

Optimal multisensory decision-making in a reaction-time task

Jan Drugowitsch^{1,2,4*}, Gregory C DeAngelis^{1†}, Eliana M Klier³, Dora E Angelaki^{3†}, Alexandre Pouget^{1,4†}

¹Department of Brain and Cognitive Sciences, University of Rochester, New York, United States; ²Institut National de la Santé et de la Recherche Médicale, École Normale Supérieure, Paris, France; ³Department of Neuroscience, Baylor College of Medicine, Houston, United States; ⁴Département des Neurosciences Fondamentales, Université de Genève, Geneva, Switzerland

Abstract Humans and animals can integrate sensory evidence from various sources to make decisions in a statistically near-optimal manner, provided that the stimulus presentation time is fixed across trials. Little is known about whether optimality is preserved when subjects can choose when to make a decision (reaction-time task), nor when sensory inputs have time-varying reliability. Using a reaction-time version of a visual/vestibular heading discrimination task, we show that behavior is clearly sub-optimal when quantified with traditional optimality metrics that ignore reaction times. We created a computational model that accumulates evidence optimally across both cues and time, and trades off accuracy with decision speed. This model quantitatively explains subjects' choices and reaction times, supporting the hypothesis that subjects do, in fact, accumulate evidence optimally over time and across sensory modalities, even when the reaction time is under the subject's control. DOI: [10.7554/eLife.03005.001](https://doi.org/10.7554/eLife.03005.001)

Introduction

Effective decision making in an uncertain, rapidly changing environment requires optimal use of all information available to the decision-maker. Numerous previous studies have examined how integrating multiple sensory cues—either within or across sensory modalities—alters perceptual sensitivity (*van Beers et al., 1996; Ernst and Banks, 2002; Battaglia et al., 2003; Fetsch et al., 2009*). These studies generally reveal that subjects' ability to discriminate among stimuli improves when multiple sensory cues are available, such as visual and tactile (*van Beers et al., 1996; Ernst and Banks, 2002*), visual and auditory (*Battaglia et al., 2003*), or visual and vestibular (*Fetsch et al., 2009*) cues. The performance gains associated with cue integration are generally well predicted by models that combine information across senses in a statistically optimal manner (*Clark and Yuille, 1990*). Specifically, we consider cue integration to be optimal if the information in the combined, multisensory condition is the sum of that available from the separate cues (see *Supplementary file 1* for formal statement) (*Clark and Yuille, 1990*).

Previous studies and models share a common fundamental limitation: they only consider situations in which the stimulus duration is fixed and subjects are required to withhold their response until the stimulus epoch expires. In natural settings, by contrast, subjects usually choose for themselves when they have gathered enough information to make a decision. In such contexts, it is possible that subjects integrate multiple cues to gain speed or to increase their proportion of correct responses (or some combination of effects), and it is unknown whether standard criteria for optimal cue integration apply. Indeed, using a reaction-time version of a multimodal heading discrimination task, we demonstrate here that human performance is markedly suboptimal when evaluated with standard criteria that ignore reaction times. Thus, the conventional framework for optimal cue integration is not applicable to behaviors in which decision times are under subjects' control.

*For correspondence: jdrugo@gmail.com

†These authors contributed equally to this work

Competing interests: See page 17


Funding: See page 17

Received: 04 April 2014

Accepted: 12 June 2014

Published: 14 June 2014

Reviewing editor: Eve Marder, Brandeis University, United States

 Copyright Drugowitsch et al. This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

eLife digest Imagine trying out a new roller-coaster ride and doing your best to figure out if you are being hurled to the left or to the right. You might think that this task would be easier if your eyes were open because you could rely on information from your eyes and also from the vestibular system in your ears. This is also what cue combination theory says—our ability to discriminate between two potential outcomes is enhanced when we can draw on more than one of the senses.

However, previous tests of cue combination theory have been limited in that test subjects have been asked to respond after receiving information for a fixed period of time whereas, in real life, we tend to make a decision as soon as we have gathered sufficient information. Now, using data collected from seven human subjects in a simulator, Drugowitsch et al. have confirmed that test subjects do indeed give more correct answers in more realistic conditions when they have two sources of information to rely on, rather than only one.

What makes this result surprising? Traditional cue combination theories do not consider that slower decisions allow us to process more information and therefore tend to be more accurate. Drugowitsch et al. show that this shortcoming causes such theories to conclude that multiple information sources might lead to worse decisions. For example, some of their test subjects made less accurate choices when they were presented with both visual and vestibular information, compared to when only visual information was available, because they made these choices very rapidly.

By developing a theory that takes into account both reaction times and choice accuracy, Drugowitsch et al. were able to show that, despite different trade-offs between speed and accuracy, test subjects still combined the information from their eyes and ears in a way that was close to ideal. As such the work offers a more thorough account of human decision making.

DOI: [10.7554/eLife.03005.002](https://doi.org/10.7554/eLife.03005.002)

On the other hand, there is a large body of empirical studies that has focused on how multisensory integration affects reaction times, but these studies have generally ignored effects on perceptual sensitivity (*Colonus and Arndt, 2001; Otto and Mamassian, 2012*). Some of these studies have reported that reaction times for multisensory stimuli are faster than predicted by 'parallel race' models (*Raab, 1962; Miller, 1982*), suggesting that multisensory inputs are combined into a common representation. However, other groups have failed to replicate these findings (*Corneil et al., 2002; Whitchurch and Takahashi, 2006*) and it is unclear whether the sensory inputs are combined optimally. Thus, multisensory integration in reaction time experiments remains poorly understood, and there is no coherent framework for evaluating optimal decision making that incorporates both perceptual sensitivity and reaction times. We address this substantial gap in knowledge both theoretically and experimentally.

For tasks based on information from a single sensory modality, diffusion models (DMs) have proven to be very effective at characterizing both the speed and accuracy of perceptual decisions, as well as speed/accuracy trade-offs (*Ratcliff, 1978; Ratcliff and Smith, 2004; Palmer et al., 2005*) (where accuracy is used in the sense of percentage of correct responses). Here, we develop a novel form of DM that not only integrates evidence optimally over time but also across different sensory cues, providing an optimal decision model for multisensory integration in a reaction-time context. The model is capable of combining cues optimally even when the reliability of each sensory input varies as a function of time. We show that this model reproduces human subjects' behavior very well, thus demonstrating that subjects near-optimally combine momentary evidence across sensory modalities. The model also predicts the counterintuitive finding that discrimination thresholds are often increased during cue combination, and demonstrates that this departure from standard cue-integration theory is due to a speed-accuracy tradeoff.

Overall, our findings provide a framework for extending cue-integration research to more natural contexts in which decision times are unconstrained and sensory cues vary substantially over time.

Results

We collected behavioral data from seven human subjects, A–G, performing a reaction-time version of a heading discrimination task (*Gu et al., 2007, 2008, 2010; Fetsch et al., 2009*) based on optic flow

alone (visual condition), inertial motion alone (vestibular condition), or a combination of both cues (combined condition, **Figure 1A**). In each stimulus condition, the subjects experienced forward translation with a small leftward or rightward deviation, and their task was to report whether they moved leftward or rightward relative to (an internal standard of) straight ahead (**Figure 1B**). In the combined condition, visual and vestibular cues were always spatially congruent, and followed temporally synchronized Gaussian velocity profiles (**Figure 1C**). Reliability of the visual cue was varied randomly across trials by changing the motion coherence of the optic flow stimulus (three coherence levels). For subjects B, D, and F, an additional experiment with six coherence levels was performed (denoted as B2, D2, F2). In contrast to previous tasks conducted with the same apparatus (**Fetsch et al., 2009; Gu et al., 2010**), subjects did not have to wait until the end of the stimulus presentation, but were allowed to respond at any time throughout the trial, which lasted up to 2 s.

For all conditions and all subjects, heading discrimination performance improved with an increase in heading direction away from straight ahead and with increased visual motion coherence. Let h denote the heading angle relative to straight ahead ($h > 0$ for right, $h < 0$ for left), and $|h|$ its magnitude. Larger values of $|h|$ simplified the discrimination task, as reflected by a larger fraction of correct choices (**Figure 2A** for subject D2, **Figure 3—figure supplement 1** for other subjects). To quantify discrimination performance, we fitted a cumulative Gaussian function to the psychometric curve for each stimulus condition and coherence. A lower discrimination threshold, given by the standard deviation of the fitted Gaussian, indicates a steeper psychometric curve and thus better performance. For both the visual and combined conditions, discrimination thresholds consistently decreased with an increase in motion coherence (**Figure 2B** for subject D2, **Figure 2—figure supplement 1** for other subjects), indicating that increasing coherence improves heading discrimination.

Sub-optimal cue combination?

Traditional cue combination models predict that the discrimination threshold in the combined condition should be smaller than that of either unimodal condition (**Clark and Yuille, 1990**). With a fixed stimulus duration, this prediction has been shown to hold for visual/vestibular heading discrimination

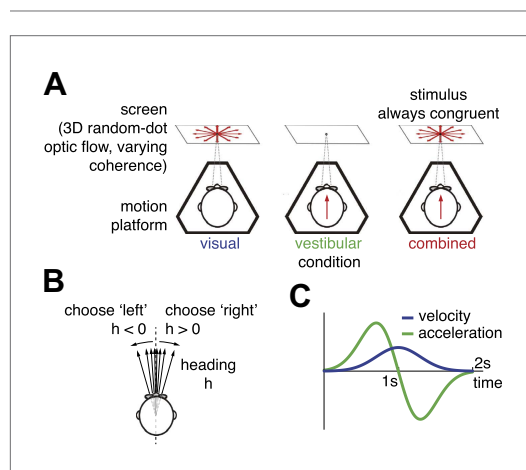


Figure 1. Heading discrimination task. **(A)** Subjects are seated on a motion platform in front of a screen displaying 3D optic flow. They perform a heading discrimination task based on optic flow (visual condition), platform motion (vestibular condition), or both cues in combination (combined condition). Coherence of the optic flow is constant within a trial but varies randomly across trials. **(B)** The subjects' task is to indicate whether they are moving rightward or leftward relative to straight ahead. Both motion direction (sign of h) and heading angle (magnitude of $|h|$) are chosen randomly between trials. **(C)** The velocity profile is Gaussian with peak velocity ~ 1 s after stimulus onset. DOI: [10.7554/eLife.03005.003](https://doi.org/10.7554/eLife.03005.003)

in both human and animal subjects (**Fetsch et al., 2009, 2011**), consistent with optimal cue combination. In contrast, the discrimination thresholds of subjects in our reaction-time task appear to be substantially sub-optimal. For the example subject of **Figure 2A**, psychometric functions in the combined condition lie between the visual and vestibular functions. Correspondingly, discrimination thresholds for the combined condition are intermediate between visual and vestibular thresholds for this subject, and for high coherences, are substantially greater than the optimal predictions (**Figure 2B**).

This pattern of results was consistent across subjects (**Figure 2C, Figure 2—figure supplement 1**). In no case did subjects feature a significantly lower discrimination threshold in the combined condition than the better of the two unimodal conditions ($p > 0.57$, one-tailed, **Supplementary file 2A**). For the largest visual motion coherence (70%), all subjects except one showed thresholds in the combined condition that were significantly greater than visual thresholds and significant greater than optimal predictions of a conventional cue-integration scheme ($p < 0.05$, **Supplementary file 2A**). These data lie in stark contrast to previous reports using fixed duration stimuli (**Fetsch et al., 2009, 2011**) in which combined thresholds were generally found to improve compared to

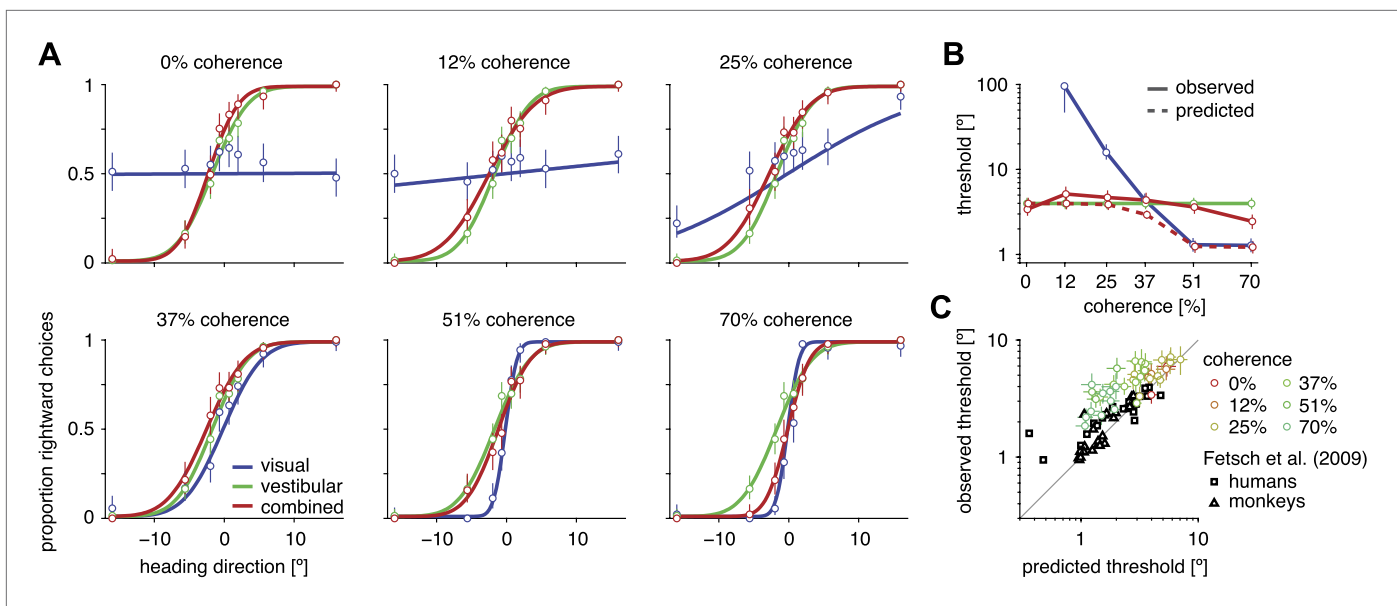


Figure 2. Heading discrimination performance. (A) Plots show the proportion of rightward choices for each heading and stimulus condition. Data are shown for subject D2, who was tested with 6 coherence levels. Error bars indicate 95% confidence intervals. (B) Discrimination threshold for each coherence and condition for subject D2 (see **Figure 2—figure supplement 1** for discrimination thresholds of all subjects). For large coherences, the threshold in the combined condition (solid red curve) lies between that of the vestibular and visual conditions, a marked deviation from the standard prediction (dashed red curve) of optimal cue integration theory. (C) Observed vs predicted discrimination thresholds for the combined condition for all subjects. Data are color coded by motion coherence. Error bars indicate 95% CIs. For most subjects, observed thresholds are significantly greater than predicted, especially for coherences greater than 25%. For comparison, analogous data from monkeys and humans (black triangles and squares, respectively) are shown from a previous study involving a fixed-duration version of the same task (Fetsch et al., 2009).

DOI: 10.7554/eLife.03005.004

The following figure supplements are available for figure 2:

Figure supplement 1. Discrimination thresholds for all subjects and conditions.

DOI: 10.7554/eLife.03005.005

the unimodal conditions, as expected by standard optimal multisensory integration models. To summarize this contrast, we compare the ratio of observed to predicted thresholds in the combined condition for our subjects to human and monkey subjects performing a similar task in a fixed duration setting (Fetsch et al., 2009). We found this ratio to be significantly greater for our subjects (Figure 2C; two-sample *t* test, $t(77) = 3.245$, $p = 0.0017$). This indicates that, with respect to predictions of standard multisensory integration models, our subjects performed significantly worse than those engaged in a similar fixed-duration task.

A different picture emerges if we take not only discrimination thresholds but also reaction times into account. Short reaction times imply that subjects gather less information to make a decision, yielding greater discrimination thresholds. Longer reaction times may decrease thresholds, but at the cost of time. Consequently, if subjects decide more rapidly in the combined condition than the visual condition, they might feature higher discrimination thresholds in the combined condition even if they make optimal use of all available information. Thus, to assess if subjects perform optimal cue combination, we need to account for the timing of their decisions.

Average reaction times depended on stimulus condition, motion coherence, and heading direction. In general, reaction times were faster for larger heading magnitudes, and reaction times in the vestibular condition were faster than those in the visual condition (Figure 3 for subject D2, Figure 3—figure supplement 1 for other subjects). In the combined condition, however, reaction times were much shorter than those seen for the visual condition and were comparable to those of the vestibular condition (Figure 3). Thus, subjects spent substantially more time integrating evidence in the visual condition, which boosted their discrimination performance when compared to the combined condition. Note also that discrimination thresholds in the combined condition were substantially smaller than vestibular thresholds, especially at 70% coherence (Figures 2 and 3). Thus, adding optic flow to a

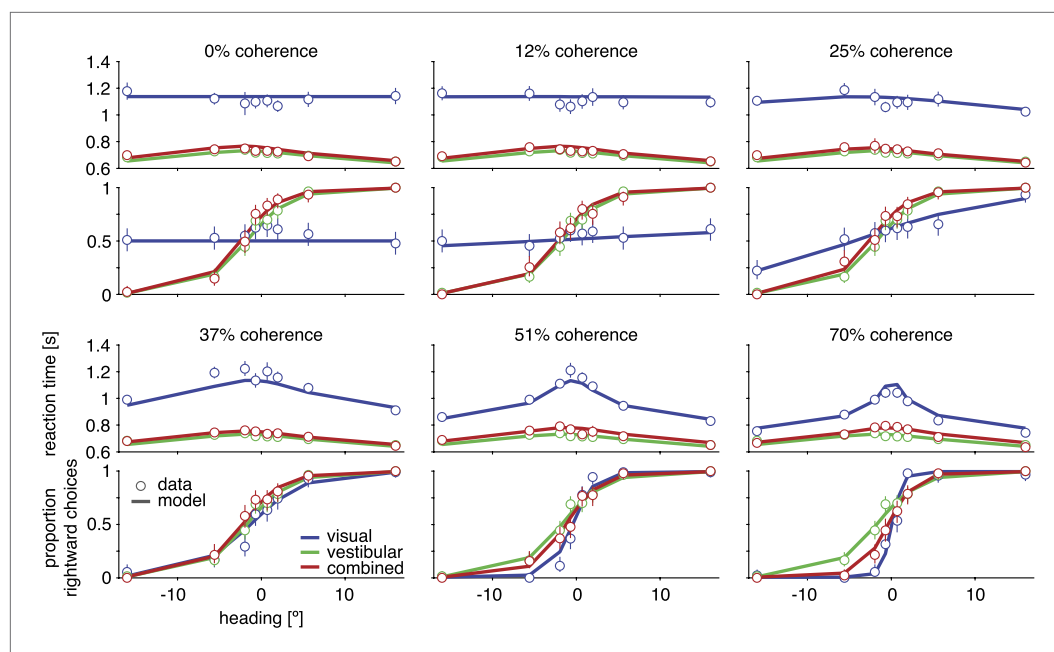


Figure 3. Discrimination performance and reaction times for subject D2. Behavioral data (symbols with error bars) and model fits (lines) are shown separately for each motion coherence. Top plot: reaction times as a function of heading; bottom plot: proportion of rightward choices as a function of heading. Mean reaction times are shown for correct trials, with error bars representing two SEM (in some cases smaller than the symbols). Error bars on the proportion rightward choice data are 95% confidence intervals. Although reaction times are only shown for correct trials, the model is fit to data from both correct and incorrect trials. See **Figure 3—figure supplement 1** for behavioral data and model fits for all subjects. **Figure 3—figure supplement 2** shows the fitted model parameters per subject.

DOI: [10.7554/eLife.03005.006](https://doi.org/10.7554/eLife.03005.006)

The following figure supplements are available for figure 3:

Figure supplement 1. Psychometric functions, chronometric functions, and model fits for all subjects.

DOI: [10.7554/eLife.03005.007](https://doi.org/10.7554/eLife.03005.007)

Figure supplement 2. Model parameters for fits of the optimal model and two alternative parameterizations.

DOI: [10.7554/eLife.03005.008](https://doi.org/10.7554/eLife.03005.008)

vestibular stimulus decreased the discrimination threshold with essentially no loss of speed. A similar overall pattern of results was observed for the other subjects (**Figure 3—figure supplement 1**). These data provide clear evidence that subjects made use of both visual and vestibular information to perform the reaction-time task, but the benefits of cue integration could not be appreciated by considering discrimination thresholds alone.

Modeling cue combination with a novel diffusion model

To investigate whether subjects accumulate evidence optimally across both time and sensory modalities, we built a model that integrates visual and vestibular cues optimally to perform the heading discrimination task, and we compare predictions of the model to data from our human subjects. The model builds upon the structure of diffusion models (DMs), which have previously been shown to account nicely for the tradeoff between speed and accuracy of decisions (**Ratcliff, 1978; Ratcliff and Smith, 2004; Palmer et al., 2005**). Additionally, DMs are known to optimally integrate evidence over time (**Laming, 1968; Bogacz et al., 2006**), given that the reliability of the evidence is time-invariant (such that, at any point in time from stimulus onset, the stimulus provides the same amount of information about the task variable). However, DMs have neither been used to integrate evidence from several sources, nor to handle evidence whose reliability changes over time, both of which are required for our purposes.

