SUPPLEMENTAL DATA

Further analysis of cue shifts (in conjunction with Suppl. Figure S1)

Previously, we found that *unsupervised* multisensory calibration is not dependent on cue reliability (Zaidel et al., 2011). In contrast, we observe here that *supervised* calibration clearly depends on cue reliability (Figures 2-4 in the main manuscript). In order to further elucidate the nature of this dependence, we plotted the magnitude of individual-cue PSE shifts vs. reliability ratio (RR) in the same way that we previously analyzed unsupervised calibration (Suppl. Fig. S1). For this analysis, positive values indicate a PSE shift towards the other cue, i.e., a shift which would reduce the cue conflict. Negative values indicate a shift away from the other cue (gray shaded region). For RR >1 the visual cue was more reliable; for RR<1 the vestibular cue was more reliable.

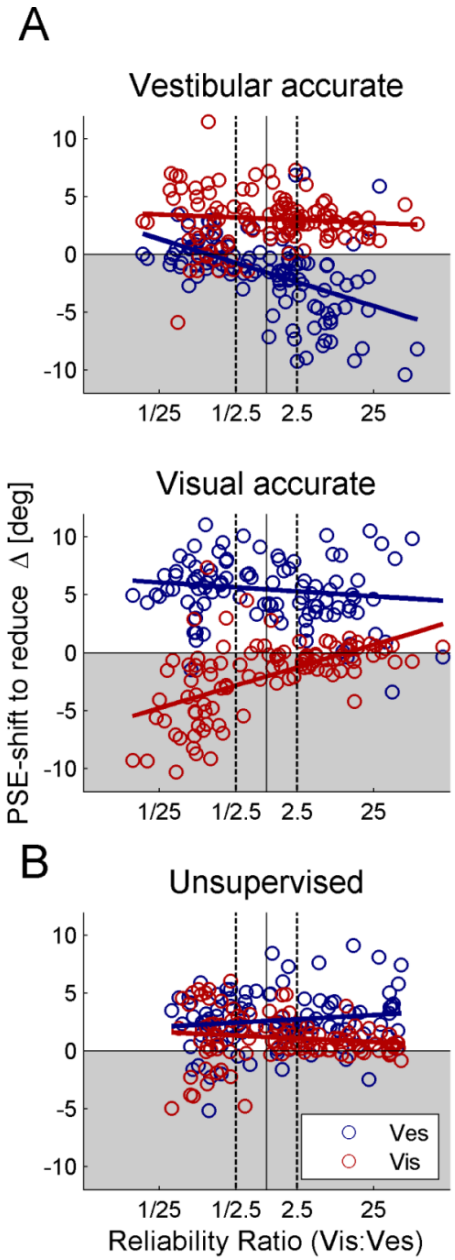
The scatter plots (and corresponding linear regression lines) demonstrate an intricate relationship between supervised cue calibration and RR. Immediately, clear differences can be observed between the types of calibration for the accurate vs. inaccurate cue. Firstly, calibration of the inaccurate cue (red and blue in Fig. S1A top and bottom, respectively) was largely independent of RR. This is demonstrated by the lack of any significant correlation between the PSE-shift and RR (Pearson correlation $p>0.1$ for both conditions). Also, the inaccurate cue shifted significantly in the direction required to reduce cue conflict, and to become more accurate ($p<0.0001$ for both conditions). All significant p-values presented here are after the Bonferroni correction for multiple comparisons. When testing for lack of significance, no correction was made.

By contrast, calibration of the accurate cue (blue and red in Fig. S1A top and bottom, respectively) was highly dependent on RR (Pearson correlation $p<0.0001$ for both conditions). For the accurate cue, we analyzed cue shifts separately for low and high RRs: when the accurate cue was also more reliable, it did not shift at all (blue, low-RR values in Fig. S1 A top, and red,

