the experiment may further exacerbate empircally determined metacognitive noise values (Maniscalco & Lau, 2012; Mueller & Weidemann, 2008). Keeping in mind the impressive differences in metacognitive noise levels between participants, an interesting direction for future research is to investigate what other factors influence our metacognitive abilities (Bang et al., 2018; Fleming et al., 2015).

Our model-based approach allowed us to examine metacognitive biases independently from metacognitive sensitivity. We found that the fitted type-2 criteria were smaller for multisensory stimuli than for either of the unisensory component stimuli by themselves. Since these criteria were fitted to the confidence judgments of participants, with identical stimulus locations for the bisensory and unisensory conditions, we concluded that participants' confidence was higher for multisensory as opposed to unisensory stimuli. Given the complexity with which mean confidence levels are affected by metacognitive sensitivity and type-2 lapse rate, we argue that differences of type-2 criteria provide the safest way to quantify confidence differences between conditions (see also Fleming & Lau, 2014).

Furthermore, the type-2 criteria across three visual reliability levels indicated that the criterion shift for multisensory stimuli was most likely based on reduced sensory uncertainty for the integrated sensory signals. In fact, the degree to which the type-2 criteria were smaller corresponded to predictions that were made based on sensory integration according to maximum likelihood estimation. We thus concluded that the confidence increase for multisensory decisions probably resulted from a sensory noise

reduction after multisensory integration. Although this may seem obvious at first thought, we argue that this is a critical finding which, as far as we are aware, has not been published before. While sensory noise reduction has widely been accepted as a fundamental beneficial property of multisensory integration (e.g. through maximum likelihood estimation), direct evidence for this proposal has been limited. In fact, despite the popularity of probabilistic models for multisensory perception (e.g. Ma, Beck, Latham & Pouget, 2006; Pouget et al., 2013), there is little direct neuroimaging/neurophysiological evidence for probabilistic perceptual encoding in the brain (van Bergen, Ma, Pratte & Jehee, 2015). Instead, most evidence supporting the sensory uncertainty reduction hypothesis for multisensory integration comes from psychophysics experiments wherein a behavioral variance reduction is observed across many trials (Ernst & Banks, 2002; Alais & Burr, 2004). Although such findings may indeed suggest that multisensory integration according to MLE has taken place, they can also be explained by simple weighted averaging of two unisensory point estimates on any one trial (Rahnev, 2018). In other words, they do not necessarily provide evidence that probability distributions were used during multisensory integration. We believe that the here shown confidence boost by multisensory noise reduction contributes new evidence in favour of distribution-based integration, because weighted averaging of point-based estimates would not be sufficient to explain the multisensory confidence boost (although more complex alternative explanations might).

In summary, the current study provides support for a hierarchical view of multisensory decision making (see also Faivre et al., 2018). At lower stages in the hierarchy unisensory signals are integrated, weighted by their reliability, leading to sensory noise reduction

for the multisensory estimates. At a second stage, metacognitive processes examine the sensory uncertainty that is associated with the multisensory estimates and assign corresponding confidence levels to the perceptual decisions. Metacognitive noise perturbs sensory uncertainty estimates causing metacognitive sensitivity to be suboptimal. Importantly, although the amount of metacognitive noise varies dramatically from person to person, the amount seems independent of sensory modality or stimulus reliability. Multisensory noise reduction thus translates into relatively higher decision confidence and the ability to introspect on sensory uncertainty is equal for unisensory and multisensorily integrated stimuli.

Appendix A: Details of experimental methods

A.1. Sample characteristics

Twenty-eight participants were initially recruited. Two participants were excluded in the first session because their auditory localization performance was inadequate (see Section A.3.1.1). One participant chose to withdraw from the study during the second session. One participant was excluded after the second session because the third and fourth sessions could not be scheduled within the next two weeks (extended time-periods between sessions may lead to inconsistent performance across sessions). Two participants were additionally excluded from data analyses post-hoc, because their mean confidence levels across the various probe locations were essentially flat lines (several conditions where the fitted meta-JND > 50°, whereas the maximum meta-JND for all other participants did not exceed 12.5°), thus indicating that these two

participants (probably) had not performed the task adequately (or improbably, that they were incapable of metacognitive evaluation of their location responses for all experimental conditions).

Behavioral data sets from twenty-two participants were included for data analysis (eight males; mean age 23 years, range 19-30 years; six participants reported left-handedness and operated the mouse with their preferred hand). All participants were university students with reportedly normal hearing, (corrected to) normal vision and no history of neurological or psychiatric disorder. Participants provided informed consent and were financially compensated. The study was approved by the human research review committee of the University of Birmingham (approval numbers ERN_11-0470AP4 & ERN_15-1458P).

A.2. Stimuli and experimental setup

Visual stimuli were greyscale circular blobs with bivariate Gaussian amplitude envelopes presented for a duration of 16.7 ms in low-contrast (20 cd/m^2 in its centre) on a darker grey background (15 cd/m^2) by means of back-projection (60Hz BenQ MW529 DLP projector) on an opaque fine-PVC fabric projector screen (127.5 cm width x 170cm height). The size of the visual stimuli was defined by the 2D Gaussian's standard deviation σ_{blob} (symmetrical in all directions) and was individually adjusted for each participant (see Section A.3.1.2).

Auditory stimuli were 16.7 ms bursts of white noise (70 dB SPL; 5 ms on/off ramp) presented by means of headphones (Sennheiser HD 280 Pro) with a playback frequency of 192 kHz. The auditory signals were convolved with standardised head-related transfer functions (Gardner & Martin, 1995;

http://sound.media.mit.edu/resources/KEMAR.html) to create illusory spatial origins along the azimuth.

Audiovisual stimuli were combinations of the above-described auditory and visual component stimuli presented simultaneously (maximum audiovisual asynchronies < 2 ms). Audiovisual stimuli could be presented as spatially congruent ($\Delta AV = 0°$) or incongruent ($\Delta AV = \pm X°$, with the visual component stimulus location shifted by $+\frac{1}{2}\Delta AV$ and the auditory component stimulus shifted by $-\frac{1}{2}\Delta AV$ relative to the reported audiovisual stimulus location). The audiovisual disparity sizes ($X°$) were individually adjusted for each participant (see Section A.3.1.2). Stimulus presentation was controlled using Psychtoolbox 3.0.12 (Brainard, 1997; Kleiner, Brainard & Pelli, 2007; www.psychtoolbox.org) running on MATLAB R2016a (www.mathworks.com).

Participants were seated behind a table in a dark room with their chin on a chinrest placed at a distance of 75 cm from the screen. Prior to the first stimulus onset participants fixated a central grey cross (1° diameter) with luminance equal to the centre of the visual stimuli (for a duration of (750 – 1250 ms, randomly jittered). Fixation was monitored by means of a desktop mount Eyelink 1000 eye tracker (www.sr-research.com) that was calibrated before the start of each block of trials. Corrective feedback regarding fixation accuracy was provided after each block.

A.3. Experimental procedure

The study consisted of four 2.5 hour sessions that were performed on four separate days. The main experiment took place in sessions 2-4, whereas session 1 was used primarily for the calibration of particular stimulus settings. In the following we will separately describe the series of experimental parts in the first and latter three sessions.

A.3.1. First session:

A.3.1.1. (Familiarization). The 2IFC task for each trial was explained to participants through a short presentation that stressed the importance of accurate confidence judgments. To enhance participants' forced-fusion assumptions they were told that the auditory and visual components of the audiovisual stimuli were always presented at the same spatial location, "as if somebody hit the back of the screen with a metal stick (where the visual blobs represent the stick's imprint during the hits), so it would be wise to localize these audiovisual events based on both sensory modalities". Participants then completed a short familiarization series that included A, V and $AV_{\Delta=0^\circ}$ trials (6 trials x 3 conditions x 12 locations: ±1°, ±4°, ±7°, ±10°, ±13°, ±15°; three visual reliability levels were pseudo-randomized with σ_{blob} either 5°, 10° or 15°). After every response participants were given immediate corrective feedback on their location response, i.e. a green/red circle was presented on the screen to indicate a correct/incorrect response (200 ms duration).

A.3.1.2. (Auditory reliability measurement). This experimental part consisted of two smaller parts. Participants first completed a series of A trials (20 trials x 13 locations: 0°, ±1°, ±2°, ±3°, ±5°, ±7°, ±10°). Participants that obtained an accuracy of less than 90% for those forty trials on which the probe was presented at ±10° azimuth were excluded from further participation in the study. For each participant we fitted a cumulative Gaussian psychometric function to the fractions of 'perceived right' responses (Palamedes toolbox 1.8.2 for Matlab; www.palamedestoolbox.org). The maximum likelihood estimate of the JND parameter served to indicate the auditory spatial reliability.

This auditory JND was subsequently used for individually adjusting the following three stimulus settings: i. Thirteen individualized probe locations were defined as: $JND * (0, \pm 0.5, \pm 1, \pm 1.5, \pm 2, \pm 2.5, \pm 3)$, rounded to 0.5° under the constraint that the 13 locations were unique. (ii) Visual stimulus sizes (σ_{blob}) were individually adjusted for each participant to match the unisensory reliabilities, i.e. $JND_V \approx JND_A$ (see Section A.3.1.3). (iii) The audiovisual disparity was set equal to the auditory JND (rounded to the nearest 1° visual angle) because this setting has been shown (through simulations; Meijer & Noppeney, 2018) to provide a good balance between being able to investigate reliability-weighted integration (outside the scope of this report) while not violating participants' forced-fusion assumptions, thus minimizing the probability that participants perform causal inference instead of simple multisensory integration (Körding, Beierholm, Ma et al., 2007; Rohe and Noppeney, 2015a).

The second part of the 'auditory reliability measurement' is a refinement of the first part's measurement by using the individualized JND-based locations (see point i above) for a second series of A trials (20 trials x 13 individualized locations). The updated auditory JND, obtained from a second psychometric function fit, was used in all further tasks (see points i, ii, and iii above).

A.3.1.3. (Visual reliability adjustment). This experimental part aimed to match the visual reliability to the auditory reliability for each participant. First we obtained the hypothetical probe locations that would have led to auditory accuracy levels of 68%, 79% and 87% (from the fitted auditory psychometric function; see above). Then we used these three location pairs ($\pm X$°) as visual probe locations at which we aimed to match the accuracy to aforementioned levels by increasing σ_{blob} if accuracy was too high and

decreasing σ_{blob} if accuracy was too low. For each location pair we used two adaptive weighted up/down staircase procedures (Kingdom and Prins, 2016), one starting at an artificially high σ_{blob} (25°) and one starting at a very low value for σ_{blob} (3°). The six staircases were presented interleaved. Each staircase was terminated after thirty reversals. For each staircase σ_{blob} was computed pooled over the last 20 reversals. For each participant we identified which of the six staircases provided the estimate that was most distant from the pooled σ_{blob} across all six staircases. To attenuate effects of potential outliers, we discarded this estimate and then computed the final pooled σ_{blob} across the remaining five staircases (i.e. $\sqrt{\frac{1}{n} * \sum({\sigma_{blob}}^2)}$, with n = 5 staircases x 20 reversals). This final σ_{blob} estimate served as the medium visual reliability level. High and low visual reliability levels were created by multiplying σ_{blob} with 0.75 and 1.25, respectively.