In the context of heading discrimination, a standard DM would operate as follows (**Figure 4A**): consider a diffusing particle with dynamics given by $\dot{x} = k \sin(h) + \eta(t)$, where h is the heading direction,

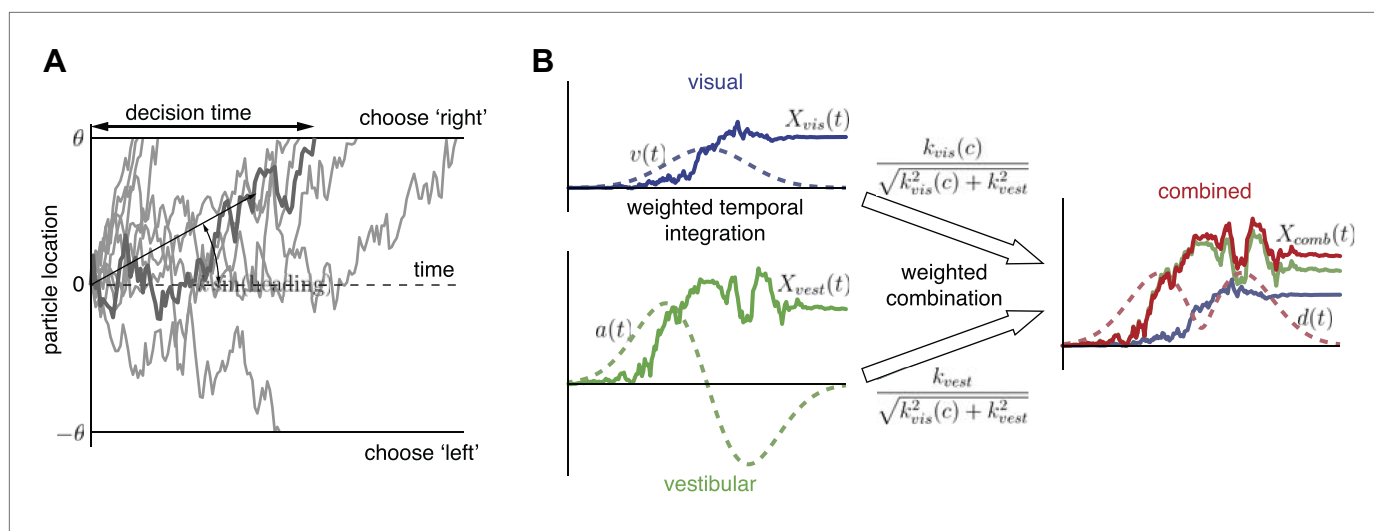


Figure 4. Extended diffusion model (DM) for heading discrimination task. **(A)** A drifting particle diffuses until it hits the lower or upper bound, corresponding to choosing 'left' or 'right' respectively. The rate of drift (black arrow) is determined by heading direction. The time at which a bound is hit corresponds to the decision time. 10 particle traces are shown for the same drift rate, corresponding to one incorrect and nine correct decisions. **(B)** Despite time-varying cue sensitivity, optimal temporal integration of evidence in DMs is preserved by weighting the evidence by the momentary measure of its sensitivity. The DM representing the combined condition is formed by an optimal sensitivity-weighted combination of the DMs of the unimodal conditions.

DOI: 10.7554/eLife.03005.009

k is a positive constant relating particle drift to heading direction, and $\eta(t)$ is unit variance Gaussian white noise. The particle starts at $x(0) = 0$, drifts with an average slope given by $k \sin(h)$, and diffuses until it hits either the upper bound θ or the lower bound $-\theta$, corresponding to rightward and leftward choices, respectively. The decision time is determined by when the particle hits a bound. Larger $|h|$'s lead to shorter decision times and more correct decisions because the drift rate is greater. Lower bound levels, $|\theta|$, also lead to shorter decision times but more incorrect decisions. Errors (hitting bound θ when $h < 0$, or hitting bound $-\theta$ when $h > 0$) can occur due to the stochasticity of particle motion, which is meant to capture the variability of the momentary sensory evidence. The Fisher information in $x(t)$ regarding h , a measure of how much information $x(t)$ provides for discriminating heading (Papoulis, 1991), is $I_x(\sin(h)) = k^2$ per second, showing that k is a measure of the subject's sensitivity to changes in heading direction. This sensitivity depends on the subject's effectiveness in estimating heading from the cue, which in turn is influenced by the reliability of the cue itself (e.g., coherence).

Now consider both a visual (vis) and a vestibular (vest) source of evidence regarding h , $\dot{x}_{vis} = k_{vis}(c) \sin(h) + \eta_{vis}(t)$ and $\dot{x}_{vest} = k_{vest} \sin(h) + \eta_{vest}(t)$, where $k_{vis}(c)$ indicates that the sensitivity to the cue in the visual modality depends on motion coherence, c . Combining these two sources of evidence by a simple sum, $\dot{x}_{vis} + \dot{x}_{vest}$, would amount to adding noise to \dot{x}_{vest} for low coherences ($k_{vis}(c) \approx 0$), which is clearly suboptimal. Rather, it can be shown that the two particle trajectories are combined optimally by weighting their rates of change in proportion to their relative sensitivities (see **Supplementary file 1** for derivation):

$$\dot{x}_{comb} = \sqrt{\frac{k_{vis}^2(c)}{k_{vis}^2(c) + k_{vest}^2}} \dot{x}_{vis} + \sqrt{\frac{k_{vest}^2}{k_{vis}^2(c) + k_{vest}^2}} \dot{x}_{vest}. \quad (1)$$

This allows us to model the combined condition by a single new DM, $\dot{x}_{comb} = k_{comb}(c) \sin(h) + \eta_{comb}(t)$, which is optimal because it preserves all information contained in both x_{vis} and x_{vest} (Figure 4B; see 'Materials and methods' and **Supplementary file 1** for a formal treatment). The sensitivity (drift rate coefficient) in the combined condition,

$$k_{comb}(c) = \sqrt{k_{vis}^2(c) + k_{vest}^2}, \quad (2)$$

is a combination of the sensitivities of the unimodal conditions and is therefore always greater than the largest unimodal sensitivity.

So far we have assumed that the reliability of each cue is time-invariant. However, as the motion velocity changes over time, so does the amount of information about h provided by each cue, and with it the subject's sensitivity to changes in h . For the vestibular and visual conditions, motion acceleration $a(t)$ and motion velocity $v(t)$, respectively, are assumed to be the physical quantities that modulate cue sensitivity ('Materials and methods' and 'Discussion'). To account for these dynamics, the DMs are modified to $\dot{x}_{\text{vest}} = a(t)k_{\text{vest}} \sin(h) + \eta_{\text{vest}}(t)$ and $\dot{x}_{\text{vis}} = v(t)k_{\text{vis}}(c) \sin(h) + \eta_{\text{vis}}(t)$. Note that once the drift rate in a DM changes with time, it generally loses its property of integrating evidence optimally over time. For example, at the beginning of each trial when motion velocity is low, \dot{x}_{vis} is dominated by noise and integrating \dot{x}_{vis} is fruitless. Fortunately, weighting the momentary visual evidence, \dot{x}_{vis} , by the velocity profile recovers optimality of the DM ('Materials and methods'). This temporal weighting causes the visual evidence to contribute more at high velocities, while the noise is downweighted at low velocities. Similarly, vestibular evidence is weighted by the time course of acceleration. The new, weighted particle trajectories are described by the DMs $\dot{X}_{\text{vis}} = v(t)\dot{x}_{\text{vis}}$ and $\dot{X}_{\text{vest}} = a(t)\dot{x}_{\text{vest}}$. The two unimodal DMs are combined as before, resulting in the combined DM given by $\dot{X}_{\text{comb}} = d(t)\dot{x}_{\text{comb}}$, where the sensitivity profile $d(t)$ is a weighted combination of the unimodal sensitivity profiles,

$$d(t) = \sqrt{\frac{k_{\text{vis}}^2(c)}{k_{\text{comb}}^2(c)} v^2(t) + \frac{k_{\text{vest}}^2}{k_{\text{comb}}^2(c)} a^2(t)}. \quad (3)$$

(Figure 4B; see [Supplementary file 1](#) for derivation). These modifications to the standard DM are sufficient to integrate evidence optimally across time and sensory modalities, even as the sensitivity to the evidence changes over time.

The model assumes that subjects know their cue sensitivities, $k_{\text{vis}}(c)$ and k_{vest} , as well as the temporal sensitivity profiles, $a(t)$ and $v(t)$, of each stimulus. In this respect, our model provides an upper bound on performance, since subjects may not have perfect knowledge of these variables, especially since stimulus modalities and visual motion coherence values are randomized across trials ('Discussion').

Quantitative assessment of cue combination performance

We tested whether subjects combined evidence optimally across both time and cues by evaluating how well the model outlined above could explain the observed behavior. The bounds, θ , of the modified DM, and the sensitivity parameters (k_{vis} , k_{vest} and k_{comb}), were allowed to vary between the visual, vestibular, and combined conditions. Varying the bound was essential to capture the deviation of the discrimination threshold in the combined condition from that predicted by traditional cue combination models (Figure 2). Indeed, this discrimination threshold is inversely proportional to bound and sensitivity (see [Supplementary file 1](#)). Since the sensitivity in the bimodal condition is not a free parameter (it is determined by Equation 2), the height of the bound is the only parameter that could modulate the discrimination thresholds.

The noise terms η_{vis} and η_{vest} play crucial roles in the model, as they relate to the reliability of the momentary sensory evidence. To specify the manner in which such noise may depend on motion coherence, we relied on fundamental assumptions about how optic flow stimuli are represented by the brain. We assumed that heading is represented by a neural population code in which neurons have heading tuning curves that, within the range of heading tested in this experiment ($\pm 16^\circ$, Figure 5A), differ in their heading preferences but have similar shapes. This is broadly consistent with data from area MSTd (Fetsch et al., 2011), but the exact location of such a code is not important for our argument. For low coherence, motion energy in the stimulus is almost uniform for all heading directions, such that all neurons in the population fire at approximately the same rate (Figure 5A, dark blue curve). For high coherence, population neural activity is strongly peaked around the actual heading direction (Figure 5A, cyan curve) (Morgan et al., 2008; Fetsch et al., 2011).

Based on this representation, and assuming that the response variability of the neurons belongs to the exponential family with linear sufficient statistics (Ma et al., 2006) (an assumption consistent with in vivo data [Graf et al., 2011]), heading discrimination can be performed optimally by a weighted sum of the activity of all neurons, with weights monotonically related to the preferred heading of each neuron. For a straight forward heading, $h = 0$, this sum should be 0, and for $h > 0$ (or $h < 0$) it should be positive (or negative), thus sharing the basic properties of the momentary evidence, \dot{x} , in our DM. This allowed us to deduce the mean and variance of the momentary evidence driving \dot{x} , based on what

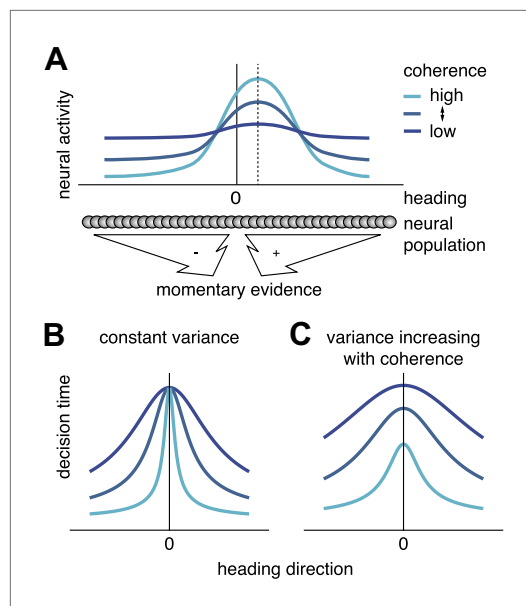


Figure 5. Scaling of momentary evidence statistics of the diffusion model (DM) with coherence. **(A)** Assumed neural population activity giving rise to the DM mean and variance of the momentary evidence, and their dependence on coherence. Each curve represents the activity of a population of neurons with a range of heading preferences, in response to optic flow with a particular coherence and a heading indicated by the dashed vertical line. **(B)** Expected pattern of reaction times if variance is independent of coherence. If neither the DM bound nor the DM variance depend on coherence, the DM predicts the same decision time for all small headings, regardless of coherence. This is due to the DM drift rate, $k_{vis}(c)\sin(h)$ being close to 0 for small headings, $h \approx 0$, independent of the DM sensitivity $k_{vis}(c)$. **(C)** Expected pattern of reaction times when variance scales with coherence. If both DM sensitivity and DM variance scale with coherence while the bound remains constant, the DM predicts different decision times across coherences, even for small headings. Greater coherence causes an increase in variance, which in turn causes the bound to be reached more quickly for higher coherences, even if the heading, and thus the drift rate, is small.

DOI: [10.7554/eLife.03005.010](https://doi.org/10.7554/eLife.03005.010)

bound $\theta_{\sigma,vest}$ in the vestibular condition do not depend on motion coherence and were thus model parameters that were fitted directly.

Observed reaction times were assumed to be composed of the decision time and some non-decision time. The decision time is the time from the start of integrating evidence until a decision is made, as predicted by the diffusion model. The non-decision time includes the motor latency and the time from stimulus onset to the start of integrating evidence. As the latter can vary between different modalities, we allowed it to differ between visual, vestibular, and combined conditions, but not for different coherences, thus introducing the model parameters $t_{nd,vis}$, $t_{nd,vest}$, and $t_{nd,comb}$. Although the fitted non-decision times were similar across stimulus conditions for most subjects (**Figure 3—figure supplement 2**), a model assuming a single non-decision time resulted in a small but significant decrease in fit quality (**Figure 7—figure supplement 2A**). Overall, 12 parameters were used to model

we know about the neural responses. First, the sensitivity, $k_{vis}(c)$, which determines how optic flow modulates the mean drift rate of \dot{x} , scales in proportion with the ‘peakedness’ of the neural activity, which in turn is proportional to coherence. We assumed a functional form of $k_{vis}(c)$ given by $a_{vis}c^{\gamma_{vis}}$, where a_{vis} and γ_{vis} are positive parameters. Second, the variance of \dot{x} is assumed to be the sum of the variances of the neural responses. Since experimental data suggest that the variance of these responses is proportional to their firing rate (Tolhurst et al., 1983), the sum of the variances is proportional to the area underneath the population activity profile (**Figure 5A**). Based on the experimental data of Britten et al. (Heuer and Britten, 2007), this area was assumed to scale roughly linearly with coherence, such that the variance of \dot{x} is proportional to $1 + b_{vis}c^{\gamma_{vis}}$ with free parameters b_{vis} and γ_{vis} , the latter of which captures possible deviations from linearity. We further assumed the DM bound to be independent of coherence, and given by $\theta_{\sigma,vis}$. Thus, the effect of motion coherence on the momentary evidence in the DM was modeled by four parameters: a_{vis} , γ_{vis} , b_{vis} , and $\theta_{\sigma,vis}$.

The above scaling of the diffusion variance by coherence, which is a consequence of the neural code for heading, makes an interesting prediction: reaction times for headings near straight ahead should be inversely proportional to coherence in the visual condition, even though the mean drift rate, $k_{vis}(c)\sin(h)$, is very close to 0. This is indeed what we observed: subjects tended to decide faster for higher coherences even when $h \approx 0$ (**Figure 3, Figure 3—figure supplement 1**). This aspect of the data can only be captured by the model if the DM variance is allowed to change with coherence (**Figure 5B,C**).

To summarize, in the combined condition, the diffusion variance was assumed to be proportional to $1 + b_{comb}c^{\gamma_{comb}}$, while the bound was fixed at $\theta_{\sigma,comb}$. By contrast, the diffusion rate (sensitivity) cannot be modeled freely but rather needs to obey $k_{comb}(c) = \sqrt{k_{vis}^2(c) + k_{vest}^2}$ in order to ensure optimal cue combination. The sensitivity k_{vest} and

cue sensitivities, bounds, variances, and non-decision times in all conditions, and these 12 parameters were used to fit 312 data points for subjects that were tested with 6 coherences (168 data points for the three-coherence version). An additional 14 parameters (8 parameters for the three-coherence version; one bias parameter per coherence/condition, one lapse parameter across all condition) controlled for biases in the motion direction percept and for lapses of attention that were assumed to lead to random choices ('Materials and methods'). Although these additional parameters were necessary to achieve good model fits (**Figure 7—figure supplement 2A**), it is critical to note that they could not account for differences in heading thresholds or reaction times across stimulus conditions. As such, the additional parameters play no role in determining whether subjects perform optimal multisensory integration. Alternative parameterizations of how drift rates and bounds depend on motion coherence yielded qualitatively similar results, but caused the model fits to worsen decisively (**Supplementary file 1; Figure 7—figure supplement 2A**).

Critically, our model predicts that the unimodal sensitivities $k_{vis}(c)$ and k_{vest} relate to the combined value by $k_{comb}^{predicted}(c) = \sqrt{k_{vis}(c)^2 + k_{vest}^2}$, if subjects accumulate evidence optimally across cues. To test this prediction, we fitted separately the unimodal and combined sensitivities, $k_{vis}(c)$, k_{vest} and k_{comb} to the complete data set from each individual subject using maximum likelihood optimization ('Materials and methods'), and then compared the fitted values of k_{comb} to the predicted values, $k_{comb}^{predicted}(c)$. Predicted and observed sensitivities for the combined condition are virtually identical (**Figure 6**), providing strong support for near-optimal cue combination across both time and cues. Remarkably, for low coherences at which optic flow provides no useful heading information, the sensitivity in the combined condition was not significantly different from that of the vestibular condition (**Figure 6**). Thus, subjects were able to completely suppress noisy visual information and rely solely on vestibular input, as predicted by the model.

Having established that cue sensitivities combine according to **Equation 2**, the model was then fit to data from each individual subject under the assumption of optimal cue combination. Model fits

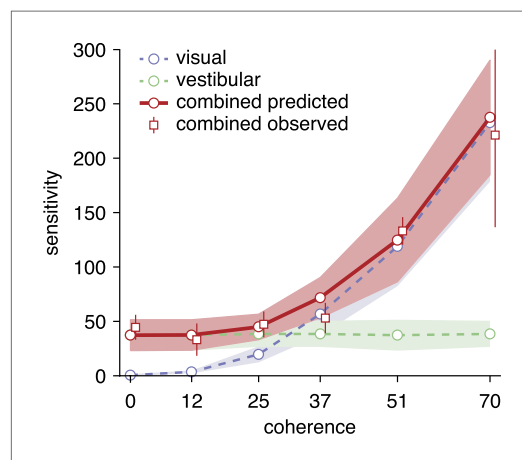


Figure 6. Predicted and observed sensitivity in the combined condition. The sensitivity parameter measures how sensitive subjects are to a change of heading. The solid red line shows predicted sensitivity for the combined condition, as computed from the sensitivities of the unimodal conditions (dashed lines). The combined sensitivity measured by fitting the model to each coherence separately (red squares) does not differ significantly from the optimal prediction, providing strong support to the hypothesis that subjects accumulate evidence near-optimally across time and cues. Data are averaged across datasets (except 0%, 12%, 51% coherence: only datasets B2, D2, F2), with shaded areas and error bars showing the 95% CIs. DOI: [10.7554/eLife.03005.011](https://doi.org/10.7554/eLife.03005.011)

are shown as solid curves for example subject D2 (**Figure 3**), as well as for all other subjects (**Figure 3—figure supplement 1**). Sensitivity parameters, bounds, and non-decision times resulting from the fits are also shown for each subject, condition, and coherence (**Figure 3—figure supplement 2**). For 8 of 10 datasets, the model explains more than 95% of the variance in the data (adjusted $R^2 > 0.95$), providing additional evidence for near-optimal cue combination across both time and cues (**Figure 7A**). The subjects associated with these datasets show a clear decrease in reaction times with larger $|h|$, and this effect is more pronounced in the visual condition than in the vestibular and combined conditions (**Figure 3, Figure 3—figure supplement 1**). The remaining two subjects (C and F) feature qualitatively different behavior and lower R^2 values of approximately 0.80 and 0.90, respectively (**Figure 3—figure supplement 1**). These subjects showed little decline in reaction times with larger values of $|h|$, and their mean reaction times were more similar across the visual, vestibular and combined conditions.

Critically, the model nicely captures the observation that the psychophysical threshold in the combined condition is typically greater than that for the visual condition, despite near-optimal combination of momentary evidence from the visual and vestibular modalities (e.g., **Figure 3, 70% coherence, Figure 2—figure supplement 1**,

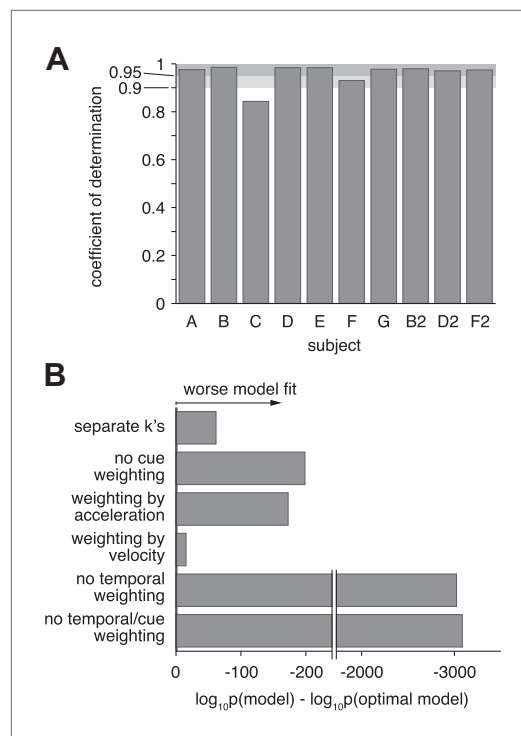


Figure 7. Model goodness-of-fit and comparison to alternative models. **(A)** Coefficient of determination (adjusted R^2) of the model fit for each of the ten datasets. **(B)** Bayes factor of alternative models compared to the optimal model. The abscissa shows the base-10 logarithm of the Bayes factor of the alternative models vs the optimal model (negative values mean that the optimal model out-performs the alternative model). The gray vertical line close to the origin (at a value of -2 on the abscissa) marks the point at which the optimal model is 100 times more likely than each alternative, at which point the difference is considered 'decisive' (Jeffreys, 1998). Only the 'separate k's' model has more parameters than the optimal model, but the Bayes factor indicates that the slight increase in goodness-of-fit does not justify the increased degrees of freedom. The 'no cue weighting' model assumes that visual and vestibular cues are weighted equally, independent of their sensitivities. The 'weighting by acceleration' and 'weighting by velocity' models assume that the momentary evidence of both cues is weighted by the acceleration and velocity profile of the stimulus, respectively. The 'no temporal weighting' model assumes that the evidence is not weighted over time according to its sensitivity. The 'no cue/temporal weighting' model lacks both weighting of cues by sensitivity and weighting by temporal profile. All of the tested alternative models explain the data decisively worse than the optimal model. **Figure 7—figure supplement 1** shows how individual subjects contribute to this model comparison, and the results of a more conservative Bayesian random-effects model comparison that supports same conclusion. **Figure 7—figure supplement 2** Figure 7. Continued on next page

Figure 3—figure supplement 1). Thus, the model fits confirm quantitatively that apparent sub-optimality in psychophysical thresholds can arise even if subjects combine all cues in a statistically optimal manner, emphasizing the need for a computational framework that incorporates both decision accuracy and speed.

Alternative models

To further assess and validate the critical design features of our modified DM, we evaluated six alternative (mostly sub-optimal) versions of the model to see if these variants are able to explain the data equally well. We compared these variants to the optimal model using Bayesian model comparison, which trades off fit quality with model complexity to determine whether additional parameters significantly improve the fit (Goodman, 1999).