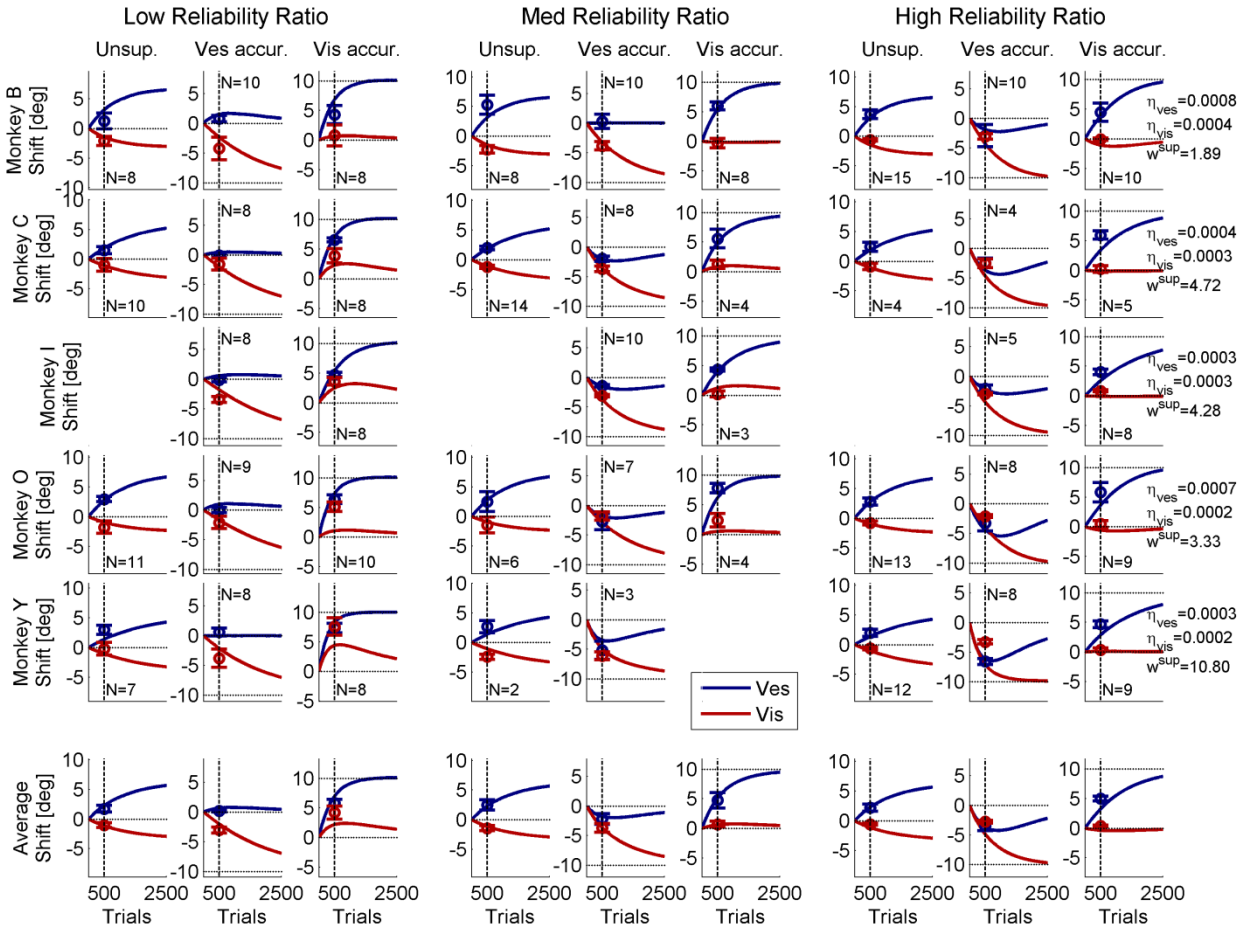
high-RR values in Fig. S1A bottom; t-test $p > 0.1$ and $p = 0.041$, respectively - we do not consider the latter significant, in view of multiple comparisons). However, when it was less reliable, it shifted significantly in the direction opposite to that required to reduce the cue conflict (blue, high-RR values in Fig. S1A top, and red, low-RR values in Fig. S1A bottom; t-test $p < 0.0001$ for both).

We also note that shift magnitude of the inaccurate cue was smaller for the vestibular-accurate paradigm (the inaccurate visual cue shifted 3.1°, Fig. S1A top) vs. visual accurate (the inaccurate vestibular cue shifted 5.5°, Fig. S1A bottom). These values were significantly different from one another (t-test $p < 0.0001$) indicating that the vestibular cue has greater plasticity than the visual cue. This parallels our finding of greater vestibular vs. visual plasticity also for unsupervised calibration (Fig. S1B; Zaidel et al., 2011). However, these supervised calibration shifts were more than twice the size and significantly greater than the unsupervised shifts ($p<0.0001$ for both visual and vestibular).