A.3.1.4. (Visual reliability confirmation). To confirm that the unisensory reliabilities were successfully matched participants completed a series of V trials with the individualized blob sizes (medium visual reliability) from the staircase procedure above (20 trials x 13 individualized locations). A psychometric function was then fitted to the visual location responses in order to obtain an estimate of the visual reliability. If the difference between the two unisensory reliabilities was too large, i.e. if ${JND_V}^2 < 0.5 * {JND_A}^2$ or ${JND_V}^2 > 2 * {JND_A}^2$, we would consider the participant to be unreliable with respect to their localization performance and therefore exclude him/her from the study at this stage (but this did not occur: unisensory reliabilities were sufficiently matched for all participants).

A.3.2. Second, third and fourth session:

A.3.2.1. (Familiarization repetition). At the beginning of sessions 2-4, participants took part in a repetition of the familiarization run (with feedback after every trial, see Section A.3.1.1) to minimize variability in perceptual reliability and task performance across sessions (6 trials x 3 conditions x 12 individualized probe locations; one of the three individualized visual reliability levels was selected at random for V and $AV_{\Delta=0^\circ}$ trials).

A.3.2.2. (Main experiment). The main experiment was subdivided into eighteen blocks of trials. Six such blocks were completed in each of three sessions (2-4). Each block consisted of three mini-blocks, where each mini-block contained trials of one modality: A, V or AV. The order of the mini-blocks within a block was pseudo-randomized. The upcoming sensory modality for each mini-block was announced on-screen immediately prior to its start. In AV mini-blocks, the order of congruent ($AV_{\Delta=0^\circ}$) and incongruent ($AV_{\Delta=\pm X^\circ}$) trials was randomized. Furthermore, in V and AV mini-blocks trials from the three visual reliability levels were also presented in random order. All A mini-blocks contained 52 trials (4 trials x 13 locations), V mini-blocks contained 78 trials (2 trials x 13 locations x 3 reliability levels), and AV mini-blocks contained 234 trials (2 trials x 13 locations x 3 reliability levels x 3 conditions: one congruent and two incongruent). Within one block (i.e. A + V + AV = 364 trials) there were three 20 second mini-breaks (after every 91 trials) wherein the participant was given the chance to relax but keep the chin on the chin-rest. Motivational quotes (unrelated to the current experiment) were shown on screen during the mini-breaks. The program continued automatically after each mini-break by announcing the upcoming sensory modality. In between two blocks of trials participants were encouraged to take longer breaks.

After each block of trials the mean confidence per location was shown on screen, separately for A, V and AV conditions (means were computed across all three reliability levels and spatially incongruent AV conditions were not included). Critically, the figure axes did not show any values (i.e. there were no tick-labels) and the plots were scaled between minimum and maximum such that the confidence levels could not be compared across conditions. These figures served as feedback for both participant and experimenter. A 'v-shape' meant that confidence was generally higher for more eccentric probe locations (averaged across correct and incorrect type-1 responses). This pattern would be expected and participants were encouraged to introspect more in successive blocks if no clear v-shapes were visible.

Importantly, only data from the main experiment was used for the analyses of confidence judgements (Section 2.2). However, we note that participants were also required to report confidence judgments (by means of mouse responses on the butterfly diagram, see Figure 1) in all other tasks. Although these confidence responses served no immediate purpose, this was done for participants to get accustomed to the response mechanism and to practice metacognitive reflection on their location responses. After finishing the auditory and visual localization tasks in session 1, see Sections A.3.1.2 and A.3.1.4, the program had shown similar v-shape feedback plots to encourage participants to introspect and to let them know that their responses were being monitored.

Appendix B: Derivation of confidence level probabilities for ideal observers and meta-JND model fitting practicalities

B.1 Audiovisual location inference

When a stimulus is presented at its true location S, a brain's internal representation of that location, x, is assumed to be corrupted by Gaussian noise. For any audiovisual stimulus, we assume that the added noise for both sensory modalities independent. The generative model thus posits that internal auditory (A) and visual (V) location representations are separately sampled from normally-distributed sensory noise probability density functions (PDFs), each with its own variance and centred on the true stimulus location: $x_V \sim N(x_V;\ S_V, {\sigma^2}_V)$ and $x_A \sim N(x_A;\ S_A, {\sigma^2}_A)$. In what follows below we will often refer to the sensory noise-induced variance as ${\sigma^2}_S$, whether the trial was auditory (${\sigma^2}_S = {\sigma^2}_A$), visual (${\sigma^2}_S = {\sigma^2}_V$), or audiovisual (${\sigma^2}_S = {\sigma^2}_{AV}$).

We assume that an ideal observer has "learned" the generative model and inverts it to obtain the best estimate of the stimulus location. Knowing internal representation x, the likelihood that a stimulus at location S has caused this particular x is given by the likelihood function (centred on x): $P(x|S) = N(S;\ x, {\sigma^2}_S)$. The audiovisual likelihood function, assuming independent sensory noise and one source location (S_{AV}) for both x_A and x_V, is the product of the two unisensory likelihood functions (Ernst & Banks, 2002):

$$P(x_{AV}|S_V, S_A) = P(x_A|S_A)P(x_V|S_V) = N(S_{AV};\ x_{AV}, {\sigma^2}_{AV})$$

Where: $x_{AV} = \frac{{\sigma^2}_A}{{\sigma^2}_A + {\sigma^2}_V} x_V + \frac{{\sigma^2}_V}{{\sigma^2}_A + {\sigma^2}_V} x_A$ and ${\sigma^2}_{AV} = \frac{{\sigma^2}_V {\sigma^2}_A}{{\sigma^2}_V + {\sigma^2}_A}$

To compute the posterior probability for S being the stimulus location that gave rise to internal representation x, Bayes' rule states that one should multiply likelihood with prior probability. We assume that observers hold a normally-distributed prior probability distribution over all locations S, with mean μ_P (e.g. 0° for a central bias; Kerzel, 2002) and variance ${\sigma^2}_P$: $P(S) = N(S; \mu_P, {\sigma^2}_P)$. The posterior probability distribution is another Gaussian:

$$P(S|x) = N(S;\ w_S x + w_P \mu_P, w_S {\sigma^2}_S)$$

Where: $w_S = \frac{{\sigma^2}_P}{{\sigma^2}_S + {\sigma^2}_P}$ and $w_P = \frac{{\sigma^2}_S}{{\sigma^2}_S + {\sigma^2}_P}$

The best estimate for the stimulus location is the maximum (mean) of this posterior distribution:

$$\hat{S} = \arg max\, P(S|x) = w_S x + w_P \mu_P$$

In a two-interval forced-choice task, such as the one used in the current study, every trial contains two stimuli, standard and probe. Importantly, the amount of sensory noise for standard and probe is identical: ${\sigma^2}_{S,st} = {\sigma^2}_{S,pr}$. Participants make responses based on the estimated difference between probe and standard (we use subscripts pr and st, respectively):

$$\widehat{\Delta S} = \hat{S}_{pr} - \hat{S}_{st} = \left(w_S x_{pr} + w_P \mu_P\right) - \left(w_S x_{st} + w_P \mu_P\right) = w_S\left(x_{pr} - x_{st}\right) = w_S \Delta x$$

Alternatively, we can express $\widehat{\Delta S}$ as the maximum (mean) of the posterior probability distribution for stimulus disparity ΔS $(= S_{pr} - S_{st})$, which is equal to the normal difference distribution of both posteriors, $P\left(S_{pr}\middle|x_{pr}\right)$ and $P(S_{st}|x_{st})$:

$$P\left(\Delta S \middle| x_{st}, x_{pr}\right) = N\left(\Delta S; \widehat{\Delta S}, 2{w_S}{\sigma^2}_S\right) \qquad \textbf{(Eq. B.1)}$$

Finally, location response z is determined by whether disparity estimate $\widehat{\Delta S}$ is greater or smaller than an internal type-1 criterion that we call C_1:

$$P\left(z = "right" \middle| x_{st}, x_{pr}\right) = \begin{cases} 1 & if \quad \widehat{\Delta S} \geq C_1 \\ 0 & if \quad \widehat{\Delta S} < C_1 \end{cases}$$

$$P\left(z = "left" \middle| x_{st}, x_{pr}\right) = \begin{cases} 1 & if \quad \widehat{\Delta S} < C_1 \\ 0 & if \quad \widehat{\Delta S} \geq C_1 \end{cases} \qquad \textbf{(Eq. B.2)}$$

B.2 Accounting for biases

We assume that participants keep their type-1 criterion constant at $C_1 = 0°$. Any left/right bias is instead modelled by a bias in the generation of internal estimates x. This is a logical choice when one considers spatially incongruent audiovisual stimuli: i.e. $S_V \neq S_A$ (see below).

Assuming statistically optimal inference (as above) the PDF for the integrated signals x_{AV} given a pair of auditory and visual stimuli locations, S_V and S_A, is equal to the product of the unisensory PDFs:

$$P(x_{AV}|S_V, S_A) = P(x_V|S_V)P(x_A|S_A) = N(x_{AV}; \frac{{\sigma^2}_A}{{\sigma^2}_A + {\sigma^2}_V} S_V + \frac{{\sigma^2}_V}{{\sigma^2}_A + {\sigma^2}_V} S_A, {\sigma^2}_{AV})$$

In the current study audiovisual standard stimuli are always presented in the centre of the screen and spatially congruent: $S_{st,V} = S_{st,A} = S_{st} = 0°$. Therefore, their PDF readily simplifies to: $P\left(x_{st,AV} \middle| S_{st}\right) = N(x_{AV};\ S_{st}, {\sigma^2}_{AV})$. However, the audiovisual probe could be presented with a small spatial disparity, ΔAV, according to: $S_{pr,V} = S_{pr} + \frac{1}{2}\Delta AV$ and $S_{pr,A} = S_{pr} - \frac{1}{2}\Delta AV$. The PDF for $x_{pr,AV}$ should thus be written as:

$$P\left(x_{pr,AV}\middle|S_{pr}\right) = N(x_{AV};\ \frac{{\sigma^2}_A}{{\sigma^2}_A+{\sigma^2}_V}(S_{pr}+\frac{1}{2}\Delta AV)+\frac{{\sigma^2}_V}{{\sigma^2}_A+{\sigma^2}_V}(S_{pr}-\frac{1}{2}\Delta AV), {\sigma^2}_{AV})$$

$$= N(x_{AV};\ S_{pr}+Bias_{pr}, {\sigma^2}_{AV})$$

Where: $$Bias_{pr} = \frac{1}{2}\Delta AV(\frac{{\sigma^2}_A-{\sigma^2}_V}{{\sigma^2}_A+{\sigma^2}_V}) \quad \textbf{(Eq. B.3)}$$

Please note that we find a left- or rightward bias (i.e. $\left|Bias_{pr}\right| > 0°$) for the internal representations of a spatially incongruent ($\Delta AV > 0°$) audiovisual probe's if ${\sigma^2}_A \neq {\sigma^2}_V$. The bias for audiovisual spatially congruent ($\Delta AV = 0°$) or unisensory (A or V) probes is assumed to be zero. Similarly, the bias for the audiovisual standard is also assumed to be zero: i.e. $Bias_{st} = 0°$.