With regard to optimality of cue integration across modalities, we examined two model variants. The first variant (also used to generate **Figure 6**) eliminates the relationship, $k_{\text{comb}}(c) = \sqrt{k_{\text{vis}}^2(c) + k_{\text{vest}}^2}$ (Equation 2), between the sensitivity parameters in the combined and single-cue conditions. Instead, this variant allows independent sensitivity parameters for the combined condition at each coherence, thus introducing one additional parameter per coherence. Since this variant is strictly more general than the optimal model, it must fit the data at least as well. However, if the subjects' behavior is near optimal, the additional degrees of freedom in this variant should not improve the fit enough to justify the addition of these parameters. This is indeed what we found by Bayesian model comparison (**Figure 7B**, 'separate k's'), which shows the optimal model to be $\sim 10^{70}$ times more likely than the variant with independent values of $k_{\text{comb}}(c)$. This is well above the threshold value that is considered to provide 'decisive' evidence in favor of the optimal model (we use Fisher's definition of decisive [Jeffreys, 1998] according to which a model is said to be decisively better if it is >100 times more likely to have generated the data). The second model variant had the same number of parameters as the optimal model, but assumed that the cues are always weighted equally. Evidence in the combined condition was given by the simple average, $\dot{X}_{\text{comb}} = \frac{1}{2}(\dot{X}_{\text{vis}} + \dot{X}_{\text{vest}})$, ignoring cue sensitivities. The resulting fits (**Figure 7B**, 'no cue weighting') are also decisively worse than those of the optimal model. Together, these model variants strongly support the hypothesis that subjects weight cues according to their relative

Figure 7. Continued

compares the proposed model to ones with alternative parameterizations.

DOI: [10.7554/eLife.03005.012](https://doi.org/10.7554/eLife.03005.012)

The following figure supplements are available for figure 7:

Figure supplement 1. Model comparison per subject, and random-effects model comparison.

DOI: [10.7554/eLife.03005.013](https://doi.org/10.7554/eLife.03005.013)

Figure supplement 2. Model comparison for models with alternative parameterization.

DOI: [10.7554/eLife.03005.014](https://doi.org/10.7554/eLife.03005.014)

decisively (**Figure 7B**, 'weighting by acceleration'). Assuming that the weighting of both modalities followed the velocity profile of the stimulus also decisively reduced fit quality (**Figure 7B**, 'weighting by velocity'), although this effect was not consistent across subjects (**Figure 7—figure supplement 1A**). If we completely removed temporal weighting of cues from the model, fits were dramatically worse than the optimal model (**Figure 7B**, 'no temporal weighting'). Finally, for completeness, we also tested a model variant that neither performs temporal weighting of cues nor considers the relative sensitivity to the cues. Again, this model variant fit the data decisively worse than the optimal model (**Figure 7B**, 'no cue/temporal weighting'). Thus, subjects seem to be able to take into account their sensitivity to the evidence across time as well as across cues. All of these model comparisons received further support from a more conservative random-effects Bayesian model comparison, shown in **Figure 7—figure supplement 1B,C**.

Finally, we also considered if a parallel race model could account for our data. The parallel race model (*Raab, 1962; Miller, 1982; Townsend and Wenger, 2004; Otto and Mamassian, 2012*) postulates that the decision in the combined condition emerges from the faster of two independent races toward a bound, one for each sensory modality. Because it does not combine information across modalities, the parallel race model predicts that decisions in the combined condition are caused by the faster modality. Consequently, choices in the combined condition are unlikely to be more correct (on average) than those of the faster unimodal condition. For all but one subject, the vestibular modality is substantially faster, even when compared to the visual modality at high coherence and controlling for the effect of heading direction (2-way ANOVA, $p < 0.0001$ for all subjects except C). Critically, all of these subjects feature significantly lower psychophysical thresholds in the combined condition than in the vestibular condition ($p < 0.039$ for all subjects except subject C, $p = 0.210$, **Supplementary file 2A**). Furthermore, we performed standard tests (Miller's bound and Grice's bound) that compare the observed distribution of reaction times with that predicted by the parallel race model (*Miller, 1982; Grice et al., 1984*). These tests revealed that all but two subjects made significantly slower decisions than predicted by the parallel race model for most coherence/heading combinations ($p < 0.05$ for all subjects except subjects F and B2; **Supplementary file 2B**), and no subject was faster than predicted ($p > 0.05$, all subjects; **Supplementary file 2B**). Based on these observations, we can reject the parallel race model as a viable hypothesis to explain the observed behavior.

Discussion

We have shown that, when subjects are allowed to choose how long to accumulate evidence in a cue integration task, their behavior no longer follows the standard predictions of optimal cue integration theory that normally apply when stimulus presentation time is controlled by the experimenter. Particularly, they feature worse discrimination performance (higher psychophysical thresholds) in the combined condition than would be predicted from the unimodal conditions—in some cases even worse than the better of the two unimodal conditions. This occurs because subjects tend to decide more quickly in the combined condition than in the more sensitive unimodal condition and thus have less time to accumulate evidence. This indicates that a more general definition of optimal cue integration must incorporate reaction times. Indeed, subjects' behavior could be reproduced by an extended diffusion model that takes into account both speed and accuracy, thus suggesting that subjects accumulate evidence across both time and cues in a statistically near-optimal manner (i.e., with minimal information loss) despite their reduced discrimination performance in the combined condition.

sensitivities, as given by **Equation 2**. These effects were largely consistent across individual subjects (**Figure 7—figure supplement 1A**).

To test the other key assumption of our model—that subjects temporally weight incoming evidence according to the profile of stimulus information—we tested three model variants that modified how temporal weighting was performed without changing the number of parameters in the model. If we assumed that the temporal weighting of both modalities followed the acceleration profile of the stimulus while leaving the model otherwise unchanged, the model fit worsened

Previous work on optimal cue integration (e.g., *Ernst and Banks, 2002; Battaglia et al., 2003; Knill and Saunders, 2003; Fetsch et al., 2009*) was based on experiments that employed fixed-duration stimuli and was thus able to ignore how subjects accumulate evidence over time. Moreover, previous work relied on the implicit assumption that subjects make use of all evidence throughout the duration of the stimulus. However, this assumption need not be true and has been shown to be violated even for short presentation durations (*Mazurek et al., 2003; Kiani et al., 2008*). Therefore, apparent sub-optimality in some previous studies of cue integration or in some individual subjects (*Battaglia et al., 2003; Fetsch et al., 2009*) might be attributable to either truly sub-optimal cue combination, to subjects halting evidence accumulation before the end of the stimulus presentation period, or to the difficulty in estimating stimulus processing time (*Stanford et al., 2010*). Unfortunately, these potential causes cannot be distinguished using a fixed-duration task. Allowing subjects to register their decisions at any time during the trial alleviates this potential confound.

We model subjects' decision times by assuming an accumulation-to-bound process. In the multisensory context, this raises the question of whether evidence accumulation is bounded for each modality separately, as assumed by the parallel race model, or whether evidence is combined across modalities before being accumulated toward a single bound, as in co-activation models and our modified diffusion model. Based on our behavioral data, we can rule out parallel race models, as they cannot explain lower psychophysical thresholds (better sensitivity) in the combined condition relative to the faster vestibular condition. Further evidence against such models is provided by neurophysiological studies which demonstrate that visual and vestibular cues to heading converge in various cortical areas, including areas MSTd (*Gu et al., 2006*), VIP (*Schlack et al., 2005; Chen et al., 2011b*), and VPS (*Chen et al., 2011a*). Activity in area MSTd can account for sensitivity-based cue weighting in a fixed-duration task (*Fetsch et al., 2011*), and MSTd activity is causally related to multi-modal heading judgments (*Britten and van Wezel, 1998, 2002; Gu et al., 2012*). These physiological studies strongly suggest that visual and vestibular signals are integrated in sensory representations prior to decision-making, inconsistent with parallel race models.

Our model makes the assumption that sensory signals are integrated prior to decision-making and is in this sense similar to co-activation models that have been used previously to model reaction times in multimodal settings (*Miller, 1982; Corneil et al., 2002; Townsend and Wenger, 2004*). However, it differs from these models in important aspects. First, co-activation models have been introduced to explain reaction times that are faster than those predicted by parallel race models (*Raab, 1962; Miller, 1982*). Our subjects, in contrast, feature reaction times that are slower than those of parallel race models in almost all conditions (*Supplementary file 2B*). We capture this effect by an elevated effective bound in the combined condition as compared to the faster vestibular condition, such that cue combination remains optimal despite longer reaction times. Second, co-activation models usually combine inputs from the different modalities by a simple sum (e.g., *Townsend and Wenger, 2004*). This entails adding noise to the combined signal if the sensitivity to one of the modalities is low, which is detrimental to discrimination performance. In contrast, we show that different cues need to be weighted according to their sensitivities to achieve statistically optimally integration of multisensory evidence at each moment in time (*Equation 2*).

Another alternative to co-activation models are serial race models, which posit that the race corresponding to one cue needs to be completed before the other one starts (e.g., *Townsend and Wenger, 2004*). These models can be ruled out by observing that they predict reaction times in the combined condition to be longer than those in the slower of the two unimodal conditions. This is clearly violated by the subjects' behavior.

Optimal accumulation of evidence over time requires the momentary evidence to be weighted according to its associated sensitivity. For the vestibular modality, we assume that the temporal profile of sensitivity to the evidence follows acceleration. This may appear to conflict with data from multimodal areas MSTd, VIP, and VPS, where neural activity in response to self-motion reflects a mixture of velocity and acceleration components (*Fetsch et al., 2010; Chen et al., 2011a*). Note, however, that the vestibular stimulus is initially encoded by otolith afferents in terms of acceleration (*Fernandez and Goldberg, 1976*). Thus, any neural representation of vestibular stimuli in terms of velocity requires a temporal integration of the acceleration signal, and this integration introduces temporal correlations into the signal. As a consequence, a neural response that is maximal at the time of peak stimulus velocity does not imply a simultaneous peak in the information coded about heading direction. Rather, information still follows the time course of its original encoding, which is in terms of acceleration.

In contrast, the time course of the sensitivity to the visual stimulus is less clear. For our model we have intuitively assumed it to follow the velocity profile of the stimulus, as information per unit time about heading certainly increases with the velocity of the optic flow field, even when there is no acceleration. This assumption is supported by a decisively worse model fit if we set the weighting of the visual momentary evidence to follow the acceleration profile (**Figure 7B**, 'weighted by acceleration'). Nonetheless, we cannot completely exclude any contribution of acceleration components to visual information (**Lisberger and Movshon, 1999; Price et al., 2005**). In any case, our model fits make clear that temporal weighting of vestibular and visual inputs is necessary to predict behavior when stimuli are time-varying.

The extended DM model described here makes the strong assumption that cue sensitivities are known before combining information from the two modalities, as these sensitivities need to be known in order to weight the cues appropriately. As only the sensitivity to the visual stimulus changes across trials in our experiment, it is possible that subjects can estimate their sensitivity (as influenced by coherence) during the initial low-velocity stimulus period (**Figure 1C**) in which heading information is minimal but motion coherence is salient. Thus, for our task, it is reasonable to assume that subjects can estimate their sensitivity to cues. We have recently begun to consider how sensitivity estimation and cue integration can be implemented neurally. The neural model (**Onken et al., 2012**. Near optimal multisensory integration with nonlinear probabilistic population codes using divisive normalization. The Society for Neuroscience annual meeting 2012) estimates the sensitivity to the visual input from motion sensitive neurons and uses this estimate to perform near-optimal multisensory integration with generalized probabilistic population codes (**Ma et al., 2006; Beck et al., 2008**) using divisive normalization. We intend to extend this model to the integration of evidence over time to predict neural responses (e.g., in area LIP) that should roughly track the temporal evolution of the decision variable ($x_{comb}(t)$, 'Materials and methods') in the DM model. This will make predictions for activity in decision-making areas that can be tested in future experiments.

In closing, our findings establish that conventional definitions of optimality do not apply to cue integration tasks in which subjects' decision times are unconstrained. We establish how sensory evidence should be weighted across modalities and time to achieve optimal performance in reaction-time tasks, and we show that human behavior is broadly consistent with these predictions but not with alternative models. These findings, and the extended diffusion model that we have developed, provide the foundation for building a general understanding of perceptual decision-making under more natural conditions in which multiple cues vary dynamically over time and subjects make rapid decisions when they have acquired sufficient evidence.

Materials and methods

Subjects and apparatus

Seven subjects (3 males) aged 23–38 years with normal or corrected-to-normal vision and no history of vestibular deficits participated in the experiments. All subjects but one were informed of the purposes of the study. Informed consent was obtained from all participants and all procedures were reviewed and approved by the Washington University Office of Human Research Protections (OHRP), Institutional Review Board (IRB; IRB ID# 201109183). Consent to publish was not obtained in writing, as it was not required by the IRB, but all subjects were recruited for this purpose and approved verbally. Of these subjects, three (subjects B, D, F; 1 male) participated in a follow-up experiment roughly 2 years after the initial data collection, with six coherence levels instead of the original three. The six-coherence version of their data is referred to as B2, D2, and F2. Procedures for the follow-up experiment were approved by the Institutional Review Board for Human Subject Research for Baylor College of Medicine and Affiliated Hospitals (BCM IRB, ID# H-29411) and informed consent and consent to publish was given again by all three subjects.

The apparatus, stimuli, and task design have been described in detail previously (**Fetsch et al., 2009; Gu et al., 2010**), and are briefly summarized here. Subjects were seated comfortably in a padded racing seat that was firmly attached to a 6-degree-of-freedom motion platform (MOOG, Inc). A 3-chip DLP projector (Galaxy 6; Barco, Kortrijk, Belgium) was mounted on the motion platform behind the subject and front-projected images onto a large (149 × 127 cm) projection screen via a mirror mounted above the subject's head. The viewing distance to the projection screen was ~70 cm, thus allowing for a field of view of ~94° × 84°. Subjects were secured to the seat using a 5-point racing

harness, and a custom-fitted plastic mask immobilized the head against a cushioned head mount. Seated subjects were enclosed in a black aluminum superstructure, such that only the display screen was visible in the darkened room. To render stimuli stereoscopically, subjects wore active stereo shutter glasses (CrystalEyes 3; RealD, Beverly Hills, CA) which restricted the field of view to $\sim 90^\circ \times 70^\circ$. Subjects were instructed to look at a centrally-located, head-fixed target throughout each trial. Sounds from the motion platform were masked by playing white noise through headphones. Behavioral task sequences and data acquisition were controlled by Matlab and responses were collected using a button box.

Visual stimuli were generated by an OpenGL accelerator board (nVidia Quadro FX1400), and were plotted with sub-pixel accuracy using hardware anti-aliasing. In the visual and combined conditions, visual stimuli depicted self-translation through a 3D cloud of stars distributed uniformly within a virtual space 130 cm wide, 150 cm tall, and 75 cm deep. Star density was $0.01/\text{cm}^3$, with each star being a $0.5 \text{ cm} \times 0.5 \text{ cm}$ triangle. Motion coherence was manipulated by randomizing the three-dimensional location of a percentage of stars on each display update while the remaining stars moved according to the specified heading. The probability of a single star following the trajectory associated with a particular heading for N video updates is therefore $(c/100)^N$, where c denotes motion coherence (ranging from 0–100%). At the largest coherence used here (70%), there is only a 3% probability that a particular star would follow the same trajectory for 10 display updates (0.17 s). Thus, it was practically not possible for subjects to track the trajectories of individual stars. This manipulation degraded optic flow as a heading cue and was used to manipulate visual cue reliability in the visual and combined conditions. 'Zero' coherence stimuli had c set to 0.1, which was practically indistinguishable from $c = 0$, but allowed us to maintain a precise definition of the correctness of the subject's choice.

Behavioral task

In all stimulus conditions, the task was a single-interval, two-alternative forced choice (2AFC) heading discrimination task. In each trial, human subjects were presented with a translational motion stimulus in the horizontal plane (Gaussian velocity profile; peak velocity, 0.403 m/s; peak acceleration, 0.822 m/s^2 ; total displacement, 0.3 m; maximum duration, 2 s). Heading was varied in small steps around straight ahead ($\pm 0.686^\circ$, $\pm 1.96^\circ$, $\pm 5.6^\circ$, $\pm 16^\circ$) and subjects were instructed to report (by a button press) their perceived heading (leftward or rightward relative to an internal standard of straight ahead) as quickly and accurately as possible. In the visual and combined conditions, cue reliability was varied across trials by randomly choosing the motion coherence of the visual stimulus from among either a group of three values (25%, 37%, and 70%, subjects A–G) or a group of six values (0%, 12%, 25%, 37%, 51%, and 70%, subjects B2, D2, F2). A coherence of 25% means that 25% of the dots move in a direction consistent with the subject's heading, whereas the remaining 75% of the dots are relocated randomly within the dot cloud. In the combined condition, visual and vestibular stimuli always specified the same heading (there was no cue conflict).

During the main phase of data collection, subjects were not informed about the correctness of their choices (no feedback). In the vestibular and combined conditions, platform motion was halted smoothly but rapidly immediately following registration of the decision, and the platform then returned to its original starting point. In the visual condition, the optic flow stimulus disappeared from the screen when a decision was made. In all conditions, 2.5 s after the decision, a sound informed the subjects that they could initiate the next trial by pushing a third button. Once a trial was initiated, the stimulus onset occurred following a randomized delay period (truncated exponential; mean, 987 ms). Prior to data collection, subjects were introduced to the task for 1–2 week 'training' sessions, in which they were informed about the correctness of their choices by either a low-frequency (incorrect) or a high-frequency (correct) sound. The training period was terminated once their behavior stabilized across consecutive training sessions. During training, subjects were able to adjust their speed-accuracy trade-off based on feedback. During subsequent data collection, we did not observe any clear changes in the speed-accuracy trade-off exhibited by subjects.

Data analysis

Analyses and statistical tests were performed using MATLAB R2013a (The Mathworks, MA, USA).

For each subject, discrimination thresholds were determined separately for each combination of stimulus modality (visual-only, vestibular-only, combined) and coherence (25%, 37%, and 70% for

subjects A–G; 0%, 12%, 25%, 37%, 51%, and 70% for subjects B2, D2, F2) by plotting the proportion of rightward choices as a function of heading direction (**Figure 2A**). The psychophysical discrimination threshold was taken as the standard deviation of a cumulative Gaussian function, fitted by maximum likelihood methods. We assumed a common lapse rate (proportion of random choices) across all stimulus conditions, but allowed for a separate bias parameter (horizontal shift of the psychometric function) for each modality/coherence. Confidence intervals for threshold estimates were obtained by taking 5000 parametric bootstrap samples (**Wichmann and Hill, 2001**). These samples also form the basis for statistical comparisons of discrimination thresholds: two thresholds were compared by computing the difference between their associated samples, leading to 5000 threshold difference samples. Subsequently, we determined the fraction of differences that were below or above zero, depending on the directionality of interest. This fraction determined the raw significance level for accepting the null hypothesis (no difference). The reported significance levels are Bonferroni-corrected for multiple comparisons. All comparisons were one-tailed. Following traditional cue combination analyses (**Clark and Yuille, 1990**), the optimal threshold $\sigma_{pred,c}$ in the combined condition for coherence c was predicted from the visual threshold $\sigma_{vis,c}$ and the vestibular threshold σ_{vest} by $\sigma_{pred,c}^2 = \sigma_{vis,c}^2 \sigma_{vest}^2 / (\sigma_{vis,c}^2 + \sigma_{vest}^2)$. Confidence intervals and statistical tests were again based on applying this formula to individual bootstrap samples of the unimodal threshold estimates. **Supplementary file 2A** reports the p-values for all subjects and all comparisons.

For each dataset, we evaluated the absolute goodness-of-fit of the optimal model (**Figure 7A**) by finding the set of model parameters φ that maximized the likelihood of the observed choices and reaction times, and then computing the average coefficient of determination, $R^2(D\varphi) = \frac{1}{2}(R_{psych}^2(\varphi) + R_{chron}^2(\varphi))$. Here, $R_{psych}^2(\varphi)$ and $R_{chron}^2(\varphi)$ denote the adjusted R^2 values for the psychometric and chronometric functions, respectively, across all modalities/coherences. The value of R_{psych}^2 for the psychometric function was based on the probability of making a correct choice across all heading angles, coherences, and conditions, weighted by the number of observations, and adjusted for the number of model parameters. The same procedure, based on the mean reaction times, was used to find R_{chron}^2 , but we additionally distinguished between mean reaction times for correct and incorrect choices, and fitted both weighted by their corresponding number of observations (see SI for expressions for $R_{psych}^2(\varphi)$ and $R_{chron}^2(\varphi)$).

We compared different variants of the full model (**Figure 7B**) by Bayesian model comparison based on Bayes factors, which were computed as follows. First, we found for each model \mathcal{M} and subject s the set of parameters φ that maximized the likelihood, $\varphi_{s,\mathcal{M}}^* = \arg \max_{\varphi} p(\text{data of subj } s | \varphi, \mathcal{M})$. Second, we approximated the Bayesian model evidence, measuring the model posterior probability while marginalizing over the parameters, up to a constant by the Bayesian information criterion, $\ln p(\mathcal{M} | s) \approx -\frac{1}{2} \text{BIC}(s, \mathcal{M})$ with $\text{BIC}(s, \mathcal{M}) = -2 \ln p(s | \varphi_{s,\mathcal{M}}^*, \mathcal{M}) + k_{\mathcal{M}} \ln N_s$. Here, $k_{\mathcal{M}}$ is the number of parameters of model \mathcal{M} , and N_s is the number of trials for dataset s , respectively. Based on this, we computed the Bayes factor of model \mathcal{M} vs the optimal model \mathcal{M}_{opt} by pooling the model evidence over datasets, resulting in $\sum_s (\ln p(\mathcal{M} | s) - \ln p(\mathcal{M}_{opt} | s))$. These values, converted to a base-10 logarithm, are shown in **Figure 7B**. In this case, a negative \log_{10} -difference of 2 implies that the optimal model is 100 times more likely given the data than the alternative model, a difference that is considered decisive in favor of the optimal model (**Jeffreys, 1998**).

To determine the faster stimulus modality for each subject, we compared reaction times for the vestibular condition with those for the visual condition at 70% coherence. We tested the difference in the logarithm of these reaction times by a 2-way ANOVA with stimulus modality and heading direction as the two factors, and we report the main effect of stimulus modality on reaction times. Although we performed a log-transform of the reaction times to ensure their normality, a Jarque–Bera test revealed that normality did not hold for some heading directions. Thus, we additionally performed a Friedman test on subsampled data (to have the same number of trials per modality/heading) which supported the ANOVA result at the same significance level. In the main text, we only report the main effect of stimulus modality on reaction time from the 2-way ANOVA. Detailed results of the 2-way ANOVA, the Jarque–Bera test, and the Friedman test are reported for each subject in **Supplementary file 2C**.