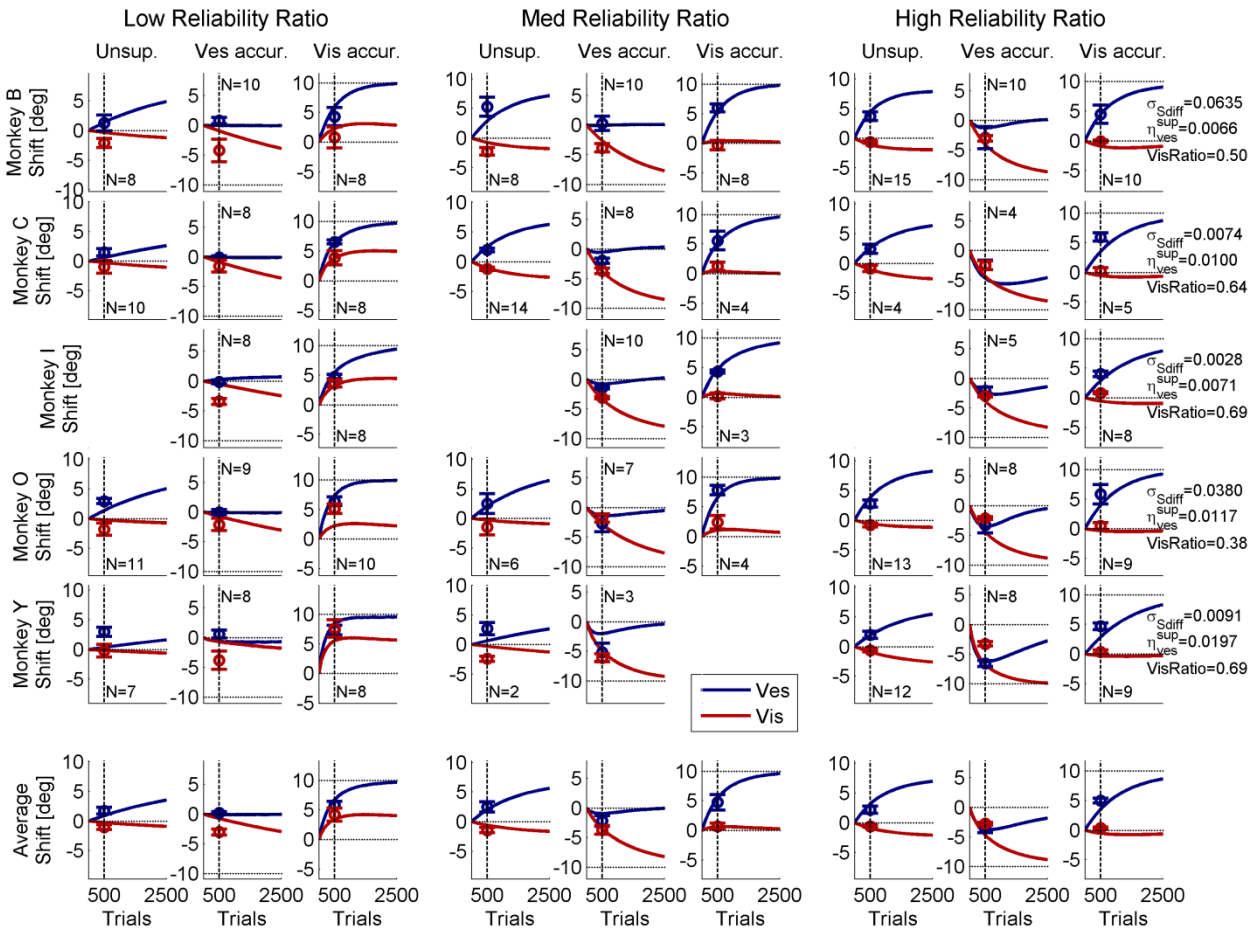
Interestingly, shifts of the accurate (but less reliable) cues were: -3.7º for vestibular accurate (Fig. S1A top) and -4.2º for visual accurate (Fig. S1A bottom). Here, not only was a significant difference no longer evident ($p>0.1$) but the effect was reversed – greater visual vs. vestibular shift. This observation, that shift magnitude is coupled between the cues (e.g., a larger shift of the inaccurate cue is associated with a larger shift of the accurate cue) reflects cue yoking.



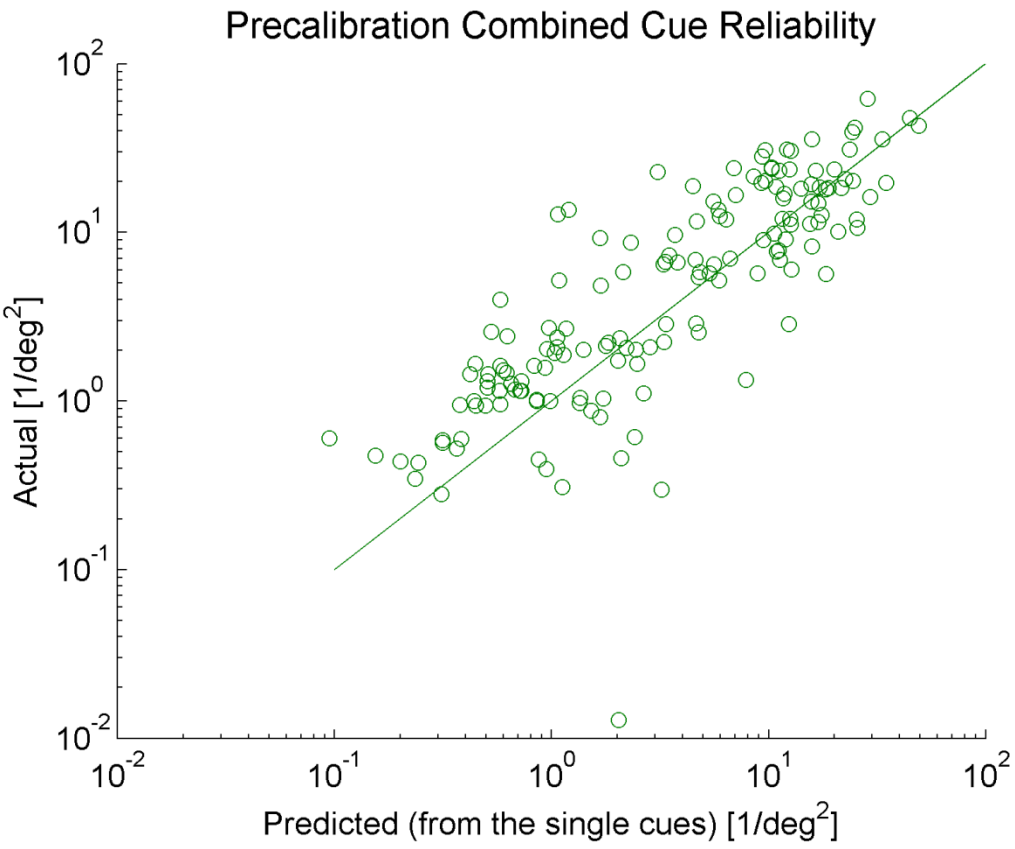
Supplemental Figure S1: PSE shifts, related to Figure 3. Shifts in the PSE (point of subjective equality) for vestibular (blue) and visual (red) cues are plotted as a function of reliability ratio (RR, ratio of visual to vestibular reliability) for (A) supervised and (B) unsupervised calibration conditions. PSE shifts were normalized by the direction of cue discrepancy, such that positive values represent a shift in the direction required to reduce the cue discrepancy, and negative values represent a shift in the opposite direction (gray shaded region). The blue and red regression lines represent the linear fit of the visual and vestibular PSE shifts, respectively. Black dashed vertical lines mark the RRs used to sort the data by low, medium and high RR, and the solid vertical line marks RR=1 (equal reliability).