B.3 Bayesian confidence judgements

An ideal performer's confidence report is given by the posterior probability of being correct (Drugowitsch et al., 2014, Hangya et al., 2016; Pouget et al., 2016): i.e. the probability that the true stimulus disparity corresponds to the response that was given.

$$Confidence =$$

$$\begin{cases} P\left(\Delta S \geq C_1\middle|z = "right", x_{st}, x_{pr}\right) = \int_{C_1}^{\infty} P\left(\Delta S\middle|z = "right", x_{st}, x_{pr}\right)d\Delta S \\ P\left(\Delta S < C_1\middle|z = "left", x_{st}, x_{pr}\right) = \int_{-\infty}^{C_1} P\left(\Delta S\middle|z = "left", x_{st}, x_{pr}\right)d\Delta S \end{cases}$$

The abovementioned $P\left(\Delta S\middle|z, x_{st}, x_{pr}\right)$, i.e. posterior probability over stimulus disparities ΔS given internal representations x_{st}, x_{pr} and response z, is equal to $P\left(\Delta S\middle|x_{st}, x_{pr}\right)$ for all $P(z|x_{st}, x_{pr}) \neq 0$ because $P(z|x_{st}, x_{pr})$ and $P(\Delta S|x_{st}, x_{pr})$ are conditionally independent. We conclude that an ideal performer's confidence judgment is defined by the area under the curve of the posterior distribution over ΔS (Eq. B.1) on the responded side ("right" or "left") of the type-1 criterion C_1 (Eq. B.2):

$$Confidence|\ z = "right", x_{st}, x_{pr} = \int_{C_1}^{\infty} N\left(\Delta S; \widehat{\Delta S}, 2{w_S}{\sigma^2}_S\right) d\Delta S$$

$$Confidence|\ z = "left", x_{st}, x_{pr} = \int_{-\infty}^{C_1} N\left(\Delta S; \widehat{\Delta S}, 2{w_S}{\sigma^2}_S\right) d\Delta S$$

Analytical solutions for these integrals are given by the (complementary) error function: erf and $erfc$, respectively. We use the symmetry of both functions and assumption $C_1 = 0°$ to obtain:

$$Confidence|\ x_{st}, x_{pr} = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{|\widehat{\Delta S}|}{\sqrt{2}\sqrt{2{w_S}{\sigma^2}_S}}\right)\right] \qquad \textbf{(Eq. B.4)}$$

So, confidence in a Bayesian sense, i.e. the posterior probability of being correct, is lowest when $\widehat{\Delta S} = C_1 = 0°$ (i.e. $Confidence = 0.5$) and continuously increases for both sides symmetrically (i.e. for left and right responses). The relationship between $Confidence$ and absolute stimulus disparity estimate $\left|\widehat{\Delta S}\right|$ is described by a cumulative Gaussian distribution whose slope depends on the stimulus reliability (i.e. $\frac{1}{{\sigma^2}_S}$) and on the width of the prior stimulus distribution $P(S)$ by means of the weighting term w_S.

By requiring observers to select one of several confidence levels (e.g. 1-4) we implicitly ask them to map their probabilities of being correct onto a confidence scale with discrete levels. Confidence level k is chosen when the probability of being correct is within that level's type-2 $Confidence$ criteria: e.g. "$ConfLevel = k$" if $0.7 < Confidence < 0.8$. Given that there is a continuous and unique theoretical mapping between $\left|\widehat{\Delta S}\right|$ and $Confidence$, we can alternatively describe the type-2 criteria in units of $\left|\widehat{\Delta S}\right|$. For a task that employs N confidence levels, there exist N+1 type-2 criteria $\boldsymbol{C_2}$ that group the absolute disparity estimates $\left|\widehat{\Delta S}\right|$ onto confidence levels according to:

$$"ConfLevel = k" \quad \text{if} \quad C_{2,k} \leq \left|\widehat{\Delta S}\right| \leq C_{2,k+1}$$

With: $\boldsymbol{C_2} = \{C_{2,1} = 0° \leq C_{2,2} \ldots \leq C_{2,N} \leq C_{2,N+1} = \infty°\}$

Please note that the first and last type-2 criteria are fixed at $C_{2,1} = 0°$ and $C_{2,N+1} = \infty°$, such that an individual's mapping between $\left|\widehat{\Delta S}\right|$ is determined by N-1 type-2 criteria.

B.4 Predicting probabilities of location responses

Given the above-described inference model for statistically optimal location responses and confidence judgments, we can now derive the predicted probability with which an ideal observer would make a particular response, e.g. $z =$ "$right$" and "$ConfLevel = k$", for a certain stimulus disparity $\Delta S = S_{pr} - S_{st}$, type-1 performance measures, $Bias_{pr}$ and ${\sigma^2}_S$, and type-2 criteria $\boldsymbol{C_2}$.

We start with the predicted probabilities for location responses $P\left(z\middle|S_{st}, S_{pr}\right)$. Since the location responses depend entirely on the stimulus disparity estimates $\widehat{\Delta S} = \hat{S}_{pr} - \hat{S}_{st}$ (see Eq. B.2), we will first derive the (across-trials) probability distributions for single stimuli estimates (i.e. $\hat{S}_{st}$ or $\hat{S}_{pr}$). These can be obtained by integrating out (i.e. marginalizing over) x:

$$P\left(\hat{S}\middle|S\right) = \int P\left(\hat{S}\middle|x\right)P(x|S)dx$$

where $P(x|S)$ is the sensory noise PDF centred on S plus a potentially non-zero bias (see Eq. B.3), and $P\left(\hat{S}\middle|x\right)$ can be expressed by means of the delta function: i.e. $P(\hat{S}|x) = 1$ if $\hat{S} = w_S x + w_P \mu_P$, and 0 otherwise. Solving the integral results in:

$$P\left(\hat{S}\middle|S\right) = \int \delta\left(w_S x + w_P \mu_P - \hat{S}\right)N\left(x; S + Bias_{pr}, {\sigma^2}_S\right)dx$$

$$= N(\hat{S}; w_S\left(S + Bias_{pr}\right) + w_P \mu_P,\ {w^2}_S {\sigma^2}_S)$$

Since both $P(\widehat{S_{pr}}|S_{pr})$ and $P(\widehat{S_{st}}|S_{st})$ are normal distributions, their difference distribution $P(\widehat{\Delta S}|S_{st}, S_{pr})$ will also be normally distributed:

$$P(\widehat{\Delta S}|S_{st}, S_{pr}) = N(\widehat{\Delta S}; (w_S(S_{pr} + Bias_{pr}) + w_P\mu_P) - (w_S S_{st} + w_P\mu_P), 2{w^2}_S{\sigma^2}_S)$$

$$= N(\widehat{\Delta S}; w_S(\Delta S + Bias_{pr}), 2{w^2}_S{\sigma^2}_S) \qquad \textbf{(Eq. B.5)}$$

Finally, the probability of a "left" or "right" response given both stimulus locations is computed as the partial integral of $P(\widehat{\Delta S}|S_{st}, S_{pr})$ to the left or right of $C_1 = 0°$:

$$P(z = "right"|S_{st}, S_{pr}) = P(\widehat{\Delta S} \geq C_1|S_{st}, S_{pr}) = \int_0^\infty P(\widehat{\Delta S}|S_{st}, S_{pr}) d\widehat{\Delta S}$$

$$= \frac{1}{2}\left[1 + \text{erf}(\frac{w_S(\Delta S + Bias_{pr})}{\sqrt{2}\sqrt{2{w^2}_S{\sigma^2}_S}})\right] = \frac{1}{2}\left[1 + \text{erf}(\frac{\Delta S - PSE}{\sqrt{2}JND})\right]$$

$$P(z = "left"|S_{st}, S_{pr}) = P(\widehat{\Delta S} < C_1|S_{st}, S_{pr}) = \int_{-\infty}^0 P(\widehat{\Delta S}|S_{st}, S_{pr}) d\widehat{\Delta S}$$

$$= \frac{1}{2}\text{erfc}(\frac{w_S(\Delta S + Bias_{pr})}{\sqrt{2}\sqrt{2{w^2}_S{\sigma^2}_S}}) = \frac{1}{2}\text{erfc}(\frac{\Delta S - PSE}{\sqrt{2}JND}) \qquad \textbf{(Eq. B.6)}$$

Where we have substituted $JND = \sqrt{2}\sigma_S$ (for the just-noticeable difference) and $PSE = -Bias_{pr}$ (for the point of subjective equality). Equation B.6 is commonly known as the psychometric function. We note that the psychometric function is not affected by the prior stimulus distribution $P(S)$ because the weighting term w_S drops out in the last step (Acuna et al., 2015).

B.5 Predicting probabilities of confidence judgments

The main question that we set out to answer is: Given a certain location response z for true stimulus locations S_{st} and S_{pr}, what is the probability with which an ideal observer

would choose any of the confidence levels? In other words, we want to know $P("ConfLevel = k"|z, S_{st}, S_{pr})$.

Since the decision for a certain confidence level depends directly on $\widehat{\Delta S}$, the probability for "$ConfLevel = k$" is equal to the partial integral of $P(\widehat{\Delta S}|z, S_{st}, S_{pr})$ over all $\widehat{\Delta S}$ where $P(\widehat{\Delta S}|z) > 0$ and $C_{2,k} \leq |\widehat{\Delta S}| \leq C_{2,k+1}$. We use Bayes' Rule to find:

$$P(\widehat{\Delta S}|z, S_{st}, S_{pr}) = \frac{P(z|\widehat{\Delta S}, S_{st}, S_{pr}) * P(\widehat{\Delta S}|S_{st}, S_{pr})}{P(z|S_{st}, S_{pr})} \qquad \textbf{(Eq. B.7)}$$

Where two of the right-hand terms, $P(\widehat{\Delta S}|S_{st}, S_{pr})$ and $P(z|S_{st}, S_{pr})$ are known (see Eqs B.5 & B.6) and the third term can be shown to equal $P(z|x_{st}, x_{pr})$, which is also known (see Eq. B.2):

$$P(z|\widehat{\Delta S}, S_{st}, S_{pr}) = \frac{P(z, \widehat{\Delta S}|S_{st}, S_{pr})}{P(\widehat{\Delta S}|S_{st}, S_{pr})} = \frac{P(z|\widehat{\Delta S}) * P(\widehat{\Delta S}|S_{st}, S_{pr})}{P(\widehat{\Delta S}|S_{st}, S_{pr})} = P(z|\widehat{\Delta S}) = P(z|x_{st}, x_{pr})$$