The extended diffusion model

Here we outline the critical extensions to the diffusion model. Detailed derivations and properties of the model are described in the **Supplementary file 1**.

Discretizing time into small steps of size Δ allows us to describe the particle trajectory $x(t)$ in a DM by a random walk, $x(t) = \sum_{n \in 1:t} \delta x_n$, where each of the steps $\delta x_n \sim (k \sin(h) \Delta, \Delta)$, called the momentary evidence, are normally distributed with mean $k \sin(h) \Delta$ and variance Δ ($1:t$ denotes the set of all steps up to time t). This representation is exact in the sense that it recovers the diffusion model, $\dot{x} = k \sin(h) + \eta(t)$, in the limit of $\Delta \rightarrow 0$.

For the standard diffusion model, the posterior probability of $\sin(h)$ after observing the stimulus for t seconds, and under the assumption of a uniform prior, is given by Bayes rule

$$p(\sin(h) | \delta x_{1:t}) \propto \prod_{n \in 1:t} p(\delta x_n | \sin(h)) \propto N\left(\sin(h) \left| \frac{x(t)}{kt}, \frac{1}{k^2 t} \right.\right), \quad (4)$$

where $\delta x_{1:t}$ is the momentary evidence up to time t . From this we can derive the belief that heading is rightward, resulting in

$$p(h > 0 | \delta x_{1:t}) = p(\sin(h) > 0 | \delta x_{1:t}) = \int_0^\pi p(\sin(h) | \delta x_{1:t}) dh = \Phi\left(\frac{x(t)}{\sqrt{t}}\right), \quad (5)$$

where $\Phi(\cdot)$ denotes the standard cumulative Gaussian function. This shows that both the posterior of the actual heading angle, as well as the belief about ‘rightward’ being the correct choice, only depend on $x(t)$ rather than the whole trajectory $\delta x_{1:t}$.

The above formulation assumes that evidence is constant over time, which is not the case for our stimuli. Considering the visual cue and assuming that its associated sensitivity varies with velocity $v(t)$, the momentary evidence $\delta x_{vis,n} \sim N(v_n k_{vis}(c) \sin(h) \Delta, \Delta)$ is Gaussian with mean $v_n k_{vis}(c) \sin(h) \Delta$, where v_n is the velocity at time step n , and variance Δ . Using Bayes rule again to find the posterior of $\sin(h)$, it is easy to shown that $x_{vis}(t)$ is no longer sufficient to determine the posterior distribution. Rather, we need to perform a velocity-weighted accumulation, $X_{vis}(t) = \sum_{n \in 1:t} v_n \delta x_{vis,n}$ to replace $x_{vis}(t)$, and replace time t with $V(t) = \sum_{n \in 1:t} v_n^2 \Delta$, resulting in the following expression for the posterior

$$p(\sin(h) | \delta x_{vis,1:t}) = p(\sin(h) | X_{vis}(t), V(t)) = N\left(\sin(h) \left| \frac{X_{vis}(t)}{k_{vis}(c)V(t)}, \frac{1}{k_{vis}^2(c)V(t)} \right.\right). \quad (6)$$

Consequently, the belief about ‘rightward’ being correct can also be fully expressed by $X_{vis}(t)$ and $V(t)$. This shows that optimal accumulation of evidence with a single-particle diffusion model with time-varying evidence sensitivity requires the momentary evidence to be weighted by its momentary sensitivity. A similar formulation holds for the posterior over heading based on the vestibular cue, however the vestibular cue is assumed to be weighted by the temporal profile of stimulus acceleration, instead of velocity.

When combining multiple cues into a single DM, $\dot{X}_{comb} = d(t)(d(t)k_{comb} \sin(h) + \eta_{comb}(t))$, we aim to find expressions for k_{comb} and $d(t)$ that keep the posterior over $\sin(h)$ unchanged, that is

$$p(\sin(h) | \delta x_{comb,1:t}) = p(\sin(h) | \delta x_{vis,1:t}, \delta x_{vest,1:t}). \quad (7)$$

$\delta x_{comb,1:t}$ is the sequence of momentary evidence in the combined condition, following $\delta x_{comb,n} \sim N(d_n k_{comb}(c) \sin(h) \Delta, \Delta)$. Expanding the probabilities reveals the equality to hold if the combined sensitivity is given by $k_{comb}^2(c) = k_{vis}^2(c) + k_{vest}^2$, and $d(t)$ is expressed by **Equation 3**, leading to **Equation 1** for optimally combining the momentary evidence (see **Supplementary file 1** for derivation).

Model fitting

The model used to fit the behavioral data is described in the main text. We never averaged data across subjects as they feature qualitatively different behavior, due to different speed-accuracy tradeoffs. Furthermore, for subjects performing both the three-coherence and the six-coherence version of the experiment, we treated either version as a separate data set. For each modality/coherence combination (7 combinations for 3 coherences, 13 combinations for 6 coherences) we fitted one bias parameter that prevents behavioral biases from influencing model fits. The fact that performance of subjects often fails to reach 100% correct even for the highest coherences and largest heading angles was modeled by a lapse rate, which describes the frequency with which the subject makes a random choice rather than one based on accumulated evidence. This lapse rate was assumed to be independent of

stimulus modality or coherence, and so a single lapse rate parameter is shared among all modality/coherence combinations.

All model fits sought to find the model parameters ϕ that maximize the likelihood of the observed choices and reaction times for each dataset. As in *Palmer et al. (2005)*, we have assumed the likelihood of the choices to follow a binomial distribution, and the reaction times of correct and incorrect choices to follow different Gaussian distributions centered on the empirical means and spread according to the standard error. Model predictions for choice fractions and reaction times for correct and incorrect choices were computed from the solution to integral equations describing first-passage times of bounded diffusion processes (*Smith, 2000*). See **Supplementary file 1** for the exact form of the likelihood function that was used.

To avoid getting trapped in local maxima of this likelihood, we utilized a three-step maximization procedure. First, we found a (possibly local) maximum by pseudo-gradient ascent on the likelihood function. Starting from this maximum, we used a Markov Chain Monte Carlo procedure to draw 44,000 samples from the parameter posterior under the assumption of a uniform, bounded prior. After this, we used the highest-likelihood sample, which is expected to be close to the mode of this posterior, as a starting point to find the posterior mode by pseudo-gradient ascent. The resulting parameter vector is taken as the maximum-likelihood estimate. All pseudo-gradient ascent maximizations were performed with the Optimization Toolbox of Matlab R2013a (Mathworks), using stringent stopping criteria (TolFun = TolX = 10^{-20}) to prevent premature convergence.

Additional information

Competing interests

DEA: Reviewing editor, *eLife*. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
National Institutes of Health	R01 DC007620	Dora E Angelaki
National Institutes of Health	R01 EY016178	Gregory C DeAngelis
National Science Foundation	BCS0446730	Alexandre Pouget
U.S. Army Research Laboratory	Multidisciplinary University Research Initiative, N00014-07-1-0937	Alexandre Pouget
Air Force Office of Scientific Research	FA9550-10-1-0336	Alexandre Pouget
James S. McDonnell Foundation		Alexandre Pouget

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

JD, Conception and design, Analysis and interpretation of data, Drafting or revising the article; GCDA, AP, Conception and design, Drafting or revising the article; EMK, Acquisition of data, Drafting or revising the article; DEA, Conception and design, Acquisition of data, Drafting or revising the article

Ethics

Human subjects: Informed consent was obtained from all participants and all procedures were reviewed and approved by the Washington University Office of Human Research Protections (OHRP), Institutional Review Board (IRB; IRB ID# 201109183). Consent to publish was not obtained in writing, as it was not required by the IRB, but all subjects were recruited for this purpose and approved verbally. Of the initial seven subjects, three participated in a follow-up experiment roughly 2 years after the initial data collection. Procedures for the follow-up experiment were approved by the Institutional Review Board for Human Subject Research for Baylor College of Medicine and Affiliated Hospitals (BCM IRB, ID# H-29411) and informed consent and consent to publish was given again by all three subjects.

Additional files

Supplementary files

- Supplementary file 1. Detailed model derivation and description.
DOI: [10.7554/eLife.03005.015](https://doi.org/10.7554/eLife.03005.015)
- Supplementary file 2. Outcome of additional statistical hypothesis tests.
DOI: [10.7554/eLife.03005.016](https://doi.org/10.7554/eLife.03005.016)

References

- Battaglia PW**, Jacobs RA, Aslin RN. 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**:1391–1397. doi: [10.1364/JOSAA.20.001391](https://doi.org/10.1364/JOSAA.20.001391).
- Beck JM**, Ma WJ, Kiani R, Hanks T, Churchland AK, Roitman J, Shadlen MN, Latham PE, Pouget A. 2008. Probabilistic population codes for Bayesian decision making. *Neuron* **60**:1142–1152. doi: [10.1016/j.neuron.2008.09.021](https://doi.org/10.1016/j.neuron.2008.09.021).
- Bogacz R**, Brown E, Moehlis J, Holmes P, Cohen JD. 2006. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review* **113**:700–765. doi: [10.1037/0033-295X.113.4.700](https://doi.org/10.1037/0033-295X.113.4.700).
- Britten KH**, van Wezel RJ. 1998. Electrical microstimulation of cortical area MST biases heading perception in monkeys. *Nature Neuroscience* **1**:59–63. doi: [10.1038/259](https://doi.org/10.1038/259).
- Britten KH**, Van Wezel RJ. 2002. Area MST and heading perception in macaque monkeys. *Cerebral Cortex* **12**:692–701. doi: [10.1093/cercor/12.7.692](https://doi.org/10.1093/cercor/12.7.692).
- Chen A**, DeAngelis GC, Angelaki DE. 2011a. A comparison of vestibular spatiotemporal tuning in macaque parietoinsular vestibular cortex, ventral intraparietal area, and medial superior temporal area. *The Journal of Neuroscience* **31**:3082–3094. doi: [10.1523/JNEUROSCI.4476-10.2011](https://doi.org/10.1523/JNEUROSCI.4476-10.2011).
- Chen A**, DeAngelis GC, Angelaki DE. 2011b. Representation of vestibular and visual cues to self-motion in ventral intraparietal cortex. *The Journal of Neuroscience* **31**:12036–12052. doi: [10.1523/JNEUROSCI.0395-11.2011](https://doi.org/10.1523/JNEUROSCI.0395-11.2011).
- Clark JJ**, Yuille AL. 1990. *Data fusion for sensory information processing systems*. Boston: Kluwer Academic.
- Colonus H**, Arndt P. 2001. A two-stage model for visual-auditory interaction in saccadic latencies. *Perception & Psychophysics* **63**:126–147. doi: [10.3758/BF03200508](https://doi.org/10.3758/BF03200508).
- Corneil BD**, Van Wanrooij M, Munoz DP, Van Opstal AJ. 2002. Auditory-visual interactions subserving goal-directed saccades in a complex scene. *Journal of Neurophysiology* **88**:438–454.
- Ernst MO**, Banks MS. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**:429–433. doi: [10.1038/415429a](https://doi.org/10.1038/415429a).
- Fernandez C**, Goldberg JM. 1976. Physiology of peripheral neurons innervating otolith organs of the squirrel monkey. III. Response dynamics. *Journal of Neurophysiology* **39**:996–1008.
- Fetsch CR**, Pouget A, DeAngelis GC, Angelaki DE. 2011. Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience* **15**:146–154. doi: [10.1038/nn.2983](https://doi.org/10.1038/nn.2983).
- Fetsch CR**, Rajguru SM, Karunaratne A, Gu Y, Angelaki DE, Deangelis GC. 2010. Spatiotemporal properties of vestibular responses in area MSTd. *Journal of Neurophysiology* **104**:1506–1522. doi: [10.1152/jn.91247.2008](https://doi.org/10.1152/jn.91247.2008).
- Fetsch CR**, Turner AH, DeAngelis GC, Angelaki DE. 2009. Dynamic reweighting of visual and vestibular cues during self-motion perception. *The Journal of Neuroscience* **29**:15601–15612. doi: [10.1523/JNEUROSCI.2574-09.2009](https://doi.org/10.1523/JNEUROSCI.2574-09.2009).
- Goodman SN**. 1999. Toward evidence-based medical statistics. 2: the Bayes factor. *Annals of Internal Medicine* **130**:1005–1013. doi: [10.7326/0003-4819-130-12-199906150-00019](https://doi.org/10.7326/0003-4819-130-12-199906150-00019).
- Graf AB**, Kohn A, Jazayeri M, Movshon JA. 2011. Decoding the activity of neuronal populations in macaque primary visual cortex. *Nature Neuroscience* **14**:239–245. doi: [10.1038/nn.2733](https://doi.org/10.1038/nn.2733).
- Grice GR**, Canham L, Boroughs JM. 1984. Combination rule for redundant information in reaction time tasks with divided attention. *Perception & Psychophysics* **35**:451–463. doi: [10.3758/BF03203922](https://doi.org/10.3758/BF03203922).
- Gu Y**, DeAngelis GC, Angelaki DE. 2007. A functional link between area MSTd and heading perception based on vestibular signals. *Nature Neuroscience* **10**:1038–1047. doi: [10.1038/nn1935](https://doi.org/10.1038/nn1935).
- Gu Y**, Angelaki DE, Deangelis GC. 2008. Neural correlates of multisensory cue integration in macaque MSTd. *Nature Neuroscience* **11**:1201–1210. doi: [10.1038/nn.2191](https://doi.org/10.1038/nn.2191).
- Gu Y**, Deangelis GC, Angelaki DE. 2012. Causal links between dorsal medial superior temporal area neurons and multisensory heading perception. *The Journal of Neuroscience* **32**:2299–2313. doi: [10.1523/JNEUROSCI.5154-11.2012](https://doi.org/10.1523/JNEUROSCI.5154-11.2012).
- Gu Y**, Fetsch CR, Adeyemo B, Deangelis GC, Angelaki DE. 2010. Decoding of MSTd population activity accounts for variations in the precision of heading perception. *Neuron* **66**:596–609. doi: [10.1016/j.neuron.2010.04.026](https://doi.org/10.1016/j.neuron.2010.04.026).
- Gu Y**, Watkins PV, Angelaki DE, DeAngelis GC. 2006. Visual and nonvisual contributions to three-dimensional heading selectivity in the medial superior temporal area. *The Journal of Neuroscience* **26**:73–85. doi: [10.1523/JNEUROSCI.2356-05.2006](https://doi.org/10.1523/JNEUROSCI.2356-05.2006).
- Heuer HW**, Britten KH. 2007. Linear responses to stochastic motion signals in area MST. *Journal of Neurophysiology* **98**:1115–1124. doi: [10.1152/jn.00083.2007](https://doi.org/10.1152/jn.00083.2007).
- Jeffreys H**. 1998. *Theory of probability*. Oxford: Clarendon Press.

- Kiani R**, Hanks TD, Shadlen MN. 2008. Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *The Journal of Neuroscience* **28**:3017–3029. doi: [10.1523/JNEUROSCI.4761-07.2008](https://doi.org/10.1523/JNEUROSCI.4761-07.2008).
- Knill DC**, Saunders JA. 2003. Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research* **43**:2539–2558. doi: [10.1016/S0042-6989\(03\)00458-9](https://doi.org/10.1016/S0042-6989(03)00458-9).
- Laming DRJ**. 1968. *Information theory of choice-reaction times*. London: Academic Press.
- Lisberger SG**, Movshon JA. 1999. Visual motion analysis for pursuit eye movements in area MT of macaque monkeys. *The Journal of Neuroscience* **19**:2224–2246.
- Ma WJ**, Beck JM, Latham PE, Pouget A. 2006. Bayesian inference with probabilistic population codes. *Nature Neuroscience* **9**:1432–1438. doi: [10.1038/nn1790](https://doi.org/10.1038/nn1790).
- Mazurek ME**, Roitman JD, Ditterich J, Shadlen MN. 2003. A role for neural integrators in perceptual decision making. *Cerebral Cortex* **13**:1257–1269. doi: [10.1093/cercor/bhg097](https://doi.org/10.1093/cercor/bhg097).
- Miller J**. 1982. Divided attention: evidence for coactivation with redundant signals. *Cognitive Psychology* **14**:247–279. doi: [10.1016/0010-0285\(82\)90010-X](https://doi.org/10.1016/0010-0285(82)90010-X).
- Morgan ML**, Deangelis GC, Angelaki DE. 2008. Multisensory integration in macaque visual cortex depends on cue reliability. *Neuron* **59**:662–673. doi: [10.1016/j.neuron.2008.06.024](https://doi.org/10.1016/j.neuron.2008.06.024).
- Onken A**, Drugowitsch J, Kanitscheider I, DeAngelis GC, Beck JM, Pouget A. 2012. Near optimal multisensory integration with nonlinear probabilistic population codes using divisive normalization. *The Society for Neuroscience annual meeting 2012*.
- Otto TU**, Mamassian P. 2012. Noise and correlations in parallel perceptual decision making. *Current Biology* **22**:1391–1396. doi: [10.1016/j.cub.2012.05.031](https://doi.org/10.1016/j.cub.2012.05.031).
- Palmer J**, Huk AC, Shadlen MN. 2005. The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision* **5**:376–404. doi: [10.1167/5.5.1](https://doi.org/10.1167/5.5.1).
- Papoulis A**. 1991. *Probability, random variables, and stochastic processes*. New York, London: McGraw-Hill.
- Price NS**, Ono S, Mustari MJ, Ibbotson MR. 2005. Comparing acceleration and speed tuning in macaque MT: physiology and modeling. *Journal of Neurophysiology* **94**:3451–3464. doi: [10.1152/jn.00564.2005](https://doi.org/10.1152/jn.00564.2005).
- Raab DH**. 1962. Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences* **24**:574–590. doi: [10.1111/j.2164-0947.1962.tb01433.x](https://doi.org/10.1111/j.2164-0947.1962.tb01433.x).
- Ratcliff R**. 1978. Theory of memory retrieval. *Psychological Review* **85**:59–108. doi: [10.1037/0033-295X.85.2.59](https://doi.org/10.1037/0033-295X.85.2.59).
- Ratcliff R**, Smith PL. 2004. A comparison of sequential sampling models for two-choice reaction time. *Psychological Review* **111**:333–367. doi: [10.1037/0033-295X.111.2.333](https://doi.org/10.1037/0033-295X.111.2.333).
- Schlack A**, Sterbing-D'Angelo SJ, Hartung K, Hoffmann KP, Bremmer F. 2005. Multisensory space representations in the macaque ventral intraparietal area. *The Journal of Neuroscience* **25**:4616–4625. doi: [10.1523/JNEUROSCI.0455-05.2005](https://doi.org/10.1523/JNEUROSCI.0455-05.2005).
- Smith PL**. 2000. Stochastic dynamic models of response time and accuracy: a foundational primer. *Journal of Mathematical Psychology* **44**:408–463. doi: [10.1006/jmps.1999.1260](https://doi.org/10.1006/jmps.1999.1260).
- Stanford TR**, Shankar S, Massoglia DP, Costello MG, Salinas E. 2010. Perceptual decision making in less than 30 milliseconds. *Nature Neuroscience* **13**:379–385. doi: [10.1038/nn.2485](https://doi.org/10.1038/nn.2485).
- Stephan KE**, Penny WD, Daunizeau J, Moran RJ, Friston KJ. 2009. Bayesian model selection for group studies. *NeuroImage* **46**:1004–1017. doi: [10.1016/j.neuroimage.2009.03.025](https://doi.org/10.1016/j.neuroimage.2009.03.025).
- Tolhurst DJ**, Movshon JA, Dean AF. 1983. The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Research* **23**:775–785. doi: [10.1016/0042-6989\(83\)90200-6](https://doi.org/10.1016/0042-6989(83)90200-6).
- Townsend JT**, Wenger MJ. 2004. A theory of interactive parallel processing: new capacity measures and predictions for a response time inequality series. *Psychological Review* **111**:1003–1035. doi: [10.1037/0033-295X.111.4.1003](https://doi.org/10.1037/0033-295X.111.4.1003).
- van Beers RJ**, Sittig AC, Denier van der Gon JJ. 1996. How humans combine simultaneous proprioceptive and visual position information. *Experimental Brain Research* **111**:253–261. doi: [10.1016/S0079-6123\(08\)60413-6](https://doi.org/10.1016/S0079-6123(08)60413-6).
- Whitchurch EA**, Takahashi TT. 2006. Combined auditory and visual stimuli facilitate head saccades in the barn owl (*Tyto alba*). *Journal of Neurophysiology* **96**:730–745. doi: [10.1152/jn.00072.2006](https://doi.org/10.1152/jn.00072.2006).
- Wichmann FA**, Hill NJ. 2001. The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics* **63**:1314–1329. doi: [10.3758/BF03194545](https://doi.org/10.3758/BF03194545).

Optimal multisensory decision-making in a reaction-time task

Detailed model derivation and description

Jan Drugowitsch, Gregory C. DeAngelis, Eliana M. Klier,
Dora E. Angelaki, Alexandre Pouget

Optimal Accumulation of Evidence across Time and Cues

The subjects are seated on a motion platform in front of a display screen. They receive information about self-motion direction either through vestibular cues (platform motion), visual cues (visual flow field), or through both cues in combination. They perform a heading discrimination task, and are instructed to indicate as quickly and as accurately as possible if they are moving rightward or leftward relative to straight ahead. Below, we develop – based on a combination of Bayesian inference and diffusion models – an ideal observer model for this task.

Denote h as the heading direction, with $h = 0$ being straight ahead, and $h > 0$ being rightward motion relative to straight ahead. Each sensory cue provides, at each point in time, noisy momentary evidence about this heading direction. The task of the decision maker is to accumulate this evidence over time and cues, and, after some time, decide whether $h \geq 0$ or $h < 0$ based on the belief:

$$p(h \geq 0 | \text{all momentary evidence}). \quad (1)$$

Below, we first describe for each cue how we assume this momentary evidence to encode h , and then we show how to compute the above belief for various, increasingly realistic cases. We begin by assuming a single sensory cue that provides evidence whose reliability is constant over time. This leads to a formulation similar to diffusion models (Ratcliff 1978; Bogacz, Brown et al. 2006), and we utilize this relationship to model the

speed/accuracy trade-off. We then show how to extend this formulation to cases in which the reliability of the momentary evidence changes with time, as is the case in both of our unimodal conditions. After that, we derive how a single diffusion model can optimally accumulate evidence across cues, even if the reliability of the evidence differs between the cues. We will again initially assume the reliability of the evidence to be constant in time, and then deal with the general case of time-varying cue reliability in which the time-course differs between the two cues. The last case describes the ideal observer/actor model for the multi-modal condition of our task.