Supplemental Figure S2: Model 3 fit for all monkeys, related to Figure 6. Circles mark the average visual (red) and vestibular (blue) shifts and error bars represent the SEM. The data were grouped (from left to right) by low, medium and high visual to vestibular reliability ratios. Each group comprises three columns which depict (from left to right) unsupervised calibration, vestibular accurate (supervised calibration) and visual accurate (supervised calibration). Solid curves represent the model simulations as a function of calibration trials, which were fit simultaneously for all data values at 500 trials (vertical dashed lines). Horizontal dotted lines mark the externally accurate solution for the supervised conditions and zero shift for the unsupervised. 'N' indicates the number of session repeats. The parameter fits are presented in the rightmost column.



Supplemental Figure S3: Model 4 fit for all monkeys, related to Figure 6. All conventions are the same as Supplementary Figure S2, however here Model 4 is presented.



Supplemental Figure S4: Precalibration combined cue reliability. A scatter plot of the actual combined cue reliabilities (calculated from the psychometric curves) versus those predicted from the single cues (by Eq. 2) confirms that they were similar. The solid line represents the diagonal (y=x).

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Human experiment

Supervised multisensory calibration was tested in humans, in a similar way to the monkey experiments. The study was approved by the internal review board at Baylor College of Medicine and subjects signed informed consent. The apparatus have been previously described (Fetsch et al., 2009; Zaidel et al., 2011). In short, subjects were seated in a cockpit-style chair and restrained safely with a five-point racing harness. They wore a custom made thermoplastic mask which was attached to the back of the chair for head stabilization and active three-dimensional glasses (CrystalEyes 3; RealD, Beverly Hills, CA) in order to provide stereoscopic depth cues. A display screen (located ~70 cm in front of the eyes) and projector (Galaxy 6; Barco, Kortrijk, Belgium) moved together with the chair on top of a motion platform (6DOF2000E; Moog, East Aurora, NY) to provide synchronized visual and vestibular input.

The subjects' task was to discriminate heading direction (two-alternative forced-choice, right or left of straight ahead), after presentation of a single-interval stimulus. The stimulus velocity followed a 3-sigma Gaussian profile with duration 1s and total displacement 13cm. Peak velocity was 0.26 m/s and peak acceleration was 0.78 m/s^2. Subjects were instructed to get as many trials as they can correct and advised that they would receive feedback (as described below). Specifically, we tested the conditions in which we found cue yoking in the monkeys, i.e., when the less reliable cue was accurate.

Like the monkey experiments, each human session comprised three blocks (pre-calibration, calibration and post-calibration) and used the method of constant stimuli. *Pre-calibration*, no choice feedback was given (no beep was sounded after making a choice). This block comprised 10 repetitions of 9 logarithmically spaced headings (4 to each side and straight ahead) for visual and vestibular cues (180 trials in total). During *calibration*, only combined cues were presented, with a cue discrepancy ($\Delta=\pm10°$, as described above). Here auditory feedback was given through

headphones: a high tone signified a correct choice and a low tone signified an incorrect choice. Similar to the monkey experiments, feedback was contingent on one of the cues (visual or vestibular). This block comprised 25 repetitions of 10 logarithmically spaced headings (5 to each side; 250 trials in total). *Post-calibration*, a shift of the individual (visual/vestibular) cues was measured by single-cue trials, interleaved with the combined-cue trials (with $\Delta = \pm 10^{o}$ as in the calibration block). This block comprised 260 trials (180 single cue trials and 80 combined cue trials). In total each human session comprised 690 trials and lasted approximately 1 hour. Four conditions were tested: vestibular less reliable but accurate ($\Delta = \pm 10°$) and visual less reliable but accurate ($\Delta = \pm 10°$). We present data from four subjects (2 male) across the four conditions (N=16).