The numerator on the right-hand side of equation B.7 is the product of a step function and a Gaussian distribution over $\widehat{\Delta S}$. Division by the denominator ensures that this truncated Gaussian is normalized to form the required probability distribution $P(\widehat{\Delta S}|z, S_{st}, S_{pr})$. The probability for a particular confidence level k, given response z, can thus be computed as a partial integral of the truncated Gaussian probability distribution. By substituting B.5 & B.6 into Eq. B.7 we find:

$$P("ConfLevel = k"|z = "right", S_{st}, S_{pr}) = \int_{C_{2,k}}^{C_{2,k+1}} \frac{N(\widehat{\Delta S}; w_S(\Delta S + Bias_{pr}), 2w^2{}_S\sigma^2{}_S)}{\frac{1}{2}\left[1 + \mathrm{erf}(\frac{\Delta S - PSE}{\sqrt{2}JND})\right]} d\widehat{\Delta S}$$

$$= \frac{\frac{1}{2}\left[1+\mathrm{erf}\left(\frac{C_{2,k+1}-w_S(\Delta S+Bias_{pr})}{\sqrt{2}\sqrt{2w^2{}_S\sigma^2{}_S}}\right)\right]-\frac{1}{2}\left[1+\mathrm{erf}\left(\frac{C_{2,k}-w_S(\Delta S+Bias_{pr})}{\sqrt{2}\sqrt{2w^2{}_S\sigma^2{}_S}}\right)\right]}{\frac{1}{2}\left[1+\mathrm{erf}\left(\frac{\Delta S+Bias_{pr}}{\sqrt{2}\sqrt{2}\sigma_S}\right)\right]}$$

$$= \frac{\mathrm{erf}\left(\frac{\frac{1}{w_S}C_{2,k+1}-\Delta S+PSE}{\sqrt{2}JND}\right)-\mathrm{erf}\left(\frac{\frac{1}{w_S}C_{2,k}-\Delta S+PSE}{\sqrt{2}JND}\right)}{1+\mathrm{erf}\left(\frac{\Delta S-PSE}{\sqrt{2}JND}\right)} \qquad \textbf{(Eq. B.8)}$$

Likewise, for $z = "left"$ responses the probability of a "$ConfLevel = k$" judgment is:

$$P("ConfLevel = k"|z = "left", S_{st}, S_{pr}) = \int_{-C_{2,k+1}}^{-C_{2,k}} \frac{N\left(\widehat{\Delta S}; w_S(\Delta S+Bias_{pr}), 2w^2{}_S\sigma^2{}_S\right)}{\frac{1}{2}\mathrm{erfc}\left(\frac{\Delta S-PSE}{\sqrt{2}JND}\right)} d\widehat{\Delta S}$$

$$= \frac{\mathrm{erf}\left(\frac{\frac{-1}{w_S}C_{2,k}-\Delta S+PSE}{\sqrt{2}JND}\right)-\mathrm{erf}\left(\frac{\frac{-1}{w_S}C_{2,k+1}-\Delta S+PSE}{\sqrt{2}JND}\right)}{\mathrm{erfc}\left(\frac{\Delta S-PSE}{\sqrt{2}JND}\right)} \qquad \textbf{(Eq. B.9)}$$

Equations B.8 & B.9 suggest that the prior stimulus distribution $P(S)$ exerts its influence on the probability of a "$ConfLevel = k$" judgment by means of the terms $\frac{1}{w_S}C_{2,k}$, $\frac{1}{w_S}C_{2,k+1}$ and $\frac{-1}{w_S}C_{2,k}$, $\frac{-1}{w_S}C_{2,k+1}$. However, we note that $\boldsymbol{C_2}$ criteria are defined in units of $\widehat{\Delta S} = w_S\Delta x$. So $\frac{1}{w_S}C_2$ is defined in units of Δx. We may thus conclude that the width of the prior stimulus distribution does *not* affect the across-trials probability for a certain confidence level. In other words: if we alternatively define a set of type-2 criteria in units of Δx , $\boldsymbol{C_{2\Delta x}}$, which differ from $\boldsymbol{C_2}$ by a factor w_S, i.e. $\boldsymbol{C_2} = w_S\boldsymbol{C_{2\Delta x}}$, then we can rewrite equations B.8 & B.9 such that the unknown weighting term w_S drops out:

$$P("ConfLevel = k"|z = "right", S_{st}, S_{pr}) = \frac{\mathrm{erf}\left(\frac{C_{2\Delta x,k+1}-\Delta S+PSE}{\sqrt{2}JND}\right)-\mathrm{erf}\left(\frac{C_{2\Delta x,k}-\Delta S+PSE}{\sqrt{2}JND}\right)}{1+\mathrm{erf}\left(\frac{\Delta S-PSE}{\sqrt{2}JND}\right)}$$

$$P(\text{"}ConfLevel = k\text{"}|z = \text{"}left\text{"}, S_{st}, S_{pr}) = \frac{\text{erf}\left(\frac{-C_{2\Delta x,k} - \Delta S + PSE}{\sqrt{2}JND}\right) - \text{erf}\left(\frac{-C_{2\Delta x,k+1} - \Delta S + PSE}{\sqrt{2}JND}\right)}{\text{erfc}\left(\frac{\Delta S - PSE}{\sqrt{2}JND}\right)}$$

(Eq. B.10)

Equation B.10 enables us to directly predict the probability with which an ideal observer would choose confidence level k for a certain stimulus disparity $\Delta S = S_{pr} - S_{st}$ and conditional on the type-1 response ($z = \text{"}right\text{"}$ or $z = \text{"}left\text{"}$), type-1 performance measures PSE and JND, and type-2 criteria $\boldsymbol{C_{2\Delta x}}$.

Note that the surrogate criteria $\boldsymbol{C_{2\Delta x}}$ differ from the actual criteria $\boldsymbol{C_2}$ by a factor $w_S = \frac{{\sigma^2}_P}{{\sigma^2}_S + {\sigma^2}_P}$. If ${\sigma^2}_P$ is large, that is if the prior is weak, $w_S \approx 1$. For stronger priors one should keep in mind that the type-2 criteria $\boldsymbol{C_{2\Delta x}}$ were scaled by $0 < w_S < 1$ relative to the true criteria. For our purposes in the current study such scaling is not important, because we make comparisons of $\boldsymbol{C_{2\Delta x}}$ differences between conditions, not between different (groups of) participants.

B.6 Practicalities of meta-JND model parameter fitting

Our aim is to fit a meta-JND to participants' confidence judgments. In practical terms this means that we predict the conditional probabilities of the confidence levels for an ideal observer using equation B.10 while we try to optimize the parameter values of JND and $\boldsymbol{C_{2\Delta x}}$ (n.b. the PSE is obtained from the type-1 psychometric function and held constant) to let the generated probabilities best match the empirical probabilities in the participant's data (by means of maximizing the likelihood, see further below). The best matching value for JND, fitted to the confidence judgments of participants, is what we call meta-JND.

Apart from meta-JND, other parameters to fit are the N-1 confidence bin criteria (in units of Δx): $C_{2\Delta x,2}$ … $C_{2\Delta x,N}$ (where $C_{2\Delta x,1}$ and $C_{2\Delta x,N+1}$ are fixed at 0 and ∞, respectively). We constrain the parameter fitting algorithm (Matlab's 'fmincon' function) such that:

$$\boldsymbol{C_{2\Delta x}} = \{C_{2\Delta x,1} = 0 \leq C_{2\Delta x,2} \ldots \leq C_{2\Delta x,N} \leq C_{2\Delta x,N+1} = \infty\}$$

Finally, we expect participants to occasionally erroneously select a confidence level at random. Therefore, we also fit one so-called type-2 lapse-rate parameter: λ_2. To compute the ideal observer's confidence bin probabilities (conditional on type-1 response z and stimulus locations S_{pr}, S_{st}) we insert the PSE, JND (i.e. the meta-JND that we're fitting), and $\boldsymbol{C_{2\Delta x}}$ parameters into equation B.10 and then adjust the obtained confidence level probability for lapses by applying:

$$P_{bin,k}(k, z, S_{st}, S_{pr}) = (1 - \lambda_2)P("ConfLevel = k"|z, S_{st}, S_{pr}) + \lambda_2 \frac{1}{N}$$

$P_{bin,k}$ is computed for each combination of confidence level k, type-1 response z and stimulus locations S_{st} and S_{pr} (given certain parameters values for λ_2, JND and $\boldsymbol{C_{2\Delta x}}$). The likelihood of a particular behavioral dataset (i.e. a set of confidence level responses conditional on z, S_{st}, S_{pr}) given the model and its parameters is the product of $P_{bin,k}$ across all trials:

$$P(data|\lambda_2, JND, \boldsymbol{C_{2\Delta x}}) = \prod_{i=1}^{nTrials} P_{bin,k}(k_i, z_i, S_{st,i}, S_{pr,i})$$

This likelihood is maximized by 'optimizing' the parameter values for JND, $\boldsymbol{C_{2\Delta x}}$, and λ_2.

CHAPTER 6: GENERAL DISCUSSION

In this thesis I have presented: 1) An evaluation of two popular computational models for multisensory integration, maximum likelihood estimation (MLE) and Bayesian causal inference (BCI), by means of a literature review. 2) A failed attempt to replicate one of the most influential multisensory integration studies supporting the MLE model (Alais & Burr, 2004). 3) Evidence that the BCI model describes human multisensory perception better, even under experimental conditions that were optimized for MLE. 4) Empirical support for a supra-modal view on metacognition in which multisensory integration leads to a boost of confidence.

In this final chapter I will elaborate on the implications of the empirical results that were presented in this thesis. Furthermore, I will address some outstanding questions and limitations of the current work with regards to multisensory integration according to BCI as we have proposed it in chapter 4. Finally, I will propose future research directions that combine metacognition and multisensory research but which we were unable to tackle in the current thesis.

1. Implications of the current research

The first empirical chapter (3) attempts to resolve a fifteen year old discrepancy in the multisensory research literature: while Alais and Burr (2004) found that human perceptual behavior for audiovisual spatial integration agrees well with the statistically

optimal strategy of maximum likelihood estimation (MLE), Battaglia, Jacobs and Aslin (2003) reported that human observers systematically overweighted the visual sensory modality relative to MLE predictions. In chapter 3 we have shown results that partially reproduce both of the previous works. Similar to Alais and Burr (2004), we found a multisensory variance reduction in line with MLE predictions, thus indicating that the human brain is able to integrate two sensory sources of information near-optimally and gain a multisensory precision improvement. However, like Battaglia et al., (2003) we also observed a clear visual overweighting pattern, which demonstrates that human observers do not exactly adhere to MLE predictions. Such a discrepant combination of an optimal variance reduction with simultaneous overweighting of one sensory modality has been reported before (e.g. Butler, Smith, Campos & Bülthoff, 2010).

At first sight, our results thus only seemed to add more confusion to the ongoing debate of whether humans perform optimal sensory integration (see Rahnev & Denison, 2018). However, we have shown through Monte Carlo simulations (Appendix D in chapter 3) that an apparently optimal variance reduction is still likely to be obtained for moderate levels of sensory overweighting (i.e. moderate suboptimality). This simulation result avoids a black-and-white judgment about which of the previously reported results is more correct (optimal vs. suboptimal integration) and instead refocuses the debate on the explanatory power (or lack thereof) in the empirical data.

Our data and supplementary simulation analyses suggest that human observers deviated slightly from MLE-optimal multisensory integration. The reason for us being able to show this small but significant deviation whereas Alais and Burr (2004) generally found a good agreement with MLE for their participants' sensory weighting scheme

probably rests on our carefully designed experimental design with individualized settings and advanced analysis methods. Prior to study execution we made use of Monte Carlo simulations to select an audiovisual spatial disparity size that optimized the sensitivity to detect sensory overweighting (Appendix B in chapter 3). It is through such simulations and detailed preparation of the psychophysical methodology that we gain a better understanding about human perceptual behavior.