In what follows, the subscripts \cdot_{vis} , \cdot_{vest} , and \cdot_{comb} refer to the visual-only, vestibular-only, and combined conditions, respectively.

Momentary Evidence

We assume that, for either modality, information about heading direction is encoded by a noisy sensory signal, called the *momentary evidence* and denoted by \dot{x} , according to

$$\dot{x} = b(t)k \sin(h) + \eta(t), \quad (2)$$

where $b(t)k$ is the *sensitivity* that determines the strength with which h influences \dot{x} , and $\eta(t)$ is a unit variance Gaussian white noise process. The only information about heading relevant to our task is the horizontal projection onto a line that is orthogonal to straight ahead. For this reason, heading h influences the momentary evidence only through this projection, as given by $\sin(h)$.

The subject's sensitivity to the evidence is characterized by the terms k and $b(t)$, where k determines how effectively each subject can make use of the incoming information, and $b(t)$ encodes how the reliability of this information changes over time. Thus, the sensitivity $b(t)k$ is a composite measure that takes into account both how reliably the momentary evidence \dot{x} reveals the heading direction (determined by experimenter), as well as how effective the subject is in using this information (property of subject). Acceleration, $a(t)$, is the physical quantity that modulates the reliability of the inertial motion (i.e, vestibular) cues, such that the subject's sensitivity follows the same time course, $b_{vest}(t) = a(t)$. We assume that, at least for the slow speeds used in the experiment, the reliability of the visual cue is mainly determined by motion velocity,

$v(t)$, such that $b_{vis}(t) = v(t)$ (Lisberger and Movshon 1999; Price, Ono et al. 2005; Schlack, Krekelberg et al. 2007).

For further development, we consider a time-discretized version of the momentary evidence, δx_n , that is related to \dot{x} by $\delta x_n = \int_{n\Delta}^{(n+1)\Delta} \dot{x}(t)dt$, where Δ denotes the short time periods into which time is discretized. Given that Δ is sufficiently small, we can assume the reliability time-course $b_n \approx b(n\Delta)$ to be constant for each n . Then, $\delta x_n \sim N(kb_n \sin(h)\Delta, \Delta)$ is distributed according to a normal distribution with mean $kb_n \sin(h)\Delta$ and variance Δ . As $\Delta \rightarrow 0$, the use of either x or δx_n becomes equivalent. Thus, all of the following is valid for either definition. The amount of information that δx_n provides for the discrimination of $\sin(h)$, as measured by the Fisher Information, is given by $I_{\delta x_n}(\sin(h)) = k^2 b_n^2 \Delta$. This confirms kb_n as measure of sensitivity to changes in the momentary evidence.

Accumulating Evidence with Constant Reliability over Time

Assume that the subject observes the stimulus for T seconds and wants to decide between $h < 0$ and $h \geq 0$ based on all momentary evidence observed up to that point. For now we assume the reliability of the evidence to be constant over time, such that $\forall t \geq 0: b(t) = 1$ and the momentary evidence becomes $\delta x_n \sim N(k \sin(h)\Delta, \Delta)$.

Given that $h \geq 0$ corresponds to $\sin(h) \geq 0$ over the range of headings of interest, we want to find the belief $p(\sin(h) > 0 | \delta x_{1:N})$, where $N \approx T/\Delta$ and $\delta x_{1:N} = \{\delta x_1, \dots, \delta x_N\}$ denotes all momentary evidence up to time $T \approx \Delta N$. In order to do so, we first compute the posterior $\sin(h)$ by Bayes' rule, resulting in

$$\begin{aligned}
p(\sin(h) | \delta x_{1:N}) &\propto \prod_{n=1}^N p(\delta x_n | \sin(h)) \\
&= \prod_{n=1}^N \mathcal{N}(\delta x_n | k \sin(h) \Delta, \Delta) \\
&\propto e^{\sin(h) k \sum_n \delta x_n - \frac{1}{2} \sin(h)^2 k^2 \sum_n \Delta} \\
&\approx e^{\sin(h) k x(T) - \frac{1}{2} \sin(h)^2 k^2 T} \\
&\propto \mathcal{N}\left(\sin(h) | \frac{x(T)}{kT}, \frac{1}{k^2 T}\right),
\end{aligned} \tag{3}$$

where we have assumed a uniform prior, $p(\sin(h)) \propto 1$ over $\sin(h) \in [-1, 1]$, and have used $\sum_n \delta x_n \approx \int_0^T \dot{x}(t) dt = x(T)$ and $\sum_n \Delta \approx T$. Furthermore, we have assumed that $|\sin(h)|$ is small (as is the case in our experiment, with $|h| \leq 16^\circ$ and thus $|\sin(h)| < 0.276$), such that we can approximate the posterior by a Gaussian despite the restriction that $\sin(h) \in [-1, 1]$. The above shows that the posterior only depends on the sufficient statistics $x(T)$ and T , rather than the whole sequence $\delta x_{1:N}$ of observations. Furthermore, the posterior variance decreases monotonically with T , as more evidence provides us with a more certain estimate.

From this posterior we find the belief of $\sin(h) \geq 0$ by

$$p(\sin(h) \geq 0 | x(T), T) = \int_0^1 p(\sin(h) = y | x(T), T) dy \approx \Phi\left(\frac{x(T)}{kT} \sqrt{k^2 T}\right) = \Phi\left(\frac{x(T)}{\sqrt{T}}\right), \tag{4}$$

where we again have assumed the posterior to be well approximated by a Gaussian that has negligible mass outside the range $[-1, 1]$, and $\Phi(\alpha) = \int_{-\infty}^{\alpha} \mathcal{N}(\beta | 0, 1) d\beta$ is the standard cumulative Gaussian, with $\Phi(\alpha) \geq \frac{1}{2}$ if $\alpha \geq 0$ and $\Phi(\alpha) < \frac{1}{2}$ otherwise. Consequently, the decision maker ought to decide in favor of $h \geq 0$ if $x(T) \geq 0$, and $h < 0$ otherwise (Drugowitsch, Moreno-Bote et al. 2012).

The above describes the optimal decision strategy given all observed momentary evidence up to some time T . However, it does not provide us with a way to choose at which time enough evidence has been collected to commit to a decision. We achieve the latter by linking this decision strategy to standard diffusion models in the next section.

Relation to Standard Diffusion Models

The framework described above can be cast as a standard diffusion model, in which accumulated momentary evidence is represented by the position of a drifting and diffusing “particle”, $x(t)$. Decisions are triggered by bounding the particle space from below at $-\theta$ and from above by θ . As soon as either of these bounds is reached, the decision maker ought to commit to the corresponding decision. Thus, diffusion models provide us with a strategy for choosing at which time to commit to a decision.

Furthermore, the analysis in the previous section shows that these decisions take into account all momentary evidence up until the point of the decision and are therefore Bayes-optimal (Laming 1968; Gold and Shadlen 2002; Bogacz, Brown et al. 2006).

From the perspective of diffusion models it seems as if the only decisive factor to commit to a decision is the particle location $x(t)$. In contrast, our Bayesian analysis above seems to additionally require information about time, t , to compute the relevant probabilities. The omission of time in diffusion models stems from partitioning the belief space into $p(\sin(h) \geq 0 | x(T), T) \geq \frac{1}{2}$ and $p(\sin(h) \geq 0 | x(T), T) < \frac{1}{2}$ while neglecting the actual magnitude of this belief. This magnitude, however, is informative about the certainty at which this decision is made. Consider, for example, that the decision maker chooses $h \geq 0$ at time t , corresponding to $x(t) = \theta$. Then, the belief of making a correct decision is given by

$$p(\sin(h) \geq 0 | x(t) = \theta, t) = \Phi\left(\frac{\theta}{\sqrt{t}}\right) \geq \frac{1}{2}, \quad (5)$$

which is a decreasing function of time. This expression follows directly from Eq. (4) despite the presence of a bound, as we (implicitly) condition on having reached the bound for the first time at time t , such that we do not need to consider the possibility of having crossed this bound before (Drugowitsch, Moreno-Bote et al. 2012). This demonstrates that, even with a constant bound in particle space, diffusion models make decisions at different levels of confidence (Kiani and Shadlen 2009). In particular, early decisions will be of high confidence, while late decisions are made at a low level of confidence. Thus, this constant bound in particle space corresponds to a collapsing bound in belief space.

Qualitatively, such a strategy has been shown to perform optimal decision-making in the sense of maximizing the reward rate (Drugowitsch, Moreno-Bote et al. 2012).

When defining the momentary evidence about heading, we have followed standard diffusion model conventions and have assumed a unit diffusion variance. We will show that this is not a restriction, as for any diffusion model with a non-unit variance we can find a diffusion model with unit variance that features exactly the same behavior. Therefore, we can assume unit variance without a loss of generality. In particular, assume a diffusion variance σ^2 , such that the momentary evidence becomes $\dot{x}_\sigma = b(t)k_\sigma \sin(h) + \sigma\eta(t)$. This evidence relates to the unit-variance evidence by $\dot{x} = \dot{x}_\sigma / \sigma$ with $k = k_\sigma / \sigma$. The same relationship $x(t) = x_\sigma(t) / \sigma$ holds between the particle locations (i.e. the accumulated momentary evidence). This shows that assuming a non-unit variance is equivalent to a re-scaling of the particle space. We can compensate for this re-scaling by re-scaling the bounds, leading to a diffusion model with exactly the same behavior as one with a unit diffusion variance. We will use this property later, to find the minimal parameterization of our model while assuming that the diffusion variance is modulated by visual motion coherence.

Accumulating Evidence with Time-Varying Reliability

In this section we will show that, as soon as the reliability of the evidence changes over time, particle location and time are no longer sufficient statistics. Instead, we need to take the changing reliability of the cues into account when accumulating the momentary evidence. This will lead to a re-definition of the particle location in diffusion models that allows us to make Bayes-optimal decisions even with time-varying reliability of the momentary evidence.

As before, we are interested in computing the belief $p(\sin(h) \geq 0 | \delta x_{1:N})$.

However, now we assume the sensitivity k to be weighted by the time-varying function $b(t)$, such that the momentary evidence is given by $\delta x_n \sim N(kb_n \sin(h)\Delta, \Delta)$. With this evidence, the posterior of $\sin(h)$ results in

$$\begin{aligned}
p(\sin(h) | \delta x_{1:T}) &\propto \prod_{n=1}^N N(\delta x_n | kb_n \sin(h)\Delta, \Delta) \\
&\propto e^{\sin(h)k \sum_n b_n \delta x_n - \frac{1}{2} \sin(h)^2 k^2 \sum_n b_n^2 \Delta} \\
&\approx e^{\sin(h)kX(T) - \frac{1}{2} \sin(h)^2 k^2 B(T)} \\
&\propto N\left(\sin(h) | \frac{X(T)}{kB(T)}, \frac{1}{k^2 B(T)}\right),
\end{aligned} \tag{6}$$

where we have defined

$$X(T) = \int_0^T b(t) \dot{x}(t) dt \approx \sum_n b_n \delta x_n, \quad B(T) = \int_0^T b(t)^2 dt \approx \sum_n b_n^2 \Delta. \tag{7}$$

Thus, $X(T)$ is the accumulated momentary evidence, weighted at each point in time by the sensitivity time course. $B(T)$ is the squared accumulated sensitivity time course, which we will call the *power* of the evidence. Comparing Eq. (3) to Eq. (6) shows that $X(T)$ replaces $x(T)$ as the particle location, and $B(T)$ becomes the new passed time, replacing T . Using $b(t) = 1$ for all t causes $B(T) = T$ and $X(T) = x(T)$ and recovers the original formulation. While it might seem that a negative $b(t)$ (for example, in the case of acceleration) causes the momentary evidence in Eq. (7) to be weighted negatively, this is in fact not the case, as $\dot{x}(t)$ is already scaled by $b(t)$ according to Eq. (2). Thus, if we replace $\dot{x}(t)$ in Eq. (7) by Eq. (2), $b(t)$ will be squared, causing its effective influence on the momentary evidence to be always non-negative.

With the above posterior, the belief becomes

$$p(\sin(h) \geq 0 | X(T), B(T)) = \Phi\left(\frac{X(T)}{\sqrt{B(T)}}\right). \tag{8}$$

Therefore, the sign of $X(T)$ now determines the decision. This confirms that $X(t)$ takes the role of the particle location in a Bayes-optimal diffusion model that triggers a decision as soon as either of the bounds is reached.

The above derivation shows that diffusion models that use the un-weighted $x(t)$ as their particle location become sub-optimal as soon as the reliability of the evidence changes with time. Consider, for example, a task in which $b(t) = 0$ for $0 \leq t \leq T$, and $b(t) = 1$ for $t > T$, such that for all $t \leq T$, the momentary information contains only noise and no information about h . If we were to use $x(t)$, we would initially only accumulate

noise while treating it as evidence, which is clearly sub-optimal. $X(t)$ avoids this problem by giving zero weight to all evidence up until T , and only starts accumulating evidence thereafter. This principle finds its parallel in the standard cue combination literature, where it is known that cues ought to be weighted according to their reliability (Clark and Yuille 1990). If one of the cues has a very low reliability, it does not contribute to the decision. The same applies here, but rather than accumulating evidence across cues, we accumulate across time.

Using $X(T)$ instead of $x(T)$ and $B(T)$ instead of T requires a re-interpretation of what a time-invariant bound on $X(T)$ means. From Eq. (8) we can see that the decision confidence at the bound (where $X(T) = \theta$) drops monotonically with $B(T)$. Thus, a constant bound on the particle location still implies a collapsing bound on belief (as $B(T)$ is monotonically increasing in time), but that the latter drops with $B(T)$ rather than with T . Thus, the rate at which this bound drops now depends on the reliability of the momentary evidence. This completes the description of the Bayes-optimal decision making model for the two unimodal conditions.

Accumulating Evidence across Time and Cues, with Constant Reliability

We now describe how to accumulate momentary evidence if information about heading direction is available from multiple cues. For now, we assume the reliability of these cues to be constant over time. In the next section, we discuss the changes required when this is not the case.

The visual and vestibular cues to heading provide momentary evidence given by $\delta x_{vis,n} \sim N(k_{vis} \sin(h)\Delta, \Delta)$ and $\delta x_{vest,n} \sim N(k_{vest} \sin(h)\Delta, \Delta)$. Here, the sensitivities k_{vis} and k_{vest} are again composite measures of how much information the momentary evidence provides about the heading, and how effective subjects are in utilizing this information. Given $\delta x_{vis,1:N}$ and $\delta x_{vest,1:N}$ up until time $T \approx N\Delta$, we find the posterior over $\sin(h)$ by Bayes rule, resulting in

$$\begin{aligned}
p(\sin(h) | \delta x_{vis,1:N}, \delta x_{vest,1:N}) &\propto \prod_n N(\delta x_{vis,n} | k_{vis} \sin(h)\Delta, \Delta) N(\delta x_{vest,n} | k_{vest} \sin(h)\Delta, \Delta) \\
&\propto e^{\sin(h)(k_{vis} \sum_n \delta x_{vis,n} + k_{vest} \sum_n \delta x_{vest,n}) - \frac{1}{2} \sin(h)^2 (k_{vis}^2 + k_{vest}^2) \sum_n \Delta} \\
&\approx e^{\sin(h)(k_{vis} x_{vis}(T) + k_{vest} x_{vest}(T)) - \frac{1}{2} \sin(h)^2 (k_{vis}^2 + k_{vest}^2) T} \\
&\propto N\left(\sin(h) | \frac{x_{comb}(T)}{k_{comb} T}, \frac{1}{k_{comb}^2 T}\right),
\end{aligned} \tag{9}$$

where we have used $\sum_n \delta x_{vis,n} \approx x_{vis}(T)$, $\sum_n \delta x_{vest,n} \approx x_{vest}(T)$, and $\sum_n \Delta \approx T$, and have defined

$$k_{comb}^2 = k_{vis}^2 + k_{vest}^2, \quad x_{comb}(T) = \frac{k_{vis}}{k_{comb}} x_{vis}(T) + \frac{k_{vest}}{k_{comb}} x_{vest}(T). \tag{10}$$

This shows that the sensitivity, k_{comb} , based on both cues is at least as large as that of the more reliable cue, such that $k_{comb} \geq \max\{k_{vis}, k_{vest}\}$. Furthermore, the combined particle location is a weighted sum of the particle locations for the two cues, with weights proportional to the sensitivity to either cue. The decision maker's belief regarding $\sin(h) \geq 0$ is again given by Eq. (4), with $x(T)$ replaced by $x_{comb}(T)$.

Accumulating Evidence across Time and Cues, with Time-Varying Reliability

In our task, the reliability of both cues varies over time. Furthermore, the time-course of variations in reliability differs between the two cues. As previously described, we assume the momentary evidence of the visual modality to be $\delta x_{vis,n} \sim N(k_{vis} v_n \sin(h)\Delta, \Delta)$, and that of the vestibular modality to be $\delta x_{vest,n} \sim N(k_{vest} a_n \sin(h)\Delta, \Delta)$, where v_n and a_n denote stimulus velocity and acceleration, respectively. Making use of momentary evidence $\delta x_{vis,1:N}$ and $\delta x_{vest,1:N}$ until time $T \approx N\Delta$, the posterior over $\sin(h)$ becomes

$$\begin{aligned}
p(\sin(h) | \delta x_{vis,1:N}, \delta x_{vest,1:N}) &\propto e^{\sin(h)(k_{vis} \sum_n v_n \delta x_{vis,n} + k_{vest} \sum_n a_n \delta x_{vest,n}) - \frac{1}{2} \sin(h)^2 (k_{vis}^2 \sum_n v_n^2 \Delta + k_{vest}^2 \sum_n a_n^2 \Delta)} \\
&\approx e^{\sin(h)(k_{vis} X_{vis}(T) + k_{vest} X_{vest}(T)) - \frac{1}{2} \sin(h)^2 (k_{vis}^2 V(T) + k_{vest}^2 A(T))} \\
&\propto N\left(\sin(h) | \frac{X_{comb}(T)}{k_{comb} D(T)}, \frac{1}{k_{comb}^2 D(T)}\right).
\end{aligned} \tag{11}$$

To derive the above, we have, as in Eq. (7), defined the sensitivity-weighted accumulated momentary evidence for each of the cues,

$$X_{vis}(T) = \int_0^T v(t) \dot{x}_{vis}(t) dt \approx \sum_n v_n \delta x_{vis,n}, \quad X_{vest}(T) = \int_0^T a(t) x_{vest}(t) dt \approx \sum_n a_n \delta x_{vest,n}, \quad (12)$$

and the accumulated power of the evidence for each cue,

$$V(T) = \int_0^T v(t)^2 dt \approx \sum_n v_n^2 \Delta, \quad A(T) = \int_0^T a(t)^2 dt \approx \sum_n a_n^2 \Delta. \quad (13)$$

Furthermore, we have left the definition of k_{comb} unchanged (see Eq. (10)), such that the total power of the evidence is given by

$$D(T) = \frac{k_{vis}^2}{k_{comb}^2} V(T) + \frac{k_{vest}^2}{k_{comb}^2} A(T). \quad (14)$$

As a consequence, the particle location for the combined diffusion model is, similar to Eq. (10), given by

$$X_{comb}(T) = \frac{k_{vis}}{k_{comb}} X_{vis}(T) + \frac{k_{vest}}{k_{comb}} X_{vest}(T), \quad (15)$$

This shows that, even if we have multiple sources of evidence whose reliability varies independently over time, we can express the process of accumulating evidence in a single diffusion model, defined by particle location $X_{comb}(t)$, with $D(t)$ being the quantity that represents the passage of time.

Based on the above formulation, we can derive how the momentary evidence of the combined diffusion model is constructed from the momentary evidence provided by the two cues. If we use $D(t) = \int_0^t d(s)^2 ds$, and replace $V(t)$ and $A(t)$ in Eq. (14) by their respective definitions in Eq. (13), we find the momentary power of the combined evidence to be given by

$$d(t)^2 = \frac{k_{vis}^2}{k_{comb}^2} v(t)^2 + \frac{k_{vest}^2}{k_{comb}^2} a(t)^2, \quad (16)$$

which is the sensitivity-weighted average of the momentary powers of the two cues. With the above, Eq. (15) shows that the combined momentary evidence, $\dot{x}_{comb}(t)$, as used in

$$X_{comb}(t) = \int_0^t d(s) \dot{x}_{comb}(s) ds, \text{ is composed of}$$

$$\dot{x}_{comb}(t) = \frac{k_{vis}v(t)}{k_{comb}d(t)}\dot{x}_{vis}(t) + \frac{k_{vest}a(t)}{k_{comb}d(t)}\dot{x}_{vest}(t), \quad (17)$$

that is, the sensitivity-weighted sum of the momentary evidence of each of the cues. With these two quantities, the momentary evidence in the combined diffusion model is given by $\dot{x}_{comb} = k_{comb}d(t)\sin(h) + \eta(t)$. This is easily verified by replacing $\dot{x}_{vis}(t)$ and $\dot{x}_{vest}(t)$ in Eq. (17) by their definitions, which leads to the above expression.

To obtain the belief regarding $\sin(h) \geq 0$, we again take the integral of the posterior $\sin(h)$ over $\sin(h) \geq 0$, which gives

$$p(\sin(h) \geq 0 | X_{comb}(T), D(T)) = \Phi\left(\frac{X_{comb}(T)}{\sqrt{D(T)}}\right). \quad (18)$$

Thus, as before, belief depends on particle location $X_{comb}(T)$ and accumulated power $D(T)$, and the decision itself is solely determined by the sign of $X_{comb}(T)$. As a consequence, we can again perform Bayes-optimal decision making by assuming bounds on $X_{comb}(t)$ at $-\theta$ and θ , and decide in favor of $\sin(h) \geq 0$ (or $\sin(h) < 0$) when the particle reaches the upper (or lower) bound. This completes the description of the Bayes-optimal decision model for the combined condition.

Optimality of evidence accumulation

The above describe how to perform Bayes-optimal decision making in different scenarios. Bayes-optimal here means that the posterior upon which the decision is based contains the information of all momentary evidence observed from either cue. This is easily demonstrated in terms of preservation of information. Formally, if $I_{\delta_x}(\sin(h))$ denotes the Fisher information that δ_x provides about changes in $\sin(h)$, then the Fisher information in the posterior (e.g. Eq. (11)) is the sum of the Fisher information of all momentary evidence across both time and cues, that is

$$\begin{aligned}
I_{X_{comb}(T)}(\sin(h)) &= k_{comb}^2 D(T) \\
&= k_{vis}^2 V(T) + k_{comb}^2 A(T) \\
&= \sum_{n=0}^N (k_{vis}^2 v_n^2 \Delta + k_{vest}^2 a_n^2 \Delta) \\
&= \sum_{n=0}^N (I_{\delta_{x_{vis},n}}(\sin(h)) + I_{\delta_{x_{vest},n}}(\sin(h))).
\end{aligned} \tag{19}$$

In the above, the first line follows from re-expressing the posterior Eq. (11) in terms of $X_{comb}(T)$ (effectively turning it back into a likelihood) and computing its Fisher Information, the second line is based on substituting Eqs. (10) and (14) for k_{comb}^2 and $D(T)$, the third line utilizes the definitions of $A(T)$ and $V(T)$ in Eq. (13), and the last line uses the expression for Fisher Information of the momentary evidence, as discussed in the Section Momentary Evidence.