Multisensory calibration - theoretical background

When presented with a stimulus S, individual cue sensors, e.g., A and B, will provide noisy measurements of S: x_A and x_B, respectively. We assume that x_A and x_B follow Gaussian distributions with standard deviations σ_A and σ_B, respectively. When both cues are presented together, the combined-cue (C) maximum-likelihood measurement is given by:

$$x_C = \frac{\frac{x_A}{\sigma_A^2} + \frac{x_B}{\sigma_B^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}} \qquad \text{(Eq. 1).}$$

Equation 1 represents optimal multisensory integration, in that it minimizes the variance of the combined-cue measurement (maximum reliability), given by:

$$\sigma_C^2 = \frac{1}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2}} \qquad \text{(Eq. 2).}$$

We confirmed that the pre-calibration combined cue reliabilities in our data were similar to those predicted by Equation 2 (see Suppl. Fig. S4).

Typically, x_A and x_B are assumed to have means equal to the true stimulus: $\mu_A(S) = S$ and $\mu_B(S) = S$, and are thus considered "externally accurate". However, a cue may be biased (inaccurate), such that its mean measurement is systematically deviated from the true stimulus: $\mu_A(S) = S + D_A$ and $\mu_B(S) = S + D_B$, where D_A and D_B are the cue deviations, respectively. Similarly, the experimenter can instill in the observer an erroneous belief that a cue is biased by providing external feedback consistent with a deviated stimulus value. The observer might respond to this bias by calibrating the cue. Since this calibration is driven by external feedback, we refer to it as supervised calibration.

In the absence of feedback, external accuracy cannot be assessed. In that case, the best a system can do is to compare the single-cue measurements to one another. If the means of the single-cue measurements are equal, $\mu_A(S) = \mu_B(S)$, then they are considered 'internally consistent'. They may or may not be accurate, i.e., they could be equally biased. If, however, they are not equal, then they are not internally consistent, and at least one of the cues is not accurate. This cue discrepancy will result in unsupervised calibration of the cues towards one-another, presumably to achieve internal consistency (Burge et al., 2010; Zaidel et al., 2011).

We model unsupervised multisensory calibration (Fig. 1A in the main manuscript) through recursive, trial-by-trial changes to the single-cue measurement means. Suppose on the n^{th} trial, the single-cue measurements are $x_{A,n}$ and $x_{B,n}$. The trial's unsupervised calibration amounts are computed as:

$$\begin{aligned} \delta_{A,n} &= \eta_A\left(x_{B,n} - x_{A,n}\right) \\ \delta_{B,n} &= \eta_B\left(x_{A,n} - x_{B,n}\right) \end{aligned} \qquad \text{(Eq. 3)},$$

where η_A and η_B are the cues' respective rates of calibration. On the next trial, the single-cue measurement means will be updated by:

$$\begin{aligned} \mu_{A,n+1}(S) &= \mu_{A,n}(S) + \delta_{A,n} \\ \mu_{B,n+1}(S) &= \mu_{B,n}(S) + \delta_{B,n} \end{aligned} \qquad \text{(Eq. 4)}$$

.

Note that during unsupervised calibration the cues shift towards one-another, in opposite directions. This will eventually achieve internal consistency. But, it does not address cue accuracy.

By contrast, the aim of supervised calibration is to improve external accuracy. In this study, we tested supervised multisensory calibration. Feedback, provided after each trial, was binary. It indicated to the monkey whether his heading choice was correct or incorrect, and thereby whether the stimulus heading was rightward or leftward. We write FB = 1 for rightward, and FB = –1 for leftward feedback. Discrepancy between a sensory estimate and FB could be due either to miscalibration or sensory noise. The former, but not latter, would warrant calibration. Hence ideally one would like to be able to infer the cause of discrepancy (Körding et al., 2007). However, on a single trial it is impossible to disambiguate whether an error comes from cue miscalibration or sensory noise. Hence our modeling strategy is based on finding the most likely miscalibration amount, given FB, with trial-by trial increments. Below we present possible models for supervised cue calibration.

Supervised calibration Model 1

In our first model, we assume that the organism has access to each cue's measurement individually. Hence each cue can be assessed and calibrated independently. A good strategy for supervised cue calibration would therefore be to find the most likely value of each cue's miscalibration (deviation from external accuracy, D_A and D_B respectively) given FB, and then to calibrate the cues accordingly. To find this, we have to first calculate the probability of receiving FB given the cues' miscalibration amounts. For each cue this is given by:

$$p(FB\,|D_A) = \int_{FB} p(x_A \mid S, D_A)p(S)dS$$
$$p(FB\,|D_B) = \int_{FB} p(x_B \mid S, D_B)p(S)dS \qquad \text{(Eq. 5)},$$

where the integral is from $-\infty$ to 0 when FB = –1, and from 0 to ∞ when FB = 1. $p(x_A \mid S, D_A)$ and $p(x_B \mid S, D_B)$ are probability distributions (assumed normal) for the cues' measurements,

given the stimulus and cue deviations: $\sim N(S + D_A, \sigma_A)$ and $\sim N(S + D_B, \sigma_B)$, respectively. $p(S)$ is the stimulus prior, also assumed normal: $\sim N(0, \sigma_S)$.