The second empirical chapter (4) goes further where the first empirical chapter (3) stopped. Though we have mentioned some possible explanations for the deviations from MLE that we observed for audiovisual spatial integration in chapter 3, we set out to systematically check and compare those hypotheses in chapter 4. The approach was two-fold: the first method was based on experimental manipulations, while the second explanatory method was performed through Bayesian model comparison. While the first method did not reveal any evidence for the influence of training or motivational reward on multisensory integration (though we acknowledge that the 2 x 2 between subjects experimental design was likely underpowered), advanced model comparison techniques (i.e. making use of Markov-Chain Monte Carlo; Acerbi, Dokka, Angelaki & Ma, 2018) helped us to identify the best explanation for the deviations from MLE: Bayesian causal inference. This result highlights the power of hypothesis-driven research and sensitive analysis methods for psychophysical research.

The finding that human observers perform Bayesian causal inference even under experimental conditions that were previously believed to invoke mandatory fusion does not only explain the empirical results in chapters 3 and 4, it could potentially also explain many previous reports of suboptimal (non-MLE) multisensory integration (extending

well outside the domain of audiovisual spatial integration; e.g. Burr, Banks & Morrone, 2009; Butler, Smith, Campos & Bülthoff, 2010; Fetsch, Turner, DeAngelis & Angelaki, 2009; Maiworm & Röder, 2011; Prsa, Gale & Blanke, 2012; Rosas, Wagemans, Ernst & Wichmann, 2005). The proposal that human observers make use of Bayesian causal inference is not new and enjoys wide support in the multisensory integration community (e.g. Beierholm, Quartz & Shams, 2009; Bosen et al., 2016; Körding, Beierholm, Ma et al., 2007; Magnotti & Beauchamp, 2017; Magnotti, Ma & Beauchamp, 2013; McGovern, Roudaia, Newell & Roach, 2016; Mendonça, Mandelli & Pulkki, 2016; Natarajan, Murray, Shams & Zemel, 2009; Odegaard & Shams, 2016; Odegaard, Wozny & Shams, 2015, 2017; Rohe & Noppeney, 2015a, 2015b, 2016, Sato, Toyoizumi and Aihara, 2007; de Winkel, Katliar & Bülthoff, 2017; Wozny, Beierholm & Shams, 2008, 2010; see section 4 in chapter 2). However, the option of causal inference has rarely been raised as a potential candidate explanation for suboptimal MLE results. The ruling idea in the field seemed to be that there were specific experimental settings in which participants applied MLE-type forced integration and other experimental paradigms in which participants relied on causal inference. The main conclusion that should be drawn from Chapter 4 is that this view does not hold in practice: human observers cannot be forced to integrate two sensory signals. Instead, they use a probability based strategy to decide whether or not to integrate (i.e. causal inference) in all situations. Experimenters should thus be aware that full multisensory integration on every trial cannot be expected.

The final empirical chapter (5) on metacognition may appear to have been a sudden change of topics (from optimal multisensory integration to metacognition), I argue instead that it is very much related to the preceding chapters because it builds on the

same computational foundation for multisensory integration: that it is a probabilistic mechanism that takes sensory uncertainties into account. In fact, I have argued that the empirical research that was presented in chapter 5 provides one of the most direct pieces of evidence to support such a probabilistic basis of the multisensory integration models that were discussed in the preceding chapters. The fact that we observed an unambiguous confidence boost for multisensory versus unisensory stimuli and that the extent of that confidence increase could be predicted by probabilistic multisensory integration modelling (based on Bayesian inference) is a clear confirmation of the fundamental ideas that underlie the most popular models for multisensory integration (i.e. MLE and BCI).

Other results that were presented in chapter 5, specifically on metacognitive noise being equal across various reliability levels and sensory modalities, further support the view that metacognition is a supra-modal process (i.e. independent of the sensory modality) that takes place after multisensory integration has completed (De Gardelle, Le Corre & Mamassian, 2016; Faivre, Filevich, Solovey, Kühn & Blanke, 2018). The current study thus contributes to the abundance of evidence in support of a hierarchical model for metacognition (Bang, Shekhar & Rahnev, 2018; De Martino, Fleming, Garrett & Dolan, 2013; Fleming et al., 2015; Jang, Wallsten & Huber, 2012; Maniscalco & Lau, 2016; Mueller & Weidemann, 2008; Rahnev, Nee, Riddle, Larson & D'Esposito, 2016; Van den Berg, Yoo & Ma, 2017).

While the above-mentioned conclusions from the third empirical chapter (5) may not be very surprising (uncertainty-based models for multisensory integration were already widely accepted) or novel (considering other reports in the literature on metacognition)

the methodology that enabled us to draw these conclusions certainly was new. In order to determine the multisensory confidence boost unambiguously we were forced to develop a new modelling approach. We extended the now popular meta-d' method by Maniscalco & Lau (2012), by deriving the predicted confidence ratings of a Bayesian ideal observer, for use in experimental paradigms that are normally used to fit psychometric functions. An evaluation of this modelling based approach, which we termed meta-JND, showed good to excellent goodness-of-fits, thereby confirming the method's validity for summarizing human metacognitive behavior. Although we applied this novel methodology to a multisensory integration experiment, it is not specific to such research and can be used by researchers in all areas of metacognition (the model fitting scripts will be made available online at a later stage).

2. Outstanding questions and limitations

In chapters 3 and 4 I have presented behavioral data that deviated only mildly from MLE predictions: i.e. a significant multisensory variance reduction was still observed in nearly all individual participants. However, in the pilot study of chapter 3 (appendix A) the audiovisual variance was higher than the auditory variance for all participants. What would explain such a major discrepancy? The answer may lie in the correlation that was observed between the difference in unisensory reliabilities ($\sigma_V - \sigma_A$) and the amount of visual overweighting ($w_{A,mle} - w_{A,emp}$), see Chapter 4 (Fig. 4). In the pilot study this difference was rather large, because the spatial reliability of the visual stimuli had been heavily degraded. According to the BCI model, such high levels of sensory noise would cause some of the internal visual estimates (x_V) to be very distant from the auditory

spatial estimates (x_A). Additionally, since the auditory reliability was relatively high, the estimated likelihood for a common source of such disparate signals would be small (see also Rohe & Noppeney, 2015a). So, I concluded in chapter 4, integration breaks down, and if the observer consistently selects the visual spatial estimate ($\hat{S}_V$) over the auditory spatial estimate ($\hat{S}_A$) then the audiovisual variance could almost be as high as the visual variance. The BCI model thus also provides a reasonable explanation for such clearly suboptimal multisensory perception.

However, the above explanation assumes that participants stick to the visual BCI estimates even though the auditory estimates are far more precise. This begs the question under what circumstances observers would switch to rely on their normally less dominant sensory modality. Studies by Jacobs and Fine (1999) and Ernst and Banks (2000) indicated that an observer's choice may be influenced by recent learning experience or task relevance (not mutually exclusive). In the experiments that were presented in this thesis, participants were required to select the audiovisual stimulus location. The task relevance hypothesis (Ernst & Banks, 2000) suggests that if we had instead asked for the location of the auditory stimulus, then we should have observed auditory overweighting relative to MLE predictions. Likewise, the experience-based hypothesis (Jacobs & Fine, 1999) suggests that if we had trained participants prior to the pilot study (Chapter 3, appendix A) to realise that the auditory stimuli were more reliable than the visual stimuli, then their audiovisual variance should have been smaller than (or similar to) the relatively small auditory variance. These are two clear predictions that can easily be tested experimentally. The fact that we have not done so yet is a

major limitation of the here-presented thesis (but see e.g. Roach, Heron & McGraw, 2006 for evidence that supports the task relevance hypothesis in the temporal domain).

While the observed divergence from MLE predictions best fitted with the BCI explanation (i.e. a violation of the forced-fusion assumption and reliance on the modality that is normally most precise, see chapter 4), we cannot altogether exclude other factors. The model that implemented visual overweighting (relative to MLE) by means of a Bayesian prior on visual reliability (Battaglia et al., 2003) performed nearly as good when compared in a fixed-effects model comparison analysis. For some participants it performed substantially better than the BCI model. It is in-principle possible that there is heterogeneity in the population, where different observers rely on different perceptual strategies. Furthermore, a combination of factors might jointly contribute to deviations from the MLE model. For example, we have discarded the correlated noise model as an explanation on its own (chapter 4), but it is plausible that some amount of supra-modal noise does contribute to the empirically measured variance across trials (Mueller & Weidemann, 2008; see also the discussion on metacognitive noise in chapter 5).

An attractive characteristic of the BCI explanation as we proposed it is its flexibility: while the Bayesian prior hypothesis would always predict overreliance on the dominant modality, the BCI-based proposal allows observers to rely more on other modalities if that would be favorable given the circumstances. In both chapter 3 and 4 we have observed a few participants who significantly overweighted audition rather than vision. Had we allowed the BCI model to individually decide whether it would select auditory or visual estimates for its responses, it would surely have surpassed fitting performance of

the Bayesian prior model by a much greater extent. The question arises again (see discussion above): why would some participants have overweighted audition rather than vision?

3. Future research combining metacognition and multisensory integration

One of the initial aims for the metacognition study (chapter 5) was to investigate whether overweighting patterns could be explained by observers' confidence responses. We hypothesized that participants with reportedly higher confidence in either unisensory modality would rely more on that modality during multisensory integration, i.e. their weights for that modality would be increased relative to MLE predictions. However, discussion of this research question was omitted from chapter 5 because, as it turned out, this particular data set proved to be the one in which deviations from MLE predictions were least pronounced (although visual overweighting was still shown to be significant in a 3 (reliability levels) by 2 (empirical vs. MLE-predicted) repeated-measures ANOVA on the auditory weights: F(1) = 5.2, p = 0.025, BF_{01} = 0.54; for statistical methods please see chapter 5). The data that was presented in chapter 5 thus may not have been ideal to test the hypothesized relationship between overconfidence and overweighting. Since overweighting was only modest, measurement noise may have obscured any existing correlation with overconfidence.

Future research that would want to address this open question might be better off using a slightly larger audiovisual disparity which would likely result in more overweighting and a better opportunity to test the hypothesized relationship with unisensory

metacognitive biases. Any such analysis with confidence ratings would then build upon the presumed fixed mapping between type-2 criteria and posterior probability of being correct (independent of sensory modalities, see Section 3.4 in chapter 5). An alternative would be to ask participants to express a confidence interval for each localization response (i.e. in units of visual angle around the location estimate), instead of confidence judgments for a two alternative forced-choice decision. A correlation analysis between sensory overweighting and unisensory confidence-interval size differences, as opposed to type-2 criteria differences, would be straightforward because of the theoretical direct relation between such confidence-interval sizes and sensory uncertainty (i.e. it avoids mapping 'posterior probabilities of being correct' onto N confidence levels).

While the reason why we observed only modest overweighting in chapter 5 relative to chapters 3 and 4 is not immediately clear, I speculate that this may be due to the experimental design in which we had asked participants to introspect on every response. Instead of defaulting to using partially-integrated visual location estimates, the explicit requirement for a confidence judgment may have prompted participants to evaluate both sensory modalities on a trial-by-trial basis. Comparison of their confidence for both auditory and visual partially-integrated estimates (i.e. $\widehat{\Delta S}_{A,BCI}$ and $\widehat{\Delta S}_{V,BCI}$) and selection of the most confident location response (e.g. if $\widehat{\Delta S}_{A,BCI}$ suggests 'left', but $\widehat{\Delta S}_{V,BCI}$ suggests 'right', then choose 'left' if $\left|\widehat{\Delta S}_{A,BCI}\right| > \left|\widehat{\Delta S}_{V,BCI}\right|$, 'right' otherwise) would have resulted in perceptual behavior that is very similar to MLE predictions. Such confidence-driven single-trial selection of the sensory modality after partial BCI integration would be an extreme example of the task-relevance based switching proposal that I have

discussed above (see 'outstanding questions and limitations'). Whether human observers do indeed have independent metacognitive access to both partially-integrated estimates remains to be investigated.