To relate this kind of optimality to that of standard diffusion models, let us consider how it compares to that of the Sequential Probability Ratio Test (SPRT, (Wald 1947; Wald and Wolfowitz 1948)), of which this diffusion model is a continuous-time implementation (Bogacz, Brown et al. 2006). The SPRT assumes that the likelihood function associated with either option (e.g. “left” and “right”) is known and time-invariant, and consists of accumulating the log-likelihood ratios of the momentary evidences to form a log-posterior. Once this log-posterior reaches a lower or upper time-invariant threshold (in units of log-odds, and thus belief/error rate), the more likely option is chosen. This procedure has been shown to be optimal in at least two senses (Wald and Wolfowitz 1948; Bogacz, Brown et al. 2006). First, if the assumptions of the SPRT are satisfied, it performs Bayes-optimal accumulation of evidence. Second, of all fixed or sequential sample tests that feature the same or lower error rate as the SPRT, the SPRT is the procedure that leads to the fastest decisions, on average.

Our decision procedure features the same optimality guarantees as the SPRT in the first sense that it is Bayes-optimal. However, it follows different underlying assumptions about the momentary evidence and how decisions are made, such that the second form of optimality relating to the speed of the decisions is not applicable. In particular, and in contrast to the SPRT, the likelihood function we utilize contains a nuisance parameter (the heading magnitude $|h|$) that we integrate out to find the belief

(e.g., in the simplest case Eq. (4)). A consequence is that the belief at decision time, and thus the error rate, is, unlike in the SPRT, not time-invariant (see Eq. (5)). As this error rate is a function of the decision time, it cannot be fixed, such that we cannot compare the decision speed of our procedure to that of others with the same error rate. Alternatively, we could have attempted to operate with the error rate averaged over decision times. However, given that there exists no analytical expression for the decision time distribution and that the problem cannot be addressed by the same means as the SPRT (using Wald's Martingale), this approach is unlikely to yield analytical statements. For a recent related attempt that uses numerical rather than analytical means, see (Drugowitsch, Moreno-Bote et al. 2012).

The model's psychometric function and discrimination threshold

To provide better insight into how the different model components contribute to its performance, we derive the model's psychometric function and discuss how it relates to the heading discrimination threshold. Considering for now the unimodal case with time-varying reliability, the psychometric function is formed by plotting how the fraction of choosing one of the two options changes as a function of heading, h (or, equivalently, its sine, $\sin(h)$). With respect to diffusion models, this fraction is the probability

$p(X(T) = \theta \mid \sin(h) = H, T, X(T) = \pm\theta)$ for some heading H , and at some decision time T at which a boundary has been reached (i.e. $X(T) = \pm\theta$), that the particle has reached the upper boundary, $X(T) = \theta$.

We find this probability by first relating it to the model's posterior, conditional on the heading magnitude $|h|$. Specifically, due to the symmetric nature of our task (both responses are a-priori equally likely correct) and the symmetry of our model (inverting the sign of both evidence and responses leaves the behavior unchanged), it can be shown (see Eq. (17) in (Drugowitsch, Moreno-Bote et al. 2012)) that the model's choice probability equals its posterior belief, that is

$$p(X(T) = \theta \mid \sin(h) = H, T, X(T) = \pm\theta) = p(\sin(h) = H \mid X(T) = \theta, T, \sin(h) = \pm H). \quad (20)$$

In the above, the term on the right-hand side is the model's posterior that, given that the upper bound was reached at time T , the heading magnitude was $\sin(h) = H$ rather than $\sin(h) = -H$. This posterior is found by restricting Eq. (6) to these two cases, resulting in

$$\begin{aligned} p(\sin(h) = H \mid X(T) = \theta, T, \sin(h) = \pm H) &\propto e^{Hk\theta - \frac{1}{2}H^2k^2B(T)}, \\ p(\sin(h) = -H \mid X(T) = \theta, T, \sin(h) = \pm H) &\propto e^{-Hk\theta - \frac{1}{2}H^2k^2B(T)}. \end{aligned} \quad (21)$$

Using Eq. (20) and the fact that the probabilities in Eq. (21) sum to 1, we find after some cancellation of terms that the psychometric function is given by the logistic sigmoid,

$$p(X(T) = \theta \mid \sin(h) = H, X(T) = \pm \theta) = \frac{1}{1 + e^{-2kH\theta}}. \quad (22)$$

Note that the above is increasing in k , H , and θ , as one would intuitively expect. Furthermore, it is independent of decision time T . Thus, as k and θ determine the slope of this function around $H = 0$, the psychometric curve's steepness around this point grows with both the subject's sensitivity and the height of the bound. As we have shown the multimodal case to be reducible to a single diffusion model, it features the same psychometric function. A similar psychometric curve for diffusion models with constant reliability over time can be derived by the use of Wald's Martingale (see, for example, (Shadlen, Hanks et al. 2006) for a derivation). This derivation can be generalized to the time-variant reliability case, leading again to Eq. (22).

To relate the model's psychometric curve to the heading discrimination threshold, we assume that the latter is determined by fitting a cumulative Gaussian $\Phi\left(\frac{H}{\sigma}\right)$ with discrimination threshold σ to this psychometric function. Furthermore, we note that the logistic sigmoid $(1 + \exp(-\beta H))^{-1}$ and the above cumulative Gaussian are cumulative distribution functions of the zero-mean Logistic distribution $\text{Logistic}(0, \beta^{-1})$ with scale β^{-1} and the zero-mean Gaussian $N(0, \sigma^2)$ with variance σ^2 , respectively, evaluated at H . These two functions are closely matched by equating the variances of their underlying random variables, which results in $\sigma \approx \frac{\pi}{\sqrt{3}\beta}$. Thus, if we use $\beta = 2k\theta$ from Eq. (22), we find that the discrimination threshold resulting from the extended diffusion model is approximately

$$\sigma \approx \frac{\pi}{\sqrt{12k\theta}}, \quad (23)$$

that is, it is inversely proportional to the sensitivity k and the bound height θ .

Model Parameterization and Fitting

In the previous sections we have described the Bayes-optimal decision model for both the combined and the unimodal conditions. Here, we show how we have fitted these models to behavioral data obtained from human subjects. We first describe the model parametrization and then describe how we have found the parameters, separately for each subject, that best explain the behavior. Finally, we describe a set of alternative, sub-optimal models that we have proposed in the main text as alternative hypotheses of how the observed behavior was generated.

Model Parameterization

The reliability of the visual cue was controlled by the percentage of dots that moved according to the current heading direction, from one video frame to the next, rather than being relocated randomly within the 3D volume. This percentage, called the motion coherence c , remained constant within a trial, but changed between trials, taking values $\{25\%, 37\%, 70\%\}$ ($c \in \{0.25, 0.37, 0.70\}$) ($\{0\%, 12\%, 25\%, 37\%, 51\%, 70\%\}$ for subjects B2, D2, F2). The subject's sensitivity to the momentary visual evidence depends on coherence, such that $k_{vis}(c)$ is a function of c . In the main text we lay out an argument, based on neurophysiological evidence, for how we believe coherence influences the sensitivity to the visual cue. Critically, this argument leads to the assumption that a change in c not only modifies the drift rate in the diffusion model, according to $k_{\sigma,vis}(c) \propto a_{vis} c^{\gamma_{vis}}$ (a_{vis} and γ_{vis} being model parameters), but also causes the diffusion variance to change according to $\sigma^2(c) \propto 1 + b_{vis} c^{\gamma_{vis}}$ (where b_{vis} is another model parameter). The decision bound $\theta_{\sigma,vis}$, on the other hand, cannot depend on coherence, as the latter is unknown to the subject. Specifying the visual-only conditions by drift rate, diffusion variance, and bound would lead to over-parameterization, as a model with diffusion variance different from unity generates the same behavior as a model that has

unit diffusion variance and drift rate and bounds adequately normalized (see above). We avoid over-parameterization by performing this normalization, resulting in

$$k_{vis}(c) = \frac{k_{\sigma,vis}(c)}{\sigma(c)} = \frac{a_{vis}c^{\gamma_{vis}}}{\sqrt{1+b_{vis}c^{\gamma_{vis}}}}, \quad \theta_{vis}(c) = \frac{\theta_{\sigma,vis}}{\sigma(c)} = \frac{\theta_{\sigma,vis}}{\sqrt{1+b_{vis}c^{\gamma_{vis}}}}. \quad (24)$$

This allows for an arbitrary proportionality constant in the relationship

$\sigma^2(c) \propto 1 + b_{vis}c^{\gamma_{vis}}$, as this constant can be absorbed into a_{vis} and $\theta_{\sigma,vis}$. To summarize, we model behavior in the visual condition for any coherence by a unit variance diffusion model with sensitivity time-course given by $v(t)$, and parameterized by

$$\{a_{vis}, \gamma_{vis}, b_{vis}, \theta_{\sigma,vis}\}.$$

In the vestibular condition, momentary evidence is not influenced by coherence, such that we can model behavior with a unit-variance diffusion model having sensitivity k_{vest} , time-course $a(t)$, and a diffusion model bound θ_{vest} . Thus, the model for the vestibular condition is parameterized by $\{k_{vest}, \theta_{vest}\}$.

In the combined condition, the coherence of the visual stimulus again influences the sensitivity to the momentary evidence. Given that visual and vestibular cues are combined optimally, the sensitivity to the evidence is completely determined by that to the two separate cues. In particular, we have $k_{comb}(c)^2 = k_{vis}(c)^2 + k_{vest}^2$ by Eq. (10) and its sensitivity time-course $d(t, c)$ given by Eq. (16) with k_{vis} replaced by $k_{vis}(c)$. As in the visual condition, we assume a constant bound $\theta_{\sigma,comb}$ and a variance that is linearly related to coherence taken to power γ_{comb} , such that the normalized bound becomes

$$\theta_{comb}(c) = \frac{\theta_{\sigma,comb}}{\sqrt{1+b_{comb}c^{\gamma_{comb}}}}. \quad (25)$$

Thus, the model for the combined condition is characterized by a unit-variance diffusion model with sensitivity determined by the unimodal conditions, and a bound that is parameterized by $\{\theta_{\sigma,comb}, b_{comb}, \gamma_{comb}\}$.

We assume that reaction times featured by the subjects are composed of the decision time, as predicted by the diffusion model, and a non-decision time that captures the initial stimulus processing delay and the motor preparation time. We require this non-

decision time to be constant for all stimuli within each of the three stimulus modalities (vestibular, visual, combined), but we allow it to vary between modalities. Thus, the non-decision time is captured by the three parameters $\{t_{nd,vis}, t_{nd,vest}, t_{nd,comb}\}$. To account for random choices due to accidental button presses or lapses of attention, we introduce a lapse probability p_{lapse} with which the decision was performed randomly (with probability $\frac{1}{2}$ for each motion direction) rather than as predicted by the diffusion model. Additionally, we captured a potential bias in heading perception (i.e., horizontal shift of the psychometric function) by one additional bias parameter $\tilde{h}_{c,cond}$ for each combination of stimulus modality and coherence. Overall, given that the diffusion model predicts mean decision times represented by $t_{DM,corr}(h, c, cond, \varphi)$ and $t_{DM,incorr}(h, c, cond, \varphi)$ for correct and incorrect decisions, respectively, with model parameters φ , and given that the probability of choosing ‘rightward’ for each combination of heading direction h , visual motion coherence $c \in \{0.25, 0.37, 0.70\}$ ($c \in \{0, 0.12, 0.25, 0.37, 0.51, 0.70\}$ for subjects B2, D2, F2) and stimulus condition $cond \in \{vis, vest, comb\}$ is represented by $p_{DM,r}(h, c, cond)$, we assumed that the subject would feature mean reaction times and choice probabilities given by

$$\begin{aligned}
t_{corr}(h, c, cond, \varphi) &= t_{DM,corr}(h + \tilde{h}_{c,cond}, c, cond, \varphi) + t_{nd,cond}, \\
t_{incorr}(h, c, cond, \varphi) &= t_{DM,incorr}(h + \tilde{h}_{c,cond}, c, cond, \varphi) + t_{nd,cond}, \\
p_r(h, c, cond, \varphi) &= (1 - p_{lapse}) p_{DM,r}(h + \tilde{h}_{c,cond}, c, cond, \varphi) + p_{lapse} \frac{1}{2}.
\end{aligned} \tag{26}$$

The diffusion model itself and the non-decision times were parameterized by 12 parameters $\{a_{vis}, \gamma_{vis}, b_{vis}, \theta_{\sigma,vis}, k_{vest}, \theta_{vest}, \gamma_{comb}, b_{comb}, \theta_{\sigma,comb}, t_{nd,vis}, t_{nd,vest}, t_{nd,comb}\}$, and an additional 8 parameters (14 parameters for subjects B2, D2, F2) captured the biases and lapse rates. With these parameters, we modeled reaction times (separately for correct and incorrect decisions) and proportions of rightward/leftward choices for 56 different combinations of heading direction h , coherence c , and stimulus condition $cond$. In total, 168 data points (312 data points for subjects B2, D2, F2) were fit with a model

containing 12 primary parameters (along with 8 or 14 additional parameters to account for biases and lapse rates).

Alternative Parameterization

To ensure that our particular choice of parameterization did not bias our results on optimal evidence accumulation, we performed the same analysis with two additional parametric forms for sensitivities and normalized bounds. As shown in Fig. 7-figure supplement 2b and 2c, neither form changed our conclusions. Furthermore, Bayesian model comparison indicated that these alternative forms introduce a larger number of parameters than justifiable by the improvement in goodness-of-fit (Fig 4a).

The first alternative parameterization questions the relation between diffusion variance and drift rate. For the visual condition in the optimal model we have assumed this variance to be proportional to $1 + b_{vis} c^{\gamma_{vis}}$ and the drift to follow $a_{vis} c^{\gamma_{vis}}$. In both cases, coherence is take to the same power γ_{vis} . To test if a different power might explain the behavior better, we left the drift rate unchanged, but modified the variance to be proportional to $1 + b_{vis} c^{\xi_{vis}}$, where ξ_{vis} is an additional parameter. Figure 3-figure supplement 2 reveals that this modification leads to slightly different fits (dashed lines), while not qualitatively changing the relation between a model assuming optimal evidence accumulation and variants that do not (Fig. 7-figure supplement 2b). However, Fig. 7-figure supplement 2a shows that introducing this additional parameter is not justified by the minor increase in goodness-of-fit.

A second alternative parameterization abolishes any functional relationship between drifts, bounds, and coherences, and instead fits these drifts and bounds for each coherence and modality separately. That is, for the visual condition, drifts and bounds are modeled by one separate $k_{vis}(c)$ and $\theta_{vis}(c)$ per coherence. The vestibular condition is modeled, as before, with two parameters, k_{vest} and θ_{vest} . In the combined condition, we assume optimal cue combination, such that $k_{comb}(c) = \sqrt{k_{vis}(c)^2 + k_{vest}^2}$, but fit the bounds $\theta_{comb}(c)$ for each coherence separately. Thus, the model still assumes optimal accumulation of evidence across both time and cues, but makes no statement about how the tradeoff between speed and accuracy depend on the visual coherence. It replaces the 9

parameters (not counting the non-decision times) of the original model by 11 parameters (20 parameters for subjects B2, D2, F2) for drifts and bounds. Figure 3-figure supplement 1 shows the drifts and bounds for full model fits for each subject, and how they relate the other two parameterizations. As shown in Fig. 7-figure supplement 2c, abolishing the function form for these model variables does not qualitatively change the relation between optimal and suboptimal models. However, they do not explain the behavior better than a model with the original parameterization (Fig. 7-figure supplement 2a).

Model Fitting

We fit the model separately to the behavior of each subject by finding the model parameters ϕ (see previous section) that maximized their likelihood given the observed behavior. As in (Palmer, Huk et al. 2005), we assumed that the fraction of correct choices followed a binomial distribution, and that the reaction times of correct and incorrect choices were distributed according to a Gaussian centered on the empirical mean and spread according to the standard error. That is, for each combination of heading h , coherence c , and condition $cond$, we assumed the likelihood of ϕ to describe the choice fraction by

$$L_{r,h,c,cond}(\phi) = \text{Bin}(\hat{p}_r(h, c, cond) n_{h,c,cond} | n_{h,c,cond}, p_r(h, c, cond, \phi)), \quad (27)$$

which is a Binomial distribution over the observed number of rightwards choices, $\hat{p}_r(h, c, cond) n_{h,c,cond}$, given a total number of $n_{h,c,cond}$ trials and the model prediction $p_r(h, c, cond, \phi)$. The likelihood terms describing the reaction times were given by the Gaussian

$$L_{corr,h,c,cond}(\phi) = N\left(\hat{t}_{corr}(h, c, cond) | t_{corr}(h, c, cond, \phi), \frac{\text{var}_{corr}(h, c, cond, \phi)}{n_{corr,h,c,cond}}\right), \quad (28)$$

for reaction times corresponding to correct choices, and an analogous term

$L_{incorr,h,c,cond}(\phi)$ for those corresponding to incorrect choices. In the above $\hat{t}_{corr}(h, c, cond)$ is the observed mean reaction time over the $n_{corr,h,c,cond}$ trials in which correct choices were made, $t_{corr}(h, c, cond, \phi)$ is the mean reaction time predicted by the model, and

$\text{var}_{corr}(h, c, cond, \varphi)$ is the variance of this prediction. Overall, the complete likelihood was given by

$$L(\varphi) \prod_{h, c, cond} L_{r, h, c, cond}(\varphi) L_{corr, h, c, cond}(\varphi) L_{incorr, h, c, cond}(\varphi). \quad (29)$$

Fitting the model consisted of finding the parameter vector φ for each subject that maximized this likelihood.

Model predictions were found by evaluating Eq. (26). For each combination of heading, coherence, and stimulus condition, we computed the diffusion model predictions by numerically evaluating the reaction time distributions for either choice in steps of 5ms, using a method described previously (Smith 2000). Based on these distributions, we computed the probability of a choosing ‘rightward’ and the mean and variance of the reaction times for either choice.

To find the maximum likelihood parameters, we acquired a three-step approach that avoided getting stuck in likelihood function plateaus or local maxima. First, we performed gradient ascent on the log-likelihood to find the initial (potentially local) maximum. We used the found parameter vector as initial sample for taking 44000 samples from the Bayesian parameter posterior by Markov Chain Monte Carlo methods, assuming a bounded uniform parameter prior. Last, we picked the highest-likelihood sample as a starting point for another gradient ascent step to find the posterior’s mode. This mode was used as the maximum likelihood parameter vector. The resulting model parameters are shown for each subject in Fig. 3-figure supplement 2. All pseudo-gradient ascent maximizations were performed with the Optimization Toolbox of Matlab R2013a (Mathworks), using stringent stopping criteria ($\text{TolFun} = \text{TolX} = 10^{-20}$) to prevent premature convergence. For posterior sampling we utilized a custom Matlab implementation of slice sampling (Neal 2003). The parameter posterior variances reported in Fig. 3-figure supplement 2 were computed from the second half of all posterior samples of the Markov Chain.

The coefficient of determination that was used to describe the overall goodness-of-fit in Fig. 7 was computed as follows. The average coefficient of determination

$$R^2(\varphi) = \frac{1}{2} \left(R_{psych}^2(\varphi) + R_{chron}^2(\varphi) \right) \text{ is for each subject the average of } R_{psych}^2(\varphi) \text{ and}$$

$R_{chron}^2(\varphi)$, that is, the adjusted coefficients of determination for the psychometric and the chronometric curves, respectively. $R_{psych}^2(\varphi)$ is computed from

$$\tilde{R}_{psych}^2(\varphi) = 1 - \frac{\sum_{h,c,cond} w_{h,c,cond} (\hat{p}_r(h,c,cond) - p_r(h,c,cond,\varphi))^2}{\sum_{h,c,cond} w_{h,c,cond} (\hat{p}_r(h,c,cond) - \bar{p}_r)^2}, \quad (30)$$

by $R_{psych}^2(\varphi) = \tilde{R}_{psych}^2(\varphi) - (1 - \tilde{R}_{psych}^2(\varphi)) \frac{k}{N_s - k - 1}$, where $\hat{p}_r(h,c,cond)$ and $p_r(h,c,cond,\varphi)$ are the same terms as in the above likelihood, \bar{p}_r is the mean probability of choosing right over all trials, $w_{h,c,cond}$ is the fraction of trials with heading h , coherence c and condition $cond$, k is the number of model parameters, and N_s is the number of trials performed by subject s . For the chronometric curve we consider reaction times for both correct and incorrect choices by computing

$$\tilde{R}_{chron}^2(\varphi) = 1 - \frac{\sum_{h,c,cond} \left(w_{corr,h,c,cond} (\hat{t}_{corr}(h,c,cond) - t_{corr}(h,c,cond,\varphi))^2 + w_{incorr,h,c,cond} (\hat{t}_{incorr}(h,c,cond) - t_{incorr}(h,c,cond,\varphi))^2 \right)}{\sum_{h,c,cond} \left(w_{corr,h,c,cond} (\hat{t}_{corr}(h,c,cond) - \bar{t})^2 + w_{incorr,h,c,cond} (\hat{t}_{incorr}(h,c,cond) - \bar{t})^2 \right)}, \quad (31)$$

where $\hat{t}_{corr}(h,c,cond,\varphi)$ and $\hat{t}_{incorr}(h,c,cond,\varphi)$ are again the same as in the likelihood, \bar{t} is the mean reaction time over all trials, and $w_{corr,h,c,cond}$ and $w_{incorr,h,c,cond}$ are the fractions of correct and incorrect trials (out of all trials, such that $w_{h,c,cond} = w_{corr,h,c,cond} + w_{incorr,h,c,cond}$), respectively, that match $h,c,cond$.

Alternative, Sub-Optimal Models

We compared the fit quality of the optimal model to that of various, mostly sub-optimal models. These models are described below.