Hence, given FB, the maximum-likelihood cue miscalibration amounts are:

$$\hat{D}_A = \underset{D_A}{\text{argmax}}\left(p(FB \mid D_A)\right)$$
$$\hat{D}_B = \underset{D_B}{\text{argmax}}\left(p(FB \mid D_B)\right) \qquad \text{(Eq. 6).}$$

We do not assume that the rate of calibration is the same for supervised vs. unsupervised calibration. Therefore, we include a parameter w^{sup}, to represent the relative weighting of supervised vs. unsupervised calibration. With single-cue measurements $x_{A,n}$ and $x_{B,n}$ on the n^{th} trial, the supervised calibration amounts (updated in the next trial by Eq. 4) are computed as:

$$\delta_{A,n} = -\eta_A w^{\text{sup}} \hat{D}_{A,n}$$
$$\delta_{B,n} = -\eta_B w^{\text{sup}} \hat{D}_{B,n} \qquad \text{(Eq. 7).}$$

Supervised calibration according to Model 1 is presented schematically in Figure 1B in the main manuscript.

Supervised calibration Model 2

As an alternative to Model 1 and to address the observed phenomenon of cue yoking in the data, we entertain the possibility that the organism only relies on the combined cue measurement for supervised calibration (Model 2). Accordingly, only x_C, but not x_A or x_B can be used. In this case, we calculate the most likely value of the combined cue's miscalibration (D_C), given FB:

$$p(FB \mid D_C) = \int_{FB} p(x_C \mid S, D_C)\, p(S)\, dS \qquad \text{(Eq. 8).}$$

Here too, the integral is from $-\infty$ to 0 when FB = – 1, and from 0 to ∞ when FB = 1. $p(x_C \mid S, D_C)$ is the probability distribution (assumed normal) of the combined cue measurement, given the stimulus and combined cue deviation: $\sim N(S + D_C, \sigma_C)$, and $p(S)$ is the stimulus prior (same as Eq. 5).

Similar to Equation 6, the maximum-likelihood combined-cue miscalibration is:

$$\hat{D}_C = \underset{D_C}{\operatorname{argmax}}\left(p\left(FB \mid D_C\right)\right) \qquad \text{(Eq. 9).}$$

Since it is the only available miscalibration information, both cues are calibrated according to $\hat{D}_C$. But we still allow each cue to shift according to its own rate of calibration. Hence, the n^{th} trial's supervised calibration amounts (updated in the next trial by Eq. 4) are computed as:

$$\begin{aligned} \delta_{A,n} &= -\eta_A w^{\text{sup}} \hat{D}_{C,n} \\ \delta_{B,n} &= -\eta_B w^{\text{sup}} \hat{D}_{C,n} \end{aligned} \qquad \text{(Eq. 10).}$$

The combined estimate, which is comprised of the individual cues (Eq. 1), will consequently also shift. Supervised calibration according to Model 2 will achieve and maintain accuracy of the combined cue. It, however, does not ensure accuracy of the individual cues. The individual cues can both be inaccurate whilst the combined cue is accurate. Supervised calibration according to Model 2 is presented schematically in Figure 1C in the main manuscript.

Supervised calibration Model 3

Our third model of supervised calibration, Model 3, is a hybrid of Model 2 and unsupervised calibration. Like Model 2, Model 3 assumes that the organism only relies on the combined cue measurement for supervised calibration. But, it maintains that in parallel to supervised calibration, the cues also shift towards one another to reduce their own cue discrepancy, like in unsupervised calibration. Hence, calibration is comprised of two components – one from Model 2 and the other from unsupervised cue calibration. According to Model 3, the n^{th} trial's calibration amounts (updated in the next trial by Eq. 4) are computed as:

$$\begin{aligned} \delta_{A,n} &= \eta_A\left(x_{B,n} - x_{A,n}\right) - \eta_A w^{\text{sup}} \hat{D}_{C,n} \\ \delta_{B,n} &= \eta_B\left(x_{A,n} - x_{B,n}\right) - \eta_B w^{\text{sup}} \hat{D}_{C,n} \end{aligned} \qquad \text{(Eq. 11),}$$

where the term on the left represents the unsupervised calibration component (taken from Eq. 3) and the term on the right represents the yoked, supervised calibration component (taken from Eq.

10). Calibration according to Model 3 is presented schematically in Figure 5 in the main manuscript.

The formulation of Model 3 presented above, in which the cues are calibrated simultaneously (in each iteration) by two components, supervised and unsupervised, implies that both components cause perceptual changes. An alternative approach for Model 3 could be that, while the unsupervised component causes perceptual changes (Zaidel et al., 2011), the supervised component can cause choice related, rather than perceptual, changes (e.g., a change in the right-left selection criterion). Accordingly, unsupervised calibration alone would represent the underlying perceptual shifts (Eq. 3). Here too, the overall PSE shifts would incorporate two superimposed components: the unsupervised, perceptual, shift and the supervised, choice related, shift. However, this approach has two differences: i) supervised calibration is common to both the cues (there would be one rate of supervised calibration, η^{sup}, instead of $\eta_A w^{\text{sup}}$ and $\eta_B w^{\text{sup}}$) and ii) the supervised and unsupervised processes would run separately, in parallel. Namely, the supervised choice related shift (updated iteratively from an initial value of zero by $\delta_n^{\text{sup}} = -\eta^{\text{sup}} \hat{D}_{C,n}$) would be superimposed on the individual cue perceptual shifts when calculating the PSEs, but it would not affect the underlying perceptual shifts (depicted by Eq. 3).

