Title: The interplay of uncertainty, relevance and learning influences auditory categorization

Authors: Janaki Sheth[1], Jared S. Collina[2], Eugenio Piasini[3], Konrad P. Kording[4,5], Yale E. Cohen[1,4,5], Maria N. Geffen[1,4,6]

Affiliations:
[1]Department of Otorhinolaryngology, University of Pennsylvania, Philadelphia, PA
[2]Neuroscience Graduate Group, University of Pennsylvania, Philadelphia, PA
[3]Department of Neuroscience, International School for Advanced Studies, Trieste, Italy
[4]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA
[5]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA
[6] Department of Neurology, University of Pennsylvania, Philadelphia, PA

Corresponding authors:
Maria N. Geffen, Ph.D.
Stemmler Hall G14
3450 Hamilton Walk
Philadelphia PA 19104
(215)898-0782
mgeffen@pennmedicine.upenn.edu

Abstract

Auditory perception requires categorizing sound sequences, such as speech, into classes, such as syllables. Auditory categorization depends not only on the sequences' acoustic waveform, but also on the listener's sensory uncertainty, any individual sound's relevance to the task, and learning the temporal statistics of the acoustic environment. However, whether and how these factors interact to shape categorization is unknown. Here, we measured human participants' performance on a multi-tone categorization task. Task-relevant tones contributed more to category choice than task-irrelevant tones, confirming that participants combined information about sensory features with task relevance. Conversely, poor estimates of tones' task relevance or high sensory uncertainty adversely impacted category choice. Learning temporal statistics of sound category also affected decisions. The magnitude of this effect correlated inversely with participants' relevance estimates. These results demonstrate that humans differentially weigh sensory uncertainty, task relevance and learning, providing a novel understanding of sensory decision-making under real-life behavioral demands.

Main

Making sensory decisions, such as identifying what a friend just said, is a complex task due to the presence of multiple forms of uncertainty[1–8] in everyday settings, for example if the conversation took place at a cocktail party versus in a quiet room. One form, called *sensory uncertainty,* relates to sensory transduction processes and neural noise[9–12]. Another form of uncertainty is *stimulus relevance*: of all stimuli within a sensory scene, only some are relevant to the current decision[13–16]. For example, when trying to identify particular sounds uttered by our friend, we have to disregard sounds uttered by other speakers, which we consider as distractors.

When presented with uncertainty, observers may rely on their *knowledge of past stimulus statistics*[17–22] to make their decisions, such as their long-term knowledge of how their friend typically talks or their short-term knowledge of what their friend has recently said[23]. Indeed, observers can learn both short- [24,25] and long-term stimulus statistics[26–28] and this learning is considered crucial for efficient decision making[29,30].

Whereas the effects of these four factors – sensory uncertainty, stimulus relevance and short- and long-term learning – on sensory decision-making, such as categorization, have previously been studied separately, in everyday situations we often encounter them at the same time. Thus, in fact, these factors may interact with each other, differentially affecting our category decisions. For example, at the cocktail party, sensory uncertainty may not only decrease our ability to categorize a sound uttered by our friend, but may also affect our estimate of the relevant versus the distractor sounds. Our expectation of what our friend is saying may play a greater role when our sensory uncertainty is high or when there are more distracting sounds by other speakers. These expectations may be reduced if we are talking to a person we just met.

To test whether and how the interplay of these four factors shapes sensory decisions, we measured the performance of human participants on an auditory categorization task[31–34]. More specifically, our goals were to test how observers' sensory uncertainty and associated estimates of the likelihood of the irrelevant stimuli mediated their sensitivity to stimulus relevance, and also how their task performance was affected by learning. Participants categorized temporal sequences of three tones, as either high or low in a two-alternative forced choice task, as we varied the number of task-irrelevant (distractor) and task-relevant (category-specific) tones across trials. Additionally, to test learning, in a subset of experimental sessions, we modulated the proportion of trials biased toward the high or the low category.