Free cue combination weights. In the optimal model, the sensitivity to the momentary evidence for the combined condition is determined by the sensitivities to the two separate

cues. In particular, $k_{comb}(c)^2 = k_{vis}(c)^2 + k_{vest}^2$ is assumed to hold. We introduced an alternative model in which $k_{comb}(c)$ is a free parameter for each coherence that is fitted independently of $k_{vis}(c)$ and k_{vest} to test two things: first, we were interested in comparing if the independently fit $k_{comb}(c)$ match those predicted from fits to the unimodal conditions. As discussed in the main text, this turns out to be the case (Fig. 6). Second, we wanted to know if loosening the optimality constraint explains the subjects' behavior better. For a fair comparison, we observe that this modification introduces one additional parameter per coherence when compared to the optimal model. Since the modified model is strictly more general than the optimal model, it is expected to fit the behavior at least as well or better than the optimal model. However, as described in the main text, Bayesian model comparison that takes into account the additional number of parameters revealed that the increased goodness-of-fit does not justify the additional degrees of freedom (Fig. 7).

Fixed cue combination weights. When performing optimal cue combination, the different cues ought to be weighted according to their respective sensitivities, as described by Eq. (17). We tested whether this was indeed the case by introducing a model variant that weights the momentary evidence of both cues equally. The evidence provided by each cue was still accumulated optimally over time according to Eq. (12), such that the momentary evidences were given by $\dot{X}_{vis}(t) = v(t)\dot{x}_{vis}(t)$ and $\dot{X}_{vest}(t) = a(t)\dot{x}_{vest}(t)$. However, rather than combining across modalities as given by Eq. (17), we performed a simple average: $\dot{X}_{comb}(t) = \frac{1}{2}(\dot{X}_{vis}(t) + \dot{X}_{vest}(t))$. In the combined condition, this resulted in a diffusion model with drift rate given by $\frac{1}{2}(v(t)^2 k_{vis}(c) + a(t)^2 k_{vest}) \sin(h)$ and diffusion variance given by $\frac{1}{4}(v(t)^2 + a(t)^2)$. The bound was left unchanged, as given by Eq. (25). The number of parameters for this model variant is the same as for the optimal model.

Weighting both cues by either acceleration or velocity. When designing the model we have assumed that the sensitivity time-course of the visual and vestibular modality is

determined by the motion velocity and acceleration, respectively. To test this choice we introduced a model variant that weights both modalities by either acceleration or velocity. For the first variant we replaced $v(t)$ in Eq. (12) and all equations that follow by $a(t)$. For the second variant we replaced $a(t)$ by $v(t)$ in all relevant equations.

No temporal weighting of momentary evidence. Our theory predicts that optimal accumulation of evidence over time requires this evidence to be weighted according to its associated momentary sensitivity, as given by Eq. (7). To test this, we introduced an additional model that did not perform this temporal weighting. Instead, we assumed the diffusion models for the unimodal conditions to feature a unit diffusion variance and un-weighted drift rates, $v(t)k_{vis}(c)\sin(h)$ and $a(t)k_{vest}(c)\sin(h)$, for the visual and vestibular conditions, respectively. The cues were still combined according to the optimal combination rule, Eq. (10), resulting in a diffusion model for the combined condition with unit variance and drift rate given by $\frac{k_{vis}^2v(t)+k_{vest}^2a(t)}{k_{comb}}\sin(h)$. The bound was left unchanged, resulting in the number of parameters to be the same as for the optimal model.

No temporal weighting and fixed cue combination weights. The last model variant discards both the assumption of temporal weighting of evidence and the assumption of sensitivity-based weighting when combining the cues across modalities. Thus, the diffusion models describing the unimodal conditions featured, as before, a unit diffusion variance and drift rates, $v(t)k_{vis}(c)\sin(h)$ and $a(t)k_{vest}(c)\sin(h)$, for the visual and vestibular conditions, respectively. In the combined condition, momentary evidence was summed according to $\dot{x}_{comb}(t) = \frac{1}{\sqrt{2}}(\dot{x}_{vis}(t) + \dot{x}_{vest}(t))$, resulting in a diffusion model with unit variance and drift rate $\frac{1}{\sqrt{2}}(v(t)k_{vis}(c) + a(t)k_{vest}(c))\sin(h)$. The $\frac{1}{\sqrt{2}}$ weighting was chosen to ensure unit variance, but any other weighting would have resulted in the same fits. The bounds were parameterized as in the optimal model, such that the number of parameters was the same as in the original model.

References

- Bogacz, R., E. Brown, et al. (2006). "The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks." Psychological Review **113**(4): 700-765.
- Clark, J. J. and A. L. Yuille (1990). Data fusion for sensory information processing systems. Boston, Kluwer Academic Publishers.
- Drugowitsch, J., R. Moreno-Bote, et al. (2012). "The Cost of Accumulating Evidence in Perceptual Decision Making." J Neurosci **32**(11): 3612-3628.
- Gold, J. I. and M. N. Shadlen (2002). "Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward." Neuron **36**(2): 299-308.
- Kiani, R. and M. N. Shadlen (2009). "Representation of confidence associated with a decision by neurons in the parietal cortex." Science **324**(5928): 759-764.
- Laming, D. R. J. (1968). Information theory of choice-reaction times. London., Academic P.
- Lisberger, S. G. and J. A. Movshon (1999). "Visual motion analysis for pursuit eye movements in area MT of macaque monkeys." J Neurosci **19**(6): 2224-2246.
- Neal, R. M. (2003). "Slice sampling." Annals of Statistics **31**(3): 705-767.
- Palmer, J., A. C. Huk, et al. (2005). "The effect of stimulus strength on the speed and accuracy of a perceptual decision." J Vis **5**(5): 376-404.
- Price, N. S., S. Ono, et al. (2005). "Comparing acceleration and speed tuning in macaque MT: physiology and modeling." J Neurophysiol **94**(5): 3451-3464.
- Ratcliff, R. (1978). "Theory of Memory Retrieval." Psychological Review **85**(2): 59-108.
- Schlack, A., B. Krekelberg, et al. (2007). "Recent history of stimulus speeds affects the speed tuning of neurons in area MT." J Neurosci **27**(41): 11009-11018.
- Shadlen, M. N., T. D. Hanks, et al. (2006). The Speed and Accuracy of a Simple Perceptual Decision: A Mathematical Primer. Bayesian Brain: Probabilistic Approaches to Neural Coding. K. Doya, S. Ishii, A. Pouget and R. P. N. Rao, MIT Press: 209-238.
- Smith, P. L. (2000). "Stochastic Dynamic Models of Response Time and Accuracy: A Foundational Primer." J Math Psychol **44**(3): 408-463.
- Wald, A. (1947). Sequential analysis. New York, London, J. Wiley & sons, Chapman & Hall.
- Wald, A. and J. Wolfowitz (1948). "Optimum Character of the Sequential Probability Ratio Test." The Annals of Mathematical Statistics **19**(3): 326-339.

Optimal multisensory decision-making in a reaction-time task

Outcome of additional statistical hypothesis tests

Jan Drugowitsch, Gregory C. DeAngelis, Eliana M. Klier,
Dora E. Angelaki, Alexandre Pouget

Subject	$\sigma_{comb}(c) < \min\{\sigma_{vis}(c), \sigma_{vest}\}$						$\sigma_{comb}(c) > \sigma_{vis}(c)$						$\sigma_{comb}(c) > \sigma_{pred}(c)$						$\sigma_{comb}(c) < \sigma_{vest}$
	0%	12%	25%	37%	51%	70%	0%	12%	25%	37%	51%	70%	0%	12%	25%	37%	51%	70%	70%
A			1	1		1			1	0.002		0			1	0		0	0
B			1	1		1			1	0		0			0.287	0		0	0
C			1	1		1			1	1		1			1	1		0.168	0.210
D			0.57	1		1			1	0		0			1	0		0	0
E			1	1		1			1	0.027		0			1	0		0	0.007
F			1	1		1			1	0.662		0.006			1	0		0	0
G			1	1		1			0.005	0		0			0	0		0	0
B2	1	1	1	1	1	1	1	1	1	0.696	0	0	1	1	0.072	0	0	0	0.039
D2	1	1	1	1	1	1	1	1	1	1	0	0	1	0.377	0.813	0.007	0	0	0
F2	1	1	1	1	1	1	1	1	1	0	0	0.042	0.319	0.325	0	0	0	0	0

Table A. p-values for comparisons of decision thresholds, for different coherence values of the visual stimulus. The performed tests are indicated in the first row (*vis* = visual-only condition, *vest* = vest-only condition, *comb* = combined condition, *pred* = predicted threshold, *c* = coherence). From left to right, the four tests conducted are: i) whether the combined threshold is significantly less than the threshold of the more sensitive modality, ii) whether the combined threshold is significantly greater than the visual threshold, iii) whether the combined threshold is significantly greater than the predicted threshold, and iv) whether the combined threshold is significantly less than the vestibular threshold. The p-values in bold indicate deviations from the general effects reported in the main text. The p-values are computed from 5000 bootstrapped samples for each threshold, and are Bonferroni-corrected for multiple one-tailed comparisons. p=0 represents p<0.0001

Subject	25% coherence				37% coherence				70% coherence			
	0.69°	1.96°	5.6°	16°	0.69°	1.96°	5.6°	16°	0.69°	1.96°	5.6°	16°
Miller's bound violations (bold = p<0.05)												
A	1	0.9473	0.9867	1	1	1	0.9996	0.9317	1	1	1	0.9978
B	1	1	1	1	1	1	1	1	1	1	1	1
C	0.8943	1	0.9537	0.7310	0.9681	0.8782	1	1	0.9250	1	0.9749	0.9973
D	0.5775	0.9606	0.9553	0.9958	1	1	0.9933	0.9919	1	0.9952	1	0.2127
E	1	1	0.6683	0.9159	0.9928	0.9839	0.9840	0.9840	1	1	1	1
F	0.8853	0.7483	1	0.8086	0.9655	0.9910	1	0.9674	0.9896	0.6454	0.9986	0.0890
G	0.9335	0.9777	0.6540	0.4549	0.7447	0.9394	0.8502	0.4544	0.9927	0.9182	0.9344	0.0358
Grice's bound violations (bold = p<0.05)												
A	0.0001	0.0346	0	0.0701	0	0.0025	0.0361	0.0373	0.0102	0.0189	0.0005	0.1195
B	0	0	0	0	0	0	0	0	0	0	0.0007	0
C	0.0070	0.0119	0.0398	0.0016	0.8851	0.0267	0.2593	0	0	0	0	0
D	0.0002	0	0.0138	0.0050	0	0	0.0171	0.0158	0	0	0.1867	0.9706
E	0.0094	0.0394	0.0039	0.0750	0.0002	0.0080	0.0377	0.0078	0	0	0	0.0016
F	0.5889	0.4669	0.2710	0.0090	0.1625	0.3408	0.0671	0.0878	0.6905	0.7897	0.8045	1
G	0.0002	0.0823	0.1451	0.1495	0.0199	0.0116	0.0053	0.0378	0	0	0	0.9364

Table B. Comparing observed reaction time distributions to those predicted by a parallel race model, subjects A-G. Parallel race models predict the reaction time distribution in the combined condition, based on those observed in the two unimodal conditions. We tested for all heading/coherence combinations if these predictions matched the observed behavior. A violation of Miller's bound (Miller 1982) or Grice's bound (Grice, Canham et al. 1984) implies that the subject reacted significantly faster or slower, respectively, than predicted by a parallel race model. The table shows the uncorrected p-values resulting from a one-sided two-sample Kolmogorov-Smirnov test between the observed distributions and those corresponding to either of the two bounds. Values in boldface are significant bound violations, based on a Bonferroni-corrected threshold at 0.05.

Subject	0.69°	1.96°	5.6°	16°	0.69°	1.96°	5.6°	16°	0.69°	1.96°	5.6°	16°
Miller's bound violations (bold = p<0.05)												
	0% coherence				12% coherence				25% coherence			
B2	0.1007	0.3100	0.2449	0.2197	0.7388	0.7375	0.5766	0.4710	0.5262	0.7974	0.8567	0.4561
D2	0.9966	0.9701	0.9573	0.9213	0.9968	0.9892	0.9908	0.9964	0.9302	0.9888	0.9933	0.8156
F2	1	1	1	1	0.9969	0.9974	0.9978	0.9981	1	1	1	1
	37% coherence				51% coherence				70% coherence			
B2	0.7311	0.9129	0.9245	0.8675	0.5534	0.7226	0.7937	0.4994	0.1135	0.4465	0.3904	0.1012
D2	1	1	1	0.9890	1	1	0.9977	0.9963	1	1	1	1
F2	1	1	1	1	0.9515	0.9592	0.9652	0.7781	0.1522	0.2088	0.2094	0.0043
Grice's bound violations (bold = p<0.05)												
	0% coherence				12% coherence				25% coherence			
B2	0.5949	0.6636	0.6041	0.5912	0.5061	0.2075	0.0961	0.1455	0.1108	0.0781	0.0162	0.0408
D2	0	0	0	0	0.0012	0	0	0	0	0	0	0
F2	0.0030	0.0006	0	0	0.0004	0	0	0	0.0033	0.0020	0.0001	0.0003
	37% coherence				51% coherence				70% coherence			
B2	0.0619	0.0201	0.0106	0.0093	0.7827	0.4873	0.1659	0.3565	0.0697	0.0012	0.0008	0.0044
D2	0	0	0	0	0	0	0	0	0	0	0	0
F2	0.0153	0.0106	0.0004	0.0002	0.0083	0.0061	0.0028	0.0179	0.9530	0.9137	0.9257	0.9347

Table B (continued). Comparing observed reaction time distributions to those predicted by a parallel race model, subjects B2, D2, F2. See previous page for details.

Subject	2-way ANOVA					JB	Friedman test		
	factor	df	df(err)	F	p		df	χ^2	p
A	mod	1	1636	1167.45	0	7	1	467.07	0
	head	3		47.00	0				
	int	3		26.26	0				
B	mod	1	2867	1847.41	0	5	1	706.80	0
	head	3		204.96	0				
	int	3		47.61	0				
C	mod	1	1805	450.91	0	7	1	347.14	0
	head	3		23.34	0				
	int	3		3.71	0.011				
D	mod	1	4320	1720.95	0	8	1	935.21	0
	head	3		316.96	0				
	int	3		107.02	0				
E	mod	1	3254	2536.39	0	5	1	872.80	0
	head	3		126.10	0				
	int	3		62.97	0				
F	mod	1	2691	41.64	0		1	35.77	0
	head	3		167.17	0				
	int	3		8.63	0				
G	mod	1	3074	1459.71	0	7	1	565.61	0
	head	3		388.60	0				
	int	3		115.67	0				
B2	mod	1	1419	1428.33	0	2	1	604.62	0
	head	3		211.32	0				
	int	3		113.57	0				
D2	mod	1	1786	2214.62	0	5	1	792.66	0
	head	3		308.53	0				
	int	3		127.79	0				
F2	mod	1	1428	67.64	0	8	1	87.26	0
	head	3		171.53	0				
	int	3		11.39	0				

Table C. Statistical tests applied to comparing reaction times between 70%-coherence visual-only condition and vestibular-only condition. 2-way ANOVA: factors are modality (mod; 70% visual vs. vestibular), heading direction (head), and interaction (int); df(err) = df of error term; F = F statistics; p = p-value (p=0 represents $p<0.0001$). The JB column reports the number of violations (out of 8; 4 x visual, 4 x vestibular) of the normality assumption (Jarque-Bera, $p<0.05$). All tests are performed on the natural logarithm of the reaction time. These tests confirm that, for all subjects, the reaction times differ significantly between the visual and vestibular conditions. Except for subject C, who responded significantly slower, all subjects responded faster in the vestibular condition than in the 70% coherence visual condition (see Figure 3-figure supplement 1)

References

- Grice, G. R., L. Canham, et al. (1984). "Combination rule for redundant information in reaction time tasks with divided attention." Percept Psychophys **35**(5): 451-463.
- Miller, J. (1982). "Divided attention: evidence for coactivation with redundant signals." Cogn Psychol **14**(2): 247-279.

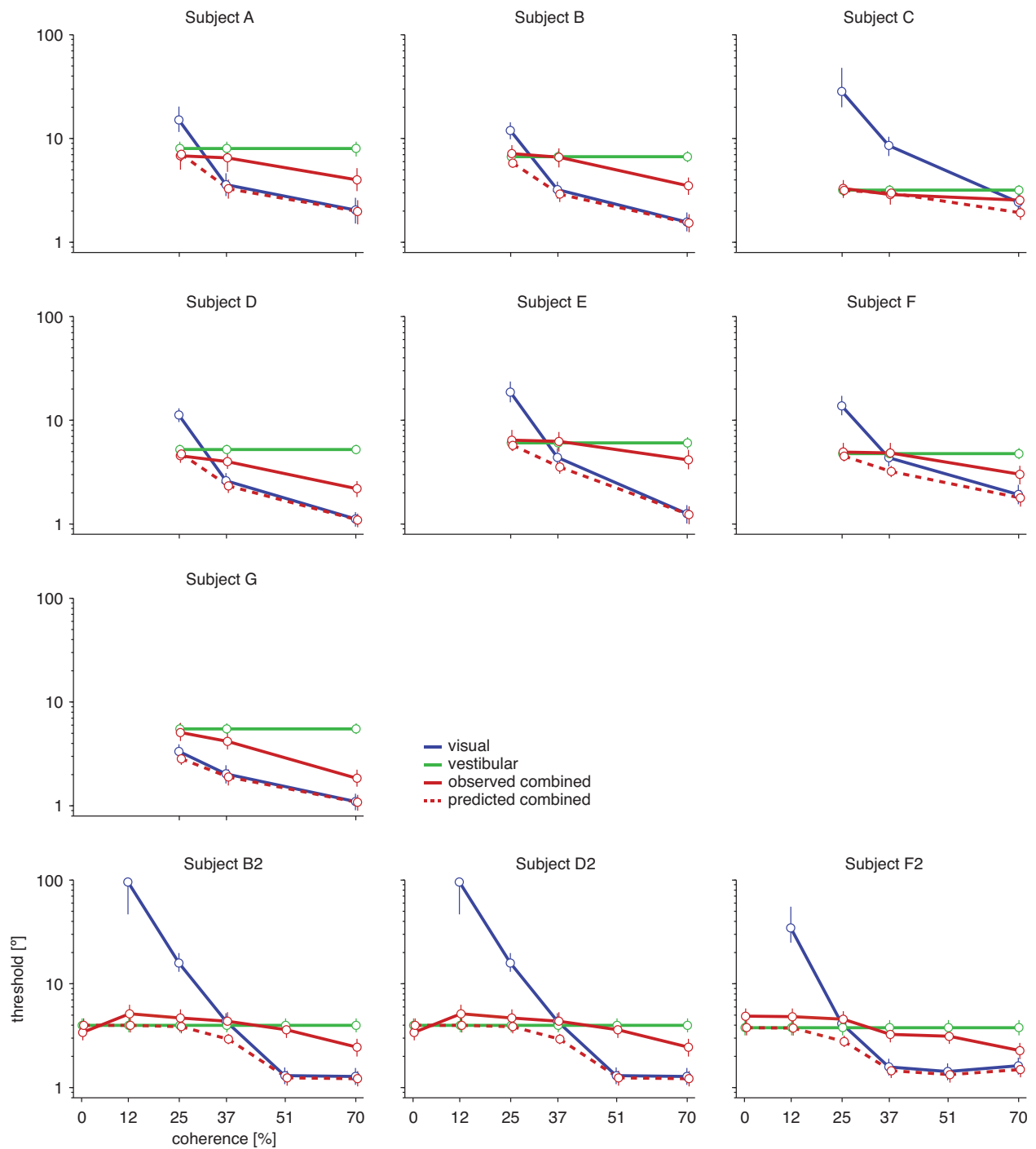


Figure 2 - figure supplement 1. Discrimination thresholds for all subjects and conditions. The psychophysical thresholds are found by fitting a cumulative Gaussian function to the psychometric curve for each condition. The predicted threshold is based on the visual and vestibular thresholds measured at the same coherence. The error bars indicate bootstrapped 95% CIs. Note that the observed thresholds in the combined condition (solid red curves) are consistently greater than the predicted thresholds (dashed red curves), especially at high coherences. For a statistical comparison between various thresholds see Supplementary Table 1.

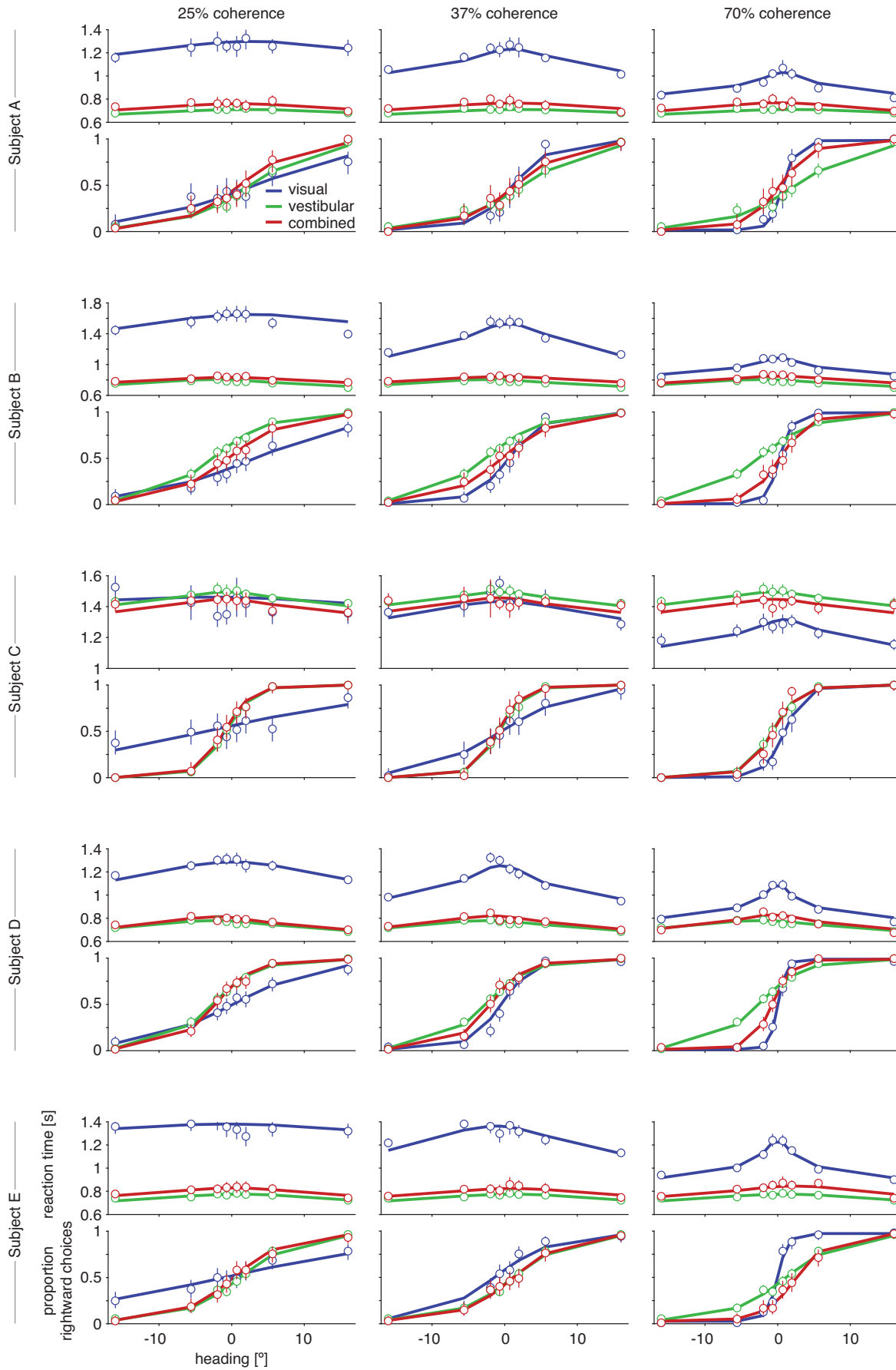


Figure 3 - figure supplement 1. Psychometric functions, chronometric functions, and model fits for all subjects. Behavioral data (symbols with error bars) and model fits (lines) are, for clarity, shown separately for each different coherence of the visual motion stimulus. The reaction time shown is the mean reaction time for correct trials, with error bars showing two SEMs (sometimes smaller than the symbols). Error bars on the proportion of rightward choices are 95% confidence intervals. Note that reaction times are shown only for correct trials, while the model is fit to both correct and incorrect trials.