Supervised calibration Model 4

Here, we extended the Ernst and Di Luca model (2011) to include external feedback. We first explain how we implemented the model in its standard form (in which there is no external feedback and thus represents only unsupervised calibration) and then we explain how we extended it to incorporate feedback. Similar to our other models, calibration is simulated across trials. However, here, in each trial, integration and calibration take place in stages. In the first stage, a posterior distribution for the cue estimates is calculated by multiplying their likelihood function, a two-dimensional Gaussian, by a diagonal coupling prior, which indicates the prior

belief that the cues represent the same world property. Thus, the posterior distribution on the n^{th} trial is represented by:

$$P(S) \propto \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(S - X)^T \Sigma^{-1}(S - X)\right) \cdot \frac{1}{\sqrt{2\pi\sigma_{Sdiff}^2}} \exp\left(-\frac{(S_{A,n} - S_{B,n})^2}{2\sigma_{Sdiff}^2}\right) \text{(Eq. 12)},$$

where $S = [S_{A,n} \quad S_{B,n}]$ are the hypothesized values of the stimuli and $X = [x_{A,n} \quad x_{B,n}]$ are the measurements, for cues A and B on trial *n*. $\Sigma = \begin{bmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_B^2 \end{bmatrix}$ represents the cue measurement variances , and σ_{Sdiff}^2 represents the prior variance of the difference between the cues (the coupling prior along the diagonal). At this point, the estimate of cue discrepancy is given by the difference between the maximum a posteriori (MAP) estimates: $\hat{D}_n^{MAP} = \hat{S}_{A,n}^{MAP} - \hat{S}_{B,n}^{MAP}$.

In the next stage, "credit" for this discrepancy is assigned to each cue according to the *bias* prior. To achieve this, the cue biases are estimated from the bias prior, given the constraint that together they must account for $\hat{D}_n^{MAP}$. This constraint can be depicted by a diagonal line (or slice) across the bias prior. Assuming that the bias prior is a zero-mean, uncorrelated, two-dimensional Gaussian with cue bias variances $\sigma_{b,A}^2$ and $\sigma_{b,B}^2$, this slice is itself a 1-dimensional Gaussian, with maximum at: $\hat{D}_{A,n}^{MAP} = \hat{D}_n^{MAP} \cdot \frac{\sigma_{b,A}^2}{\sigma_{b,A}^2 + \sigma_{b,B}^2}$ and $\hat{D}_{B,n}^{MAP} = -\hat{D}_n^{MAP} \cdot \frac{\sigma_{b,B}^2}{\sigma_{b,A}^2 + \sigma_{b,B}^2}$. Note that this maximum is determined by the ratio of the cue bias variances and not their magnitudes (multiplying both $\sigma_{b,A}^2$ and $\sigma_{b,B}^2$ by a constant does not change the result). Finally, the cues are updated by their estimated miscalibration (substituting for $\hat{D}_n^{MAP}$) according to Eq. 4:

$$\delta_{A,n} = (\hat{S}_{B,n}^{MAP} - \hat{S}_{A,n}^{MAP}) \cdot \frac{\sigma_{b,A}^2}{\sigma_{b,A}^2 + \sigma_{b,B}^2}$$

$$\delta_{B,n} = (\hat{S}_{A,n}^{MAP} - \hat{S}_{B,n}^{MAP}) \cdot \frac{\sigma_{b,B}^2}{\sigma_{b,A}^2 + \sigma_{b,B}^2} \qquad \text{(Eq. 13).}$$

We would like to point out a few characteristics of the resulting unsupervised calibration: i) the ultimate point of convergence is independent of cue reliability, and determined entirely by the bias prior. ii) Only the ratio of the cue bias variance is needed to determine this point. iii) The rate of calibration is determined by the extent to which the cues are integrated, e.g., when the cues are highly integrated, $\hat{D}_n^{MAP}$ is small and thus the cue will be updated in small increments. Specifically, we found that a very high degree of integration is required in order to achieve calibration rates sufficiently low to account for our data.

We now address how we implemented the cue update based on feedback. According to the Ernst and Di Luca model, the combined cue estimate on trial *n* is given by: $\hat{S}_{C,n}^{MAP} = \hat{S}_{A,n}^{MAP} - \hat{D}_{A,n}^{MAP} \equiv \hat{S}_{B,n}^{MAP} - \hat{D}_{B,n}^{MAP}$. Hence, the choice on trial *n* is determined by: $choice = sign(\hat{S}_{C,n}^{MAP})$, where 1 signifies a rightward choice and –1, a leftward choice. This was compared to feedback (similarly, FB = 1 for rightward feedback, and FB = –1 for leftward feedback). If the choice was not equal to feedback, a component of supervised calibration updated the cues, in addition to the unsupervised component presented above.