Because it has been suggested that the human brain optimally integrates information from relevant inputs while ignoring the irrelevant ones[35], we characterized participants' behavior using normative Bayesian models of sensory decision-making[36–41]. Bayesian formalism is a well-established framework to understand complex interactions among many factors governing decision-making[13,14,26,27,42]. On average, when categorizing auditory stimuli, participants integrated information across all three tones but weighed them by their task relevance. Individual participants' sensitivity to task relevance depended on their internal sensory uncertainty and subjective estimates of the likelihood of irrelevant stimuli. We also found population-wide evidence of both

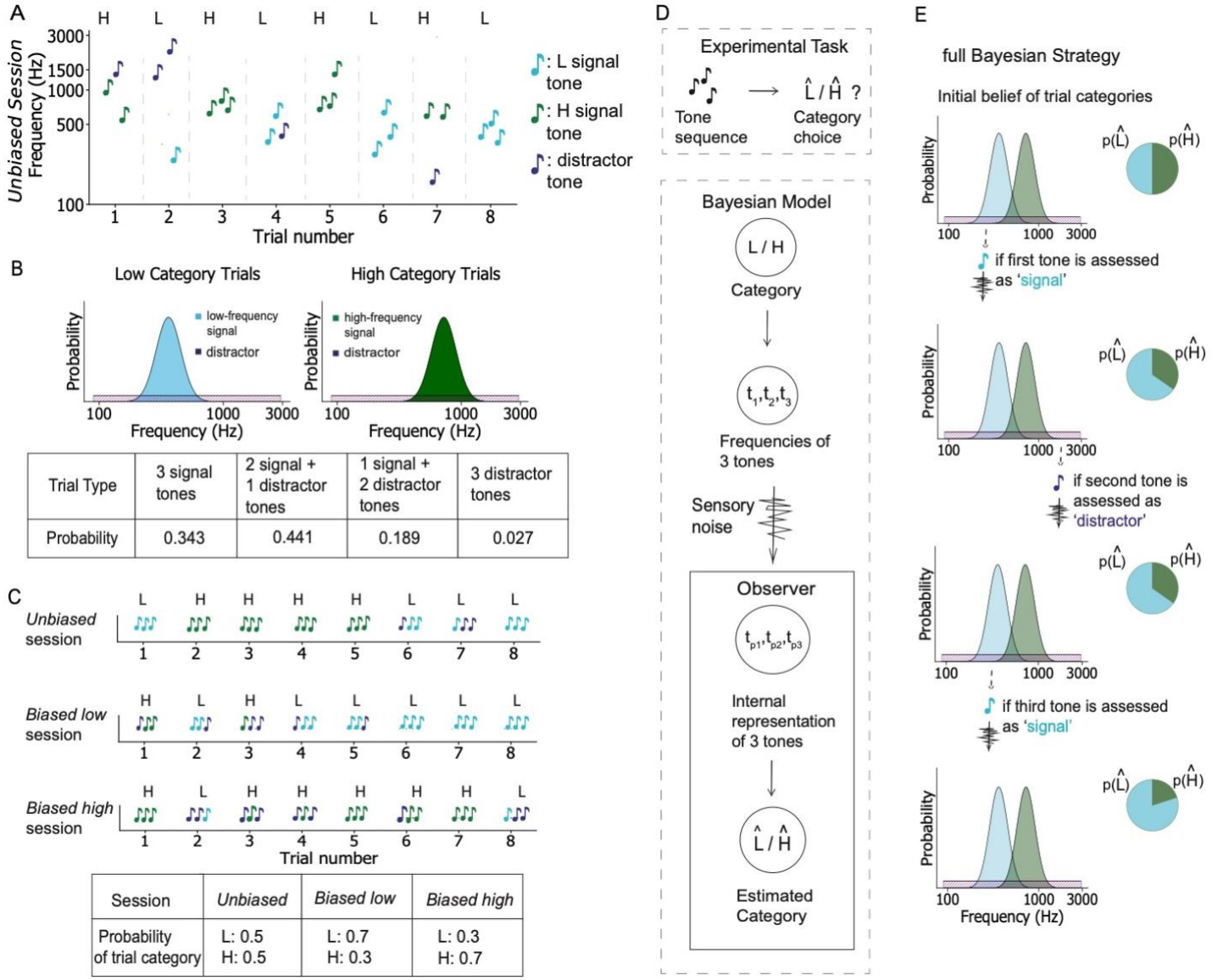
short-term learning (i.e., stimulus category of the previous trial) and long-term learning (i.e., session-level bias toward high versus low category). However, individual participants traded