Another initial research objective that was omitted from chapter 5 because the experimental design proved unsuitable to address it, was to investigate whether uncertainty about the causal structure of the world (i.e. one common AV source or two independent A and V sources) would affect the reported confidence judgments of the type-1 perceptual decisions. Previously, White et al., (2014) have shown that integrated incongruent audiovisual speech stimuli that caused illusory McGurk percepts were associated with lower confidence levels than audiovisual congruent percepts of the same syllable. This result suggests that observers are able to evaluate their causal uncertainty through metacognitive processes and use it to modify their confidence reports on perceptual decisions (Deroy, Spence & Noppeney, 2016). For the metacognition study that is described in chapter 5, we had hypothesized that increased levels of causal uncertainty would lead to lower confidence judgments for AV spatially incongruent as opposed to spatially congruent conditions (i.e. an increase of the fitted type-2 criteria). The fact that we did not observe such a confidence difference should not be considered as proof that it would not have arisen had we used larger audiovisual disparities (i.e. if we had induced more causal uncertainty). To properly examine this interesting research direction in future experiments it would be advantageous to adopt an experimental design with various audiovisual disparity levels, as is conventional in BCI studies (Körding, Beierholm, Ma et al., 2007; Rohe & Noppeney, 2015a). Moreover, it would be helpful to ask participants for a direct indication of causal uncertainty (e.g. as a

confidence judgment on multisensory congruency decisions), so that these self-reported causal uncertainty estimates can be compared against perceptual confidence reports on a trial-by-trial basis.

4. Concluding remarks

The above-described discussions and suggested directions for future research build directly on the two central topics of this thesis: 1) that an experimenter doing multisensory research cannot assume observers to automatically and completely fuse two sensory signals because 2) such observers are engaged in a complex perceptual process wherein multiple sources of uncertainty interact and necessarily need to be evaluated in order to make successful decisions. A striking example of how the empirical work and theoretical background that was presented in this thesis contributes to a better understanding of the intricate relationship between multisensory integration and perceptual decisions comes from my own experience. In my first year of this PhD I observed that participants who displayed a major cross-modal bias in a standard ventriloquist effect paradigm with two locations on opposite sides of the midline (i.e. they nearly always located the sound on the same side as the flash), were almost one hundred percent correct when the same stimuli were presented in a two-interval forced-choice task wherein they were required to identify which of the two intervals contained the spatially incongruent stimuli (the other interval contained identical but spatially congruent stimuli). As a naïve experimenter I did not yet understand that observers could exhibit partial multisensory integration that may appear as forced fusion in insensitive experimental designs, and that these same observers would be able

to use metacognitive assessments of the multisensory uncertainty to successfully decide which of the two auditory location estimates was based on a perceptual illusion.

I think it is reasonable to conclude that I learned a lot about the intricate interplay of multisensory integration, metacognition, and sensitive psychophysical methodology through the various research projects during my PhD as presented here in this thesis. I hope that this body of work enables other researchers to go through this process faster such that we can build forward together with the aim to better understand human multisensory perception and decision making.

REFERENCES

Acerbi, L., Dokka, K., Angelaki, D. E., & Ma, W. J. (2018). Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *Plos Computational Biology, 14*(7). doi:ARTN 100611010.1371/journal.pcbi.1006110

Acerbi, L. & Ma, W. J. (2017). Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. *Advances in Neural Information Processing Systems 30*, pages 1834-1844.

Acuna, D. E., Berniker, M., Fernandes, H. L., & Kording, K. P. (2015). Using psychophysics to ask if the brain samples or maximizes. *Journal of Vision, 15*(3). doi:10.1167/15.3.7

Adler, W. T., & Ma, W. J. (2018). Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS Computional Biology, 14*(11), e1006572. doi:10.1371/journal.pcbi.1006572

Adler, W. T., & Ma, W. J. (2018b). Limitations of Proposed Signatures of Bayesian Confidence. *Neural Computation 30*(12), 1-28. doi:10.1162/neco_a_01141

Adam, R., & Noppeney, U. (2010). Prior auditory information shapes visual category-selectivity in ventral occipito-temporal cortex. *Neuroimage, 52*(4), 1592-1602. doi:10.1016/j.neuroimage.2010.05.002

Ais, J., Zylberberg, A., Barttfeld, P., & Sigman, M. (2016). Individual consistency in the accuracy and distribution of confidence judgments. *Cognition, 146*, 377-386. doi:10.1016/j.cognition.2015.10.006

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making. *PLoS Computational Biology, 11*(10), e1004519. doi:10.1371/journal.pcbi.1004519

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*(3), 257-262. doi:10.1016/j.cub.2004.01.029

Aller M, Giani A, Conrad V, Watanabe M, Noppeney U. (2015). A spatially collocated sound thrusts a flash into awareness. *Frontiers in Integrative Neuroscience, 9*(16). Doi:10.3389/fnint.2015.00016

Bang, J. W., Shekhar, M., & Rahnev, D. (2019). Sensory noise increases metacognitive efficiency. *Journal of Experimental Psychology General, 148*(3), 437-452. doi:10.1037/xge0000511

Barlow, H. B. (1956). Retinal noise and absolute threshold. *Journal of the Optical Society of America, 46*(8), 634-639.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A, Optics, Image Science and Vision, 20*(7), 1391-1397.

Battaglia, P. W., Kersten, D., & Schrater, P. R. (2011). How haptic size sensations improve distance perception. *PLoS Computational Biology, 7*(6), e1002080. doi:10.1371/journal.pcbi.1002080

Beauchamp, M. S., Pasalar, S., & Ro, T. (2010). Neural substrates of reliability-weighted visual-tactile multisensory integration. *Frontiers in Systems Neuroscience, 4*, 25. doi:10.3389/fnsys.2010.00025

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron, 74*(1), 30-39. doi:10.1016/j.neuron.2012.03.016

Beierholm, U. R., Quartz, S. R., & Shams, L. (2009). Bayesian priors are encoded independently from likelihoods in human multisensory perception. *Journal of Vision, 9*(5), 23 21-29. doi:10.1167/9.5.23

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: insights from audio-visual speech perception. *PLoS One, 6*(5), e19812. doi:10.1371/journal.pone.0019812

Bentvelzen, A., Leung, J., & Alais, D. (2009). Discriminating audiovisual speed: optimal integration of speed defaults to probability summation when component reliabilities diverge. *Perception, 38*(7), 966-987. doi:10.1068/p6261

Bishop, C. W., & Miller, L. M. (2011). Speech cues contribute to audiovisual spatial integration. *PLoS One, 6*(8), e24016. doi:10.1371/journal.pone.0024016

Bosen, A. K., Fleming, J. T., Brown, S. E., Allen, P. D., O'Neill, W. E., & Paige, G. D. (2016). Comparison of congruence judgment and auditory localization tasks for assessing the spatial limits of visual capture. *Biological Cybernetics, 110*(6), 455-471. doi:10.1007/s00422-016-0706-6

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

Bresciani, J. P., Dammeier, F., & Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. *Journal of Vision, 6*(5), 554-564. doi:10.1167/6.5.2

Bresciani, J. P., Ernst, M. O., Drewing, K., Bouyer, G., Maury, V., & Kheddar, A. (2005). Feeling what you hear: auditory signals can modulate tactile tap perception. *Experimental Brain Research, 162*(2), 172-180. doi:10.1007/s00221-004-2128-2

Burnham, K.P. and Anderson, D.R. (2002) Chapter 2, Section 2.4 In: *Model Selection and Inference: A Practical Information-Theoretic Approach* (pp. 66-67). 2nd Edition, Springer-Verlag, New York. http://dx.doi.org/10.1007/b97636

Burr, D., Banks, M. S., & Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research, 198*(1), 49-57. doi:10.1007/s00221-009-1933-z

Butler, J. S., Smith, S. T., Campos, J. L., & Bulthoff, H. H. (2010). Bayesian integration of visual and vestibular signals for heading. *Journal of Vision, 10*(11), 23. doi:10.1167/10.11.23

Cecere, R., Rees, G., & Romei, V. (2015). Individual differences in alpha frequency drive crossmodal illusory perception. *Current Biology, 25*(2), 231-235. doi:10.1016/j.cub.2014.11.034

Chen, Y. C., & Spence, C. (2017). Assessing the Role of the 'Unity Assumption' on Multisensory Integration: A Review. *Frontiers in Psychology, 8*, 445. doi:10.3389/fpsyg.2017.00445

Cuppini, C., Shams, L., Magosso, E., & Ursino, M. (2017). A biologically inspired neurocomputational model for audiovisual integration and causal inference. *European Journal of Neuroscience, 46*(9), 2481-2498. doi:10.1111/ejn.13725

de Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a Common Currency between Vision and Audition. *PLoS One, 11*(1), e0147901. doi:10.1371/journal.pone.0147901

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience, 16*(1), 105-110. doi:10.1038/nn.3279

de Winkel, K. N., Katliar, M., & Bulthoff, H. H. (2017). Causal Inference in Multisensory Heading Estimation. *PLoS One, 12*(1), e0169676. doi:10.1371/journal.pone.0169676

Denison, R. N., Adler, W. T., Carrasco, M., & Ma, W. J. (2018). Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences of the U.S.A., 115*(43), 11090-11095. doi:10.1073/pnas.1717720115

Deroy, O., Spence, C., & Noppeney, U. (2016). Metacognition in Multisensory Perception. *Trends in Cognitive Sciences, 20*(10), 736-747. doi:10.1016/j.tics.2016.08.006

Di Luca, M., & Rhodes, D. (2016). Optimal Perceived Timing: Integrating Sensory Information with Dynamically Updated Expectations. *Scientific Reports, 6*, 28563. doi:10.1038/srep28563

Drugowitsch, J., Moreno-Bote, R., & Pouget, A. (2014). Relation between belief and performance in perceptual decision making. *PLoS One, 9*(5), e96511. doi:10.1371/journal.pone.0096511

Dyjas, O., Bausenhart, K. M., & Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. *Attention, Perception, & Psychophysics, 74*(8), 1819–1841. https://doi.org/10.3758/s13414-012-0362-4

Erev, I., & Roth, A. E. (2014). Maximization, learning, and economic behavior. *Proceedings of the National Academy of Sciences of the U.S.A., 111 Suppl 3*, 10818-10825. doi:10.1073/pnas.1402846111

Ernst, M. O. (2006). Chapter 6 - A Bayesian view on multimodal cue integration. In *Human Body Perception From The Inside Out* (pp. 105-131). New York: Oxford University Press.

Ernst, M. O. (2012). Optimal Multisensory Integration: Assumptions and Limits. In B. E. Stein (Ed.), *The New Handbook of Multisensory Processes* (pp. 527-544). Cambridge, Massachusetts: MIT Press.