The supervised "error" was calculated for each cue by its offset from the region consistent with feedback (namely, its distance to the closest cue estimate that would give the same feedback as that received). This is the cue estimate itself ($\hat{S}_{A,n}^{MAP}$ and $\hat{S}_{B,n}^{MAP}$). Accordingly, similar to our other models, the supervised component for updating the cues is represented by: $-\eta_A^{\sup}\hat{S}_{A,n}^{MAP}$ and $-\eta_B^{\sup}\hat{S}_{B,n}^{MAP}$, where $\eta_A^{\sup}$ and $\eta_B^{\sup}$ represent the supervised rates of calibration. Thus, together with the unsupervised component, the cue updates on trial *n* are represented by:

$$\delta_{A,n} = (\hat{S}_{B,n}^{MAP} - \hat{S}_{A,n}^{MAP}) \cdot \frac{\sigma_{b,A}^2}{\sigma_{b,A}^2 + \sigma_{b,B}^2} - \eta_A^{\sup} \hat{S}_{A,n}^{MAP}$$

$$\delta_{B,n} = (\hat{S}_{A,n}^{MAP} - \hat{S}_{B,n}^{MAP}) \cdot \frac{\sigma_{b,B}^2}{\sigma_{b,A}^2 + \sigma_{b,B}^2} - \eta_B^{\sup} \hat{S}_{B,n}^{MAP} \qquad \text{(Eq. 14).}$$

Model optimization

In order to compare the models quantitatively, we fit each to the data. All four models had 3 free parameters. For Models 1-3 these were η_{ves}, η_{vis} and $w^{\sup}$. η_{ves} and η_{vis} represent the cue calibration rates during unsupervised calibration, and $w^{\sup}$, the relative rate of supervised calibration compared to unsupervised calibration (the supervised calibration rates are simply represented by the unsupervised rates multiplied by $w^{\sup}$). For Model 4 the free parameters were σ_{Sdiff}^2 (width of the coupling prior), *VisRatio* (the relative rate of visual to vestibular calibration) and $\eta_{ves}^{\sup}$ (the vestibular rate during supervised calibration; the visual rate was: $\eta_{vis}^{\sup} = VisRatio \cdot \eta_{ves}^{\sup}$).

All models assume that cue calibration rates follow a set ratio such that each cue has an intrinsic degree of plasticity. This is represented in Models 1-3 by η_{ves} and η_{vis}, and in Model 4 by *VisRatio*. In both cases, these parameters determine the relative rate of visual to vestibular calibration both during unsupervised and supervised calibration. This is in accordance with our observation that the vestibular cue indeed shifts more than the visual cue, both during unsupervised and supervised calibration (Suppl. Data, above). In the alternative Model 3 approach, described above (that supervised calibration represents choice related changes) η_{ves} and η_{vis} only determine the unsupervised rates; the supervised rate is common to both cues and represented by $\eta^{\sup}$ (instead of $\eta_{ves} w^{\sup}$ and $\eta_{vis} w^{\sup}$; thus it too has 3 free parameters).

For the model simulations we generated random data to closely resemble the actual data: ten logarithmically spaced heading stimuli were used with the same values as those used in the experiments, and also according to the method of constant stimuli. σ_S was calculated from these headings (although when we varied σ_S we found that it had negligible effect on the results). Cue

reliabilities were taken from the data for each monkey individually as the inverse variance of the fitted psychometric functions (see section **Data analysis**). The models assume the observer has access to the individual cue reliabilities (or reliability ratio).

The model simulations were run up to 2500 calibration trials, in order to see how each model would evolve beyond the 500 calibration trials used in the experiment. However, model optimization took place at 500 trials. For each model, experimental condition and set of model parameters, simulated cue shifts were averaged over 50 random sequences. For each monkey, the model parameters were fit simultaneously across all their data, by minimizing the total sum of squared errors. Each monkey had at least 42 sessions in total, and therefore at least 84 data points (42 sessions ×2 cues) were used for each model fit (the number of session repeats is depicted for each monkey and paradigm in Supplementary Figures S2 and S3).

SUPPLEMENTAL REFERENCES

Burge J, Girshick AR, Banks MS (2010) Visual-haptic adaptation is determined by relative reliability. J Neurosci 30:7714-7721.

Ernst MO, Di Luca M (2011) Multisensory Perception: From Integration to Remapping. In: Sensory Cue Integration (Trommershauser J, Kording K, Landy MS, eds), pp 224-250. Oxford University Press.

Fetsch CR, Turner AH, DeAngelis GC, Angelaki DE (2009) Dynamic reweighting of visual and vestibular cues during self-motion perception. J Neurosci 29:15601-15612.

Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. PLoS one 2:e943.

Zaidel A, Turner AH, Angelaki DE (2011) Multisensory calibration is independent of cue reliability. J Neurosci 31:13949-13962.