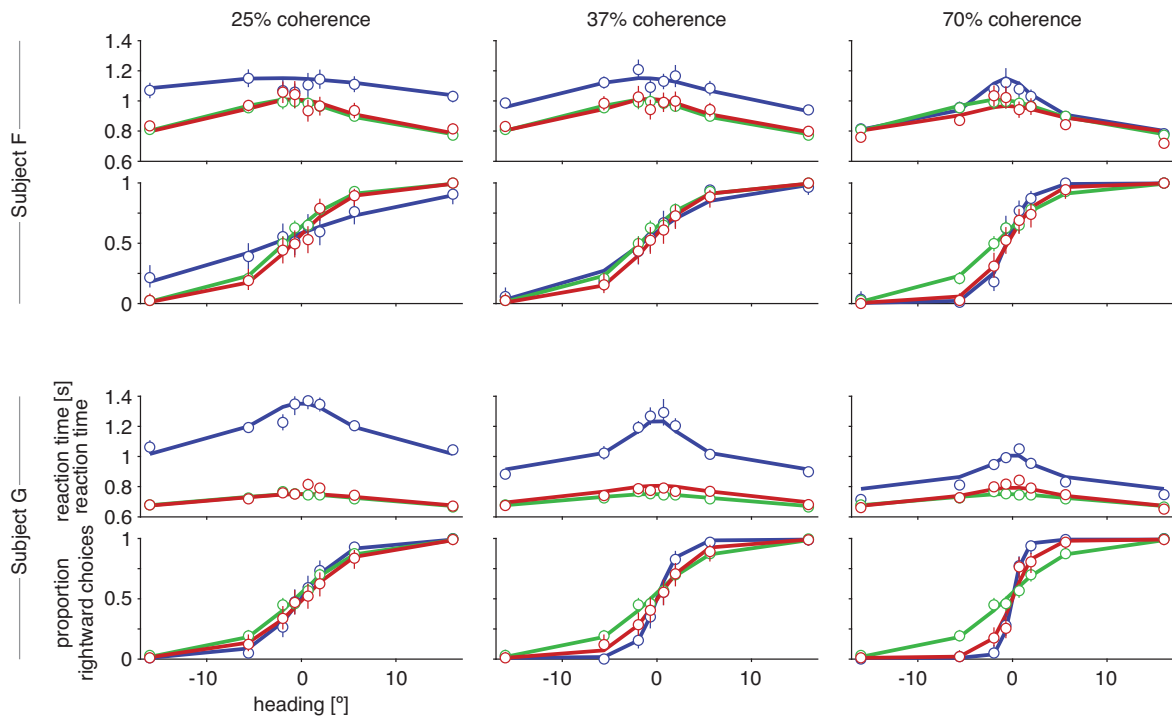


Figure 3 - figure supplement 1. Psychometric functions, chronometric functions, and model fits for all subjects (continued).

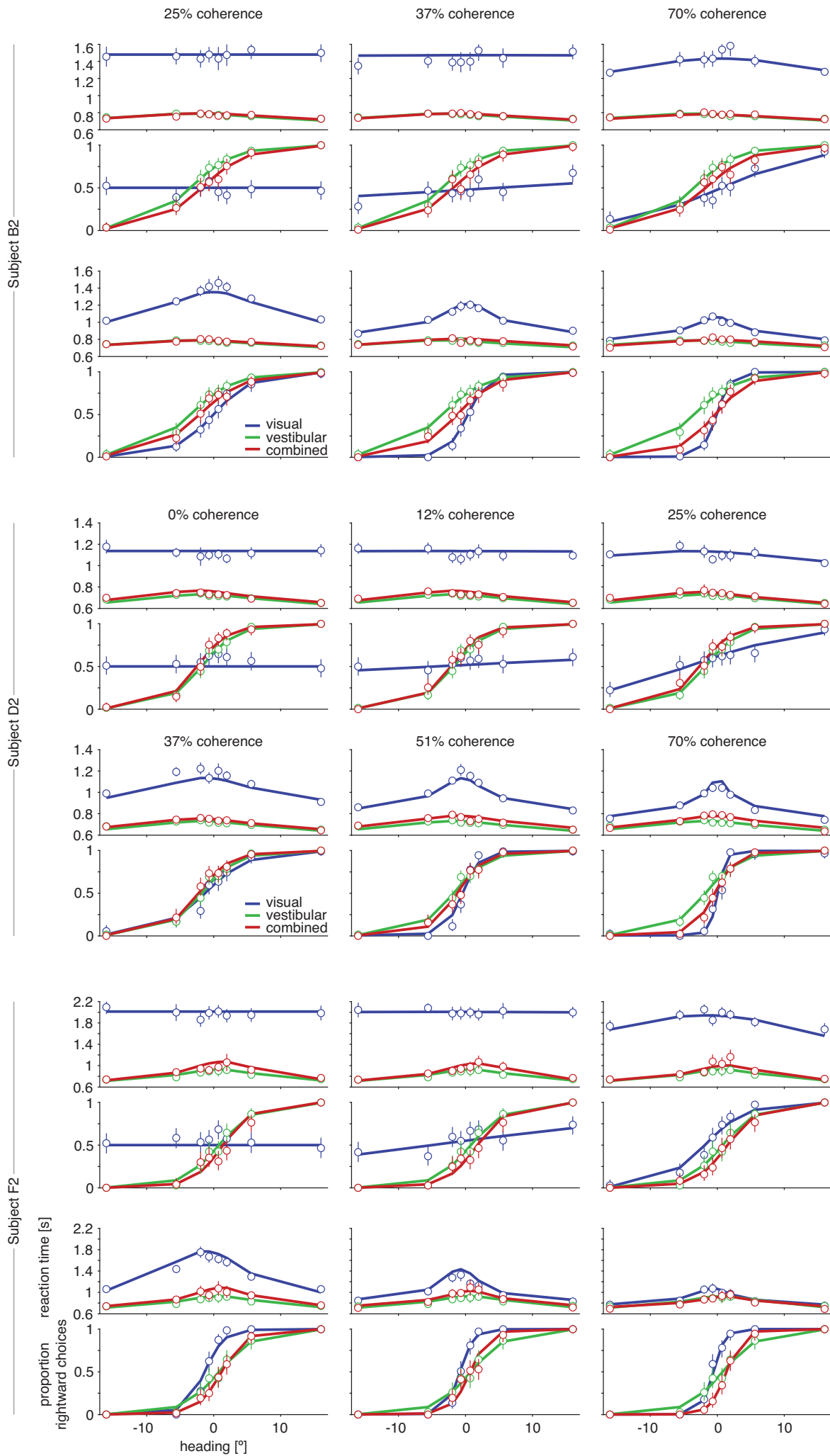


Figure 3 - figure supplement 1. Psychometric functions, chronometric functions, and model fits for all subjects (continued).

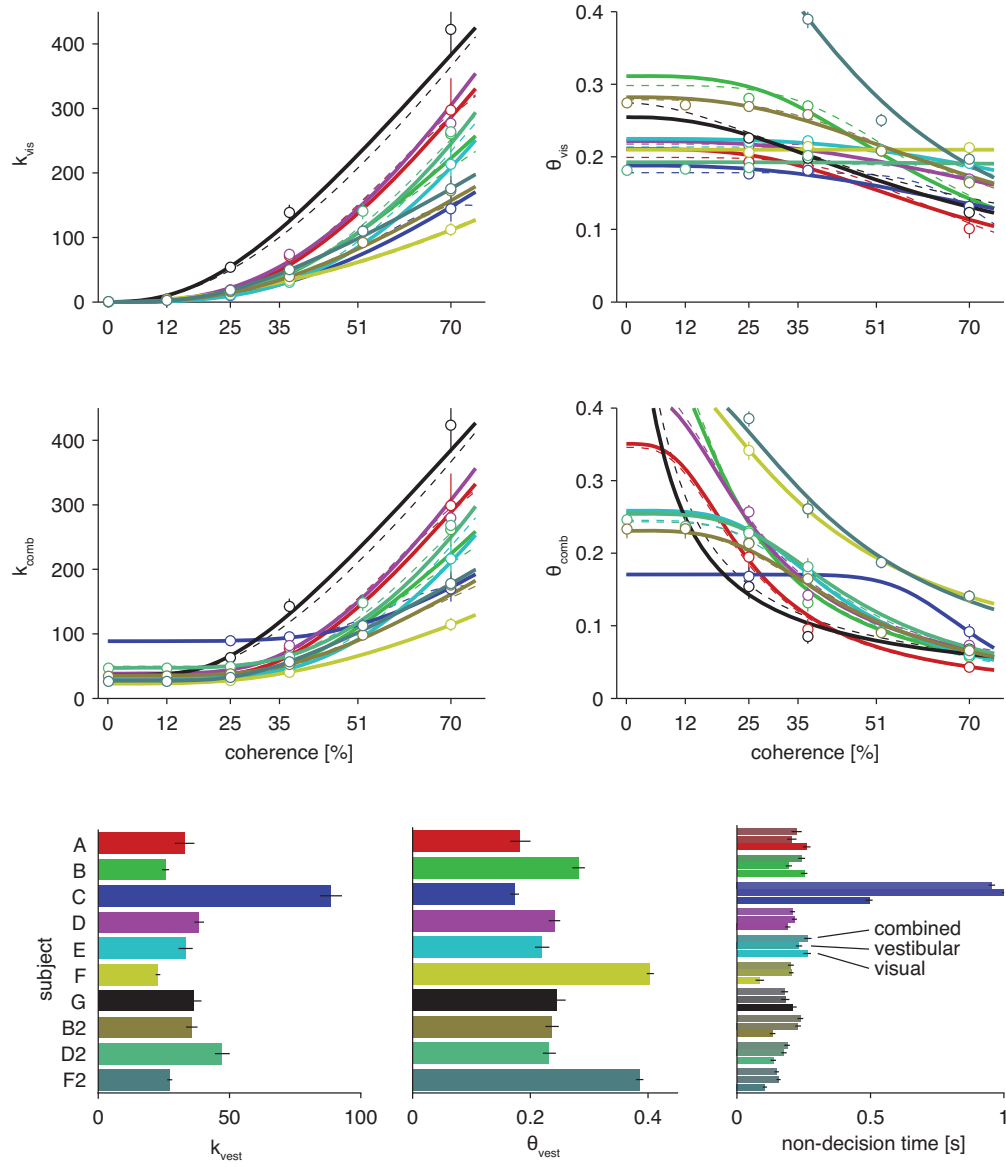


Figure 3 - figure supplement 2. Model parameters for fits of the optimal model and two alternative parameterizations. Based on the maximum likelihood parameters of full model fits for each subject, the four top plots show how drift rate and normalized bounds are assumed to depend on visual motion coherence. The solid lines show fits for the model described in the main text. The dashed lines show fits for an alternative parameterization with one additional parameter (see Supplementary Text). The circles show the fits of a model that, instead of linking them by a parametric function, fits these drifts and bounds for each coherence separately. As can be seen, the parametric functions qualitatively match these independent fits. The bottom bar graphs show drift rate and bound for the vestibular modalities and fitted non-decision times for each subject, all for the model parameterization described in the text. All error bars show ± 1 SD of the parameter posterior. Each color corresponds to a separate subject, with color scheme given by the bottom left bar graph.

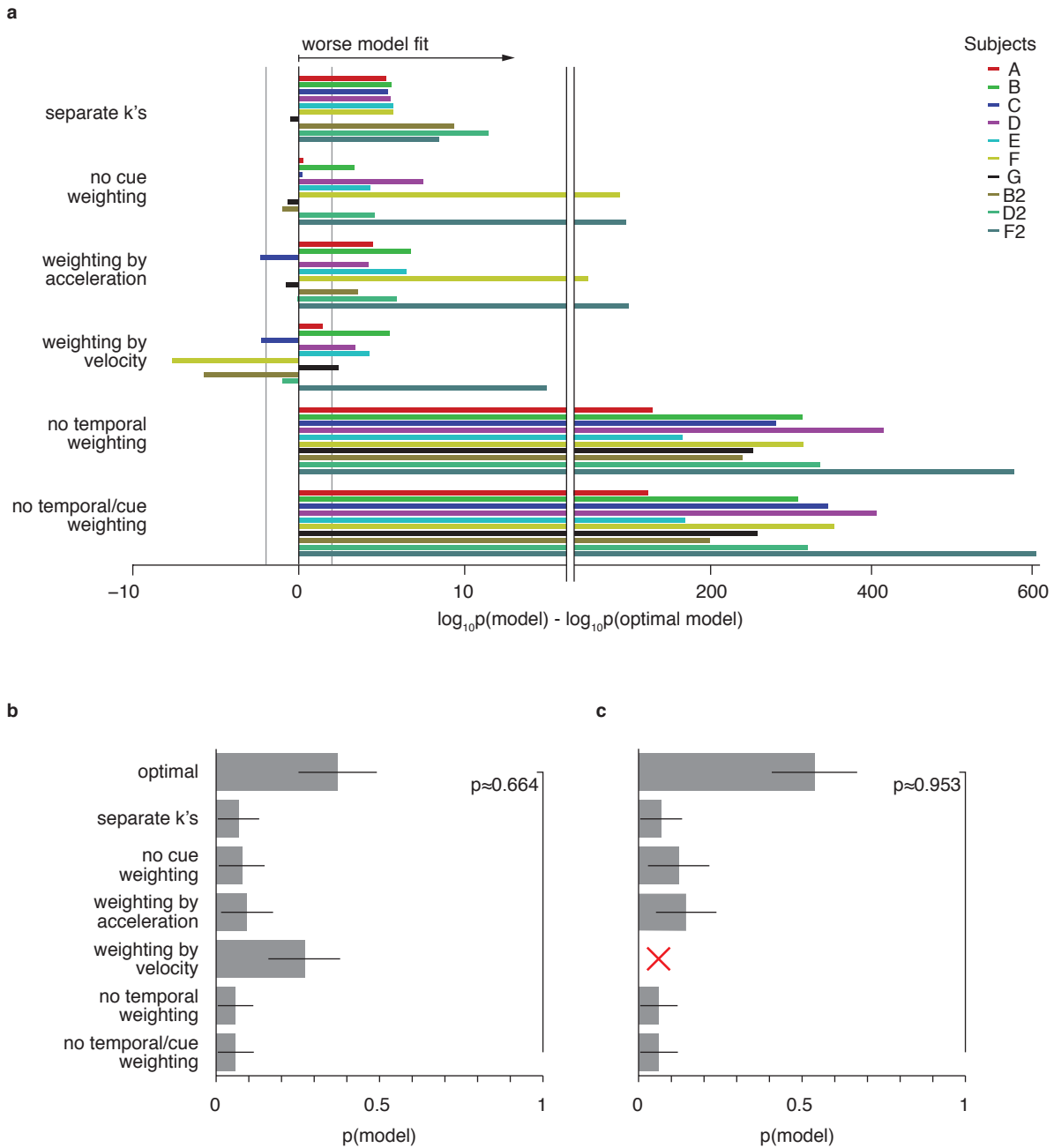


Figure 7-figure supplement 1. Model comparison per subject, and random-effects model comparison. (a) shows the contribution of each subject to the model comparison shown in Fig. 7b. As in Fig. 7b, the grey line shows the threshold above which the alternative models provide a decisively worse (if negative) or better (if positive) model fit. As can be seen, the model comparison is mostly consistent across subjects, except for models that weight both modalities either by acceleration or velocity only. Even in these cases, pooling across subjects leads to a decisively worse fit of the alternative model when compared to the optimal model (Fig. 7). (b) and (c) show the results of a random-effects Bayesian model comparison (Stephan, Penny et al. 2009). This model comparison infers the probability of each model to have generated the behavior observed for each subject, and is less sensitive to model fit outliers than the fixed-effects comparison shown in Fig. 7b (e.g., a single subject might strongly support an otherwise unsupported model, which could skew the overall comparison). (b) shows the inferred distribution over all compared models, and supports the optimal model with exceedance probability $p \approx 0.664$ (probability that the optimal model is more likely than any other model). This random-effects comparison causes models with very similar predictions to share some probability mass – in our case the optimal model and the model assuming evidence weighting by the velocity time-course. In (c) we perform the same comparison without the ‘weighting by velocity’ model, in which case the exceedance probability supporting the optimal model rises to $p \approx 0.953$.

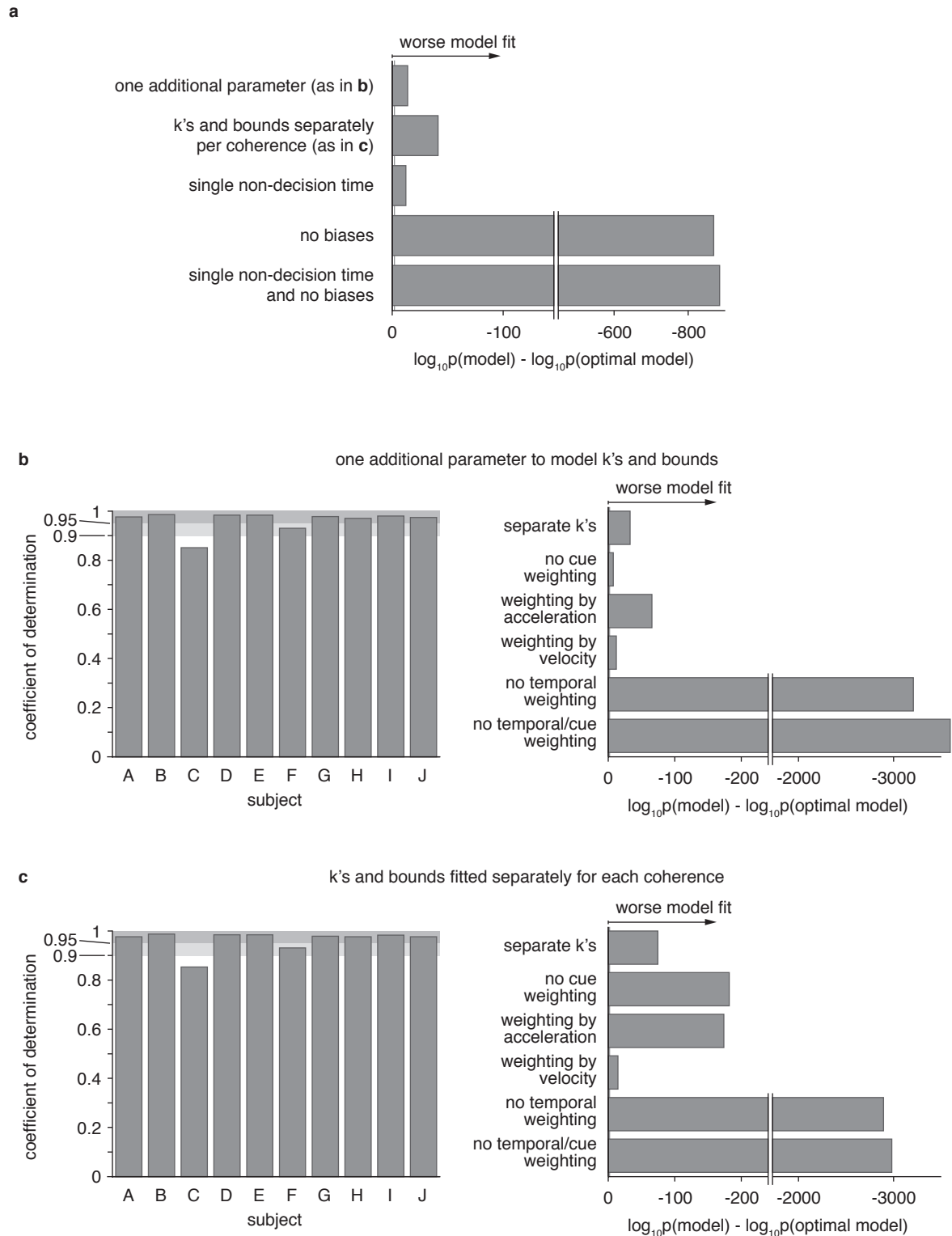


Figure 7 - figure supplement 2. Model comparison for models with alternative parameterization. (a) compares the optimal model as described in the main text to various alternative models. The first model changes how drifts and bounds relate to coherence (see Supplementary Text), and introduces one additional parameter. The second model fits drifts and bounds separately for all coherences. The other models either use a single non-decision time (instead of one or each modality), no heading biases, or a combination of both. The figure shows the Bayes factor, illustrating that in all cases the alternative models are decisively worse (grey line close to origin indicating threshold) than the original model. (b) and (c) show the overall model goodness-of-fit (left panels) of two model that used an alternative parameterization of how drifts and bounds depend on coherence (see (a)). Furthermore, it compares these models, which still perform optimal evidence accumulation across both time and cues, to sub-optimal models (right panels) that do not (except “separate k’s”, which is potentially optimal). These figures are analogous to Fig. 7 and show that neither change of parameterization qualitatively changes our conclusions.