Ernst, M. O. & Banks, M. S. (2000). What determines dominance of vision over haptics? *Proceedings of the Annual Psychonomics Meeting*.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429-433. doi:10.1038/415429a

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*(4), 162–169. https://doi.org/10.1016/j.tics.2004.02.002

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews. Neuroscience, 9*(4), 292-303. doi:10.1038/nrn2258

Faivre, N., Filevich, E., Solovey, G., Kuhn, S., & Blanke, O. (2018). Behavioral, Modeling, and Electrophysiological Evidence for Supramodality in Human Metacognition. *Journal of Neuroscience, 38*(2), 263-277. doi:10.1523/JNEUROSCI.0322-17.2017

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fetsch, C. R., Kiani, R., Newsome, W. T., & Shadlen, M. N. (2014). Effects of Cortical Microstimulation on Confidence in a Perceptual Decision. *Neuron, 84*(1), 239. doi:10.1016/j.neuron.2014.09.020

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2011). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience, 15*(1), 146-154. doi:10.1038/nn.2983

Fetsch, C. R., Turner, A. H., DeAngelis, G. C., & Angelaki, D. E. (2009). Dynamic reweighting of visual and vestibular cues during self-motion perception. *Journal of Neuroscience, 29*(49), 15601-15612. doi:10.1523/JNEUROSCI.2574-09.2009

Fleming, S. M. (2016). Changing our minds about changes of mind. *Elife, 5*, e14790. doi:10.7554/eLife.14790

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Hum Neurosci, 8*, 443. doi:10.3389/fnhum.2014.00443

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-specific disruption of perceptual confidence. *Psychological Science, 26*(1), 89-98. doi:10.1177/0956797614557697

Fründ I, Haenel NV, Wichmann FA. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision, 11*(6).

Gardner, W. G., & Martin, K. D. (1995). HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America, 97*(6), 3907–3908. https://doi.org/10.1121/1.412407

Gau, R., & Noppeney, U. (2016). How prior expectations shape multisensory perception. *Neuroimage, 124*(Pt A), 876-886. doi:10.1016/j.neuroimage.2015.09.045

Geisler, W. S., & Diehl, R. L. (2002). Bayesian natural selection and the evolution of perceptual systems. *Philosophical Transactions of the Royal Society of London, B. Biological Sciences, 357*(1420), 419-448. doi:10.1098/rstb.2001.1055

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Basics of Markov chain simulation (Chapter 11, Sections 11.4-11.5). In: *Bayesian data analysis* (3rd edition). CRC Press; 2013.

Gepshtein, S., & Banks, M. S. (2003). Viewing geometry determines how vision and haptics combine in size perception. *Current Biology, 13*(6), 483-488.

Gepshtein, S., Burge, J., Ernst, M. O., & Banks, M. S. (2005). The combination of vision and touch depends on spatial proximity. *Journal of Vision, 5*(11), 1013-1023. doi:10.1167/5.11.7

Gibaldi, A., Vanegas, M., Bex, P. J., & Maiello, G. (2017). Evaluation of the Tobii EyeX Eye tracking controller and Matlab toolkit for research. *Behavior Research Methods, 49*(3), 923-946. doi:10.3758/s13428-016-0762-9

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience, 14*(7), 926-932. doi:10.1038/nn.2831

Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development. *Developmental Sciences, 10*(3), 281-287. doi:10.1111/j.1467-7687.2007.00584.x

Gori, M., Sciutti, A., Burr, D., & Sandini, G. (2011). Direct and indirect haptic calibration of visual size judgments. *PLoS One, 6*(10), e25599. doi:10.1371/journal.pone.0025599

Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transaction of the Royal Society of London, B. Biological Sciences, 352*(1358), 1121-1127. doi:10.1098/rstb.1997.0095

Gu, Y., Angelaki, D. E., & Deangelis, G. C. (2008). Neural correlates of multisensory cue integration in macaque MSTd. *Nature Neuroscience, 11*(10), 1201-1210. doi:10.1038/nn.2191

Hangya, B., Sanders, J. I., & Kepecs, A. (2016). A Mathematical Framework for Statistical Decision Confidence. *Neural Computation, 28*(9), 1840-1858. doi:10.1162/NECO_a_00864

Hairston, W. D., Wallace, M. T., Vaughan, J. W., Stein, B. E., Norris, J. L., & Schirillo, J. A. (2003). Visual localization ability influences cross-modal bias. *Journal of Cognitive Neuroscience, 15*(1), 20-29. doi:10.1162/089892903321107792

Hartcher-O'Brien, J., Di Luca, M., & Ernst, M. O. (2014). The duration of uncertain times: audiovisual information about intervals is integrated in a statistically optimal fashion. *PLoS One, 9*(3), e89339. doi:10.1371/journal.pone.0089339

Helbig, H. B., & Ernst, M. O. (2007). Optimal integration of shape information from vision and touch. *Experimental Brain Research, 179*(4), 595-606. doi:10.1007/s00221-006-0814-y

Helbig, H. B., Ernst, M. O., Ricciardi, E., Pietrini, P., Thielscher, A., Mayer, K. M., . . . Noppeney, U. (2012). The neural mechanisms of reliability weighted integration of shape information from vision and touch. *Neuroimage, 60*(2), 1063-1072. doi:10.1016/j.neuroimage.2011.09.072

Honkanen, A., Immonen, E. V., Salmela, I., Heimonen, K., & Weckstrom, M. (2017). Insect photoreceptor adaptations to night vision. *Philosophical Transaction of the Royal Society of London, B. Biological Sciences, 372*(1717). doi:10.1098/rstb.2016.0077

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research, 39*(21), 3621-3629.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences, 6*(8), 345-350.

Jacobs, R. A., & Fine, I. (1999). Experience-dependent integration of texture and motion cues to depth. *Vision Research, 39*(24), 4062-4075.

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review, 119*(1), 186-200. doi:10.1037/a0025960

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience, 13*(8), 1020-1026. doi:10.1038/nn.2590

Kanaya, S., & Yokosawa, K. (2011). Perceptual congruency of audio-visual speech affects ventriloquism with bilateral visual stimuli. *Psychonomic Bulletin & Review, 18*(1), 123-128. doi:10.3758/s13423-010-0027-z

Kerzel, D. (2002). Memory for the position of stationary objects: disentangling foveal bias and memory averaging. *Vision Research, 42*(2), 159-167.

Kingdom, F. A. A., & Prins, N. (2016). *Psychophysics* (Second Edition). San Diego: Academic Press. http://www.sciencedirect.com/science/book/9780124071568

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron, 84*(6), 1329-1342. doi:10.1016/j.neuron.2014.12.015

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: a commentary. *Perception & Psychophysics, 63*(8), 1421-1455.
Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C., & others. (2007). What's new in Psychtoolbox-3. *Perception, 36*(14), 1.
Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS One, 2*(9), e943. doi:10.1371/journal.pone.0000943
Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427*(6971), 244-247. doi:10.1038/nature02169
Lee, H., & Noppeney, U. (2011). Long-term music training tunes how the brain temporally binds signals from multiple senses. *Proceedings of the National Academy of Sciences of the U.S.A., 108*(51), E1441-1450. doi:10.1073/pnas.1115267108
Lee, H., & Noppeney, U. (2014). Temporal prediction errors in visual and auditory cortices. *Current Biology, 24*(8), R309-310. doi:10.1016/j.cub.2014.02.007
Lewald, J., & Guski, R. (2003). Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Brain Research. Cognitive Brain Research, 16*(3), 468-478.
Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9*(11), 1432-1438. doi:10.1038/nn1790
Ma, W. J., & Rahmati, M. (2013). Towards a neural implementation of causal inference in cue combination. *Multisensory Research, 26*(1-2), 159-176.
Magnotti, J. F., & Beauchamp, M. S. (2017). A Causal Inference Model Explains Perception of the McGurk Effect and Other Incongruent Audiovisual Speech. *PLoS Computational Biology, 13*(2), e1005229. doi:10.1371/journal.pcbi.1005229
Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. *Frontiers in Psychology, 4*, 798. doi:10.3389/fpsyg.2013.00798
Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of Experimental Psychology. Human Perception and Performance, 37*(1), 245-256. doi:10.1037/a0019952
Maiworm, M., & Röder, B. (2011). Suboptimal auditory dominance in audiovisual integration of temporalcues. *Tsinghua Science and Technology, 16*(2), 121-132. doi:10.1016/S1007-0214(11)70019-0
Mamassian, P., & Landy, M. S. (2001). Interaction of visual prior constraints. *Vision Research, 41*(20), 2653-2668.
Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*(1), 422-430. doi:10.1016/j.concog.2011.09.021
Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: Meta-d', response-specific meta-d', and the unequal variance SDT model. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25-66). New York, NY, US: Springer-Verlag Publishing. http://dx.doi.org/10.1007/978-3-642-45190-4_3

Maniscalco, B., & Lau, H. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness, 2016*(1). doi:10.1093/nc/niw002

Maniscalco, B., Peters, M. A., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception & Psychophysics, 78*(3), 923-937. doi:10.3758/s13414-016-1059-x

McGovern, D. P., Roudaia, E., Newell, F. N., & Roach, N. W. (2016). Perceptual learning shapes multisensory causal inference via two distinct mechanisms. *Scientific Reports, 6*, 24673. doi:10.1038/srep24673

Meijer, D., Veselic, S., Calafiore, C., & Noppeney, U. (2019). Integration of audiovisual spatial signals is not consistent with maximum likelihood estimation. *Cortex (in press)*. Doi: 10.1016/j.cortex.2019.03.026 [See chapter 3 of this thesis]

Mendonca, C., Mandelli, P., & Pulkki, V. (2016). Modeling the Perception of Audiovisual Distance: Bayesian Causal Inference and Other Models. *PLoS One, 11*(12), e0165391. doi:10.1371/journal.pone.0165391

Mendonca, C., Santos, J. A., & Lopez-Moliner, J. (2011). The benefit of multisensory integration with biological motion signals. *Experimental Brain Research, 213*(2-3), 185-192. doi:10.1007/s00221-011-2620-4

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *Journal of Neuroscience, 7*(10), 3215-3229.

Meredith, M. A., & Stein, B. E. (1996). Spatial determinants of multisensory integration in cat superior colliculus neurons. *Journal of Neurophysiology, 75*(5), 1843-1857. doi:10.1152/jn.1996.75.5.1843

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*(4), 406–419. https://doi.org/10.1037/a0024377

Morgan, M. L., Deangelis, G. C., & Angelaki, D. E. (2008). Multisensory integration in macaque visual cortex depends on cue reliability. *Neuron, 59*(4), 662-673. doi:10.1016/j.neuron.2008.06.024

Mueller, S. T., & Weidemann, C.T. (2008). Decision noise: an explanation for observed violations of signals detection theory. *Psychonomic Bulletin & Review, 15*(3), 465-494.

Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015). Neural evidence accumulation persists after choice to inform metacognitive judgments. *Elife, 4*. doi:10.7554/eLife.11946

Nahorna, O., Berthommier, F., & Schwartz, J. L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *Journal of the Acoustic Society of America, 132*(2), 1061-1077. doi:10.1121/1.4728187

Nahorna, O., Berthommier, F., & Schwartz, J. L. (2015). Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the McGurk effect. *Journal of the Acoustic Society of America, 137*(1), 362-377. doi:10.1121/1.4904536

Natarajan, R., Murray, I., Shams, L., & Zemel, R. S. (2009). Characterizing response behavior in multisensory perception with conflicting cues. In D. Koller, D.

Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21* (pp. 1153-1160): MIT Press.
Nath, A. R., & Beauchamp, M. S. (2011). Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *Journal of Neuroscience, 31*(5), 1704-1714. doi:10.1523/JNEUROSCI.4853-10.2011
Noppeney, U., Josephs, O., Hocking, J., Price, C. J., & Friston, K. J. (2008). The effect of prior visual information on recognition of speech and sounds. *Cerebral Cortex, 18*(3), 598-609. doi:10.1093/cercor/bhm091
Odegaard, B., & Shams, L. (2016). The Brain's Tendency to Bind Audiovisual Signals Is Stable but Not General. *Psychological Sciences, 27*(4), 583-591. doi:10.1177/0956797616628860
Odegaard, B., Wozny, D. R., & Shams, L. (2015). Biases in Visual, Auditory, and Audiovisual Perception of Space. *PLoS Computational Biology, 11*(12), e1004649. doi:10.1371/journal.pcbi.1004649
Odegaard, B., Wozny, D. R., & Shams, L. (2017). A simple and efficient method to enhance audiovisual binding tendencies. *PeerJ, 5*, e3143. doi:10.7717/peerj.3143
Ohshiro, T., Angelaki, D. E., & DeAngelis, G. C. (2011). A normalization model of multisensory integration. *Nature Neuroscience, 14*(6), 775-782. doi:10.1038/nn.2815
Ohshiro, T., Angelaki, D. E., & DeAngelis, G. C. (2017). A Neural Signature of Divisive Normalization at the Level of Multisensory Integration in Primate Cortex. *Neuron, 95*(2), 399-411 e398. doi:10.1016/j.neuron.2017.06.043
Oruc, I., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research, 43*(23), 2451-2468.
Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. *Nature Communications, 7*, 11543. doi:10.1038/ncomms11543
Parise, C. V., & Spence, C. (2009). 'When birds of a feather flock together': synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS One, 4*(5), e5664. doi:10.1371/journal.pone.0005664
Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Current Biology, 22*(1), 46-49. doi:10.1016/j.cub.2011.11.039
Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience, 16*(9), 1170-1178. doi:10.1038/nn.3495
Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature Neuroscience, 19*(3), 366-374. doi:10.1038/nn.4240
Prsa, M., Gale, S., & Blanke, O. (2012). Self-motion leads to mandatory cue fusion across sensory modalities. *Journal of Neurophysiology, 108*(8), 2282-2291. doi:10.1152/jn.00439.2012
Rahnev, D. (2018) The case against full probability distributions in perceptual decision making. *bioRxiv* 108944; https://doi.org/10.1101/108944
Rahnev D, Denison RN. Suboptimality in Perceptual Decision Making. *The Behavioral and Brain Sciences.* 2018:1-107.
Rahnev, D., Nee, D. E., Riddle, J., Larson, A. S., & D'Esposito, M. (2016). Causal evidence for frontal cortex organization for perceptual decision making. *Proceedings of the*

National Academy of Sciences of the U.S.A., 113(21), 6059-6064. doi:10.1073/pnas.1522551113

Raposo, D., Sheppard, J. P., Schrater, P. R., & Churchland, A. K. (2012). Multisensory decision-making in rats and humans. *Journal of Neuroscience, 32*(11), 3726-3735. doi:10.1523/JNEUROSCI.4998-11.2012

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - revisited. *Neuroimage, 84*, 971-985. doi:10.1016/j.neuroimage.2013.08.065

Roach, N. W., Heron, J., & McGraw, P. V. (2006). Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proceedings. Biological Sciences, 273*(1598), 2159-2168. doi:10.1098/rspb.2006.3578

Roach, N. W., McGraw, P. V., Whitaker, D. J., & Heron, J. (2017). Generalization of prior information for rapid Bayesian time estimation. *Proceedings of the National Academy of Sciences of the U.S.A., 114*(2), 412-417. doi:10.1073/pnas.1610706114

Rohde, M., van Dam, L. C. J., & Ernst, M. O. (2016). Statistically Optimal Multisensory Cue Integration: A Practical Tutorial. *Multisensory Research, 29*(4-5), 279-317. doi:https://doi.org/10.1163/22134808-00002510

Rohe, T., & Noppeney, U. (2015a). Sensory reliability shapes perceptual inference via two mechanisms. *Journal of Vision, 15*(5), 22. doi:10.1167/15.5.22

Rohe, T., & Noppeney, U. (2015b). Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLoS Biology, 13*(2), e1002073. doi:10.1371/journal.pbio.1002073

Rohe, T., & Noppeney, U. (2016). Distinct Computational Principles Govern Multisensory Integration in Primary Sensory and Association Cortices. *Current Biology, 26*(4), 509-514. doi:10.1016/j.cub.2015.12.056

Rosas, P., Wagemans, J., Ernst, M. O., & Wichmann, F. A. (2005). Texture and haptic cues in slant discrimination: reliability-based cue weighting without statistically optimal cue combination. *Journal of the Optical Society of America, A. Optics, Image Science and Vision, 22*(5), 801-809.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356-374. doi:10.1016/j.jmp.2012.08.001

Samaha, J., & Postle, B. R. (2015). The Speed of Alpha-Band Oscillations Predicts the Temporal Resolution of Visual Perception. *Current Biology, 25*(22), 2985-2990. doi:10.1016/j.cub.2015.10.007

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a Statistical Computation in the Human Sense of Confidence. *Neuron, 90*(3), 499-506. doi:10.1016/j.neuron.2016.03.025

Sato, Y., Toyoizumi, T., & Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. *Neural Computation, 19*(12), 3335-3355. doi:10.1162/neco.2007.19.12.3335

Schütt HH, Harmeling S, Macke JH, Wichmann FA. Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research. 2016*;122:105-123.

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences, 14*(9), 425-432. doi:10.1016/j.tics.2010.07.001

Shams, L., Ma, W. J., & Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport, 16*(17), 1923-1927.

Shen, S., & Ma, W. J. (2016). A detailed comparison of optimality and simplicity in perceptual decision making. *Psychological Review, 123*(4), 452-480. doi:10.1037/rev0000028

Sheppard, J. P., Raposo, D., & Churchland, A. K. (2013). Dynamic weighting of multisensory stimuli shapes decision-making in rats and humans. *Journal of Vision, 13*(6). doi:10.1167/13.6.4

Siedlecka, M., Paulewicz, B., & Wierzchon, M. (2016). But I Was So Sure! Metacognitive Judgments Are Less Accurate Given Prospectively than Retrospectively. *Frontiers in Psychology, 7*, 218. doi:10.3389/fpsyg.2016.00218

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport, 12*(1), 7-10.

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition, 20*(4), 1787-1792. doi:10.1016/j.concog.2010.12.011

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of Experimental Psychology. Human Perception and Performance, 35*(2), 580-587. doi:10.1037/a0013483

Spratling, M. W. (2016). A neural implementation of Bayesian inference based on predictive coding. *Connection Science, 28*(4), 346-383. doi:10.1080/09540091.2016.1243655

Stevens, C. F. (2003). Neurotransmitter release at central synapses. *Neuron, 40*(2), 381-388.

Stevenson, R. A., Fister, J. K., Barnett, Z. P., Nidiffer, A. R., & Wallace, M. T. (2012). Interactions between the spatial and temporal stimulus factors that influence multisensory integration in human performance. *Experimental Brain Research, 219*(1), 121-137. doi:10.1007/s00221-012-3072-1

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9*(4), 578-585. doi:10.1038/nn1669

Talsma D, Senkowski D, Soto-Faraco S, Woldorff MG. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14*(9), 400-410.

Thakur, B., Mukherjee, A., Sen, A., & Banerjee, A. (2016). A dynamical framework to relate perceptual variability with multisensory information processing. *Scientific Reports, 6*, 31280. doi:10.1038/srep31280

Triesch, J., Ballard, D. H., & Jacobs, R. A. (2002). Fast temporal dynamics of visual cue integration. *Perception, 31*(4), 421-434. doi:10.1068/p3314

van Beers, R. J., Sittig, A. C., & Denier van der Gon, J. J. (1996). How humans combine simultaneous proprioceptive and visual position information. *Experimental Brain Research, 111*(2), 253-261.
van Bergen, R. S., Ma, W. J., Pratte, M. S., & Jehee, J. F. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience, 18*(12), 1728-1730. doi:10.1038/nn.4150
van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *Elife, 5*, e12192. doi:10.7554/eLife.12192
van den Berg, R., Yoo, A. H., & Ma, W. J. (2017). Fechner's law in metacognition: A quantitative model of visual working memory confidence. *Psychological Review, 124*(2), 197-214. doi:10.1037/rev0000060
van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*(3), 598-607. doi:10.1016/j.neuropsychologia.2006.01.001
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC (vol 27, pg 1413, 2017). *Statistics and Computing, 27*(5), 1433-1433. doi:10.1007/s11222-016-9709-3
Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science, 38*(4), 599-637. doi:10.1111/cogs.12101
Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research, 158*(2), 252-258. doi:10.1007/s00221-004-1899-9
Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: implications for transitivity among the spatial senses. *Perception & Psychophysics, 30*(6), 557-564.
Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience, 5*(6), 598-604. doi:10.1038/nn858
White, T. P., Wigton, R. L., Joyce, D. W., Bobin, T., Ferragamo, C., Wasim, N., Lisk, S., & Shergill, S. S. (2014). Eluding the illusion? Schizophrenia, dopamine and the McGurk effect. *Frontiers in Human Neuroscience, 8*, 565. doi:10.3389/fnhum.2014.00565
Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Attention, Perception, & Psychophysics, 63*(8), 1293–1313.
Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Attention, Perception, & Psychophysics, 63*(8), 1314–1329.
Wokke, M. E., Cleeremans, A., & Ridderinkhof, K. R. (2017). Sure I'm Sure: Prefrontal Oscillations Support Metacognitive Monitoring of Decision Making. *Journal of Neuroscience, 37*(4), 781-789. doi:10.1523/JNEUROSCI.1612-16.2016
Wozny, D. R., Beierholm, U. R., & Shams, L. (2008). Human trimodal perception follows optimal statistical inference. *Journal of Vision, 8*(3), 24 21-11. doi:10.1167/8.3.24
Wozny, D. R., Beierholm, U. R., & Shams, L. (2010). Probability matching as a computational strategy used in perception. *PLoS Computational Biology, 6*(8). doi:10.1371/journal.pcbi.1000871

Wozny, D. R., & Shams, L. (2011). Recalibration of auditory space following milliseconds of cross-modal discrepancy. *Journal of Neuroscience, 31*(12), 4607-4612. doi:10.1523/JNEUROSCI.6079-10.2011

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society of London, B. Biological Sciences, 367*(1594), 1310-1321. doi:10.1098/rstb.2011.0416

Young, M. J., Landy, M. S., & Maloney, L. T. (1993). A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision Research, 33*(18), 2685-2696.

Yu, Z., Chen, F., Dong, J., & Dai, Q. (2016). Sampling-based causal inference in cue combination and its neural implementation. *Neurocomputing, 175*, 155-165. doi:https://doi.org/10.1016/j.neucom.2015.10.045

Yuille, A. L., & Bulthoff, H. H. (1996). Bayesian decision theory and psychophysics. In C. K. David & R. Whitman (Eds.), *Perception as Bayesian inference* (pp. 123-161): Cambridge University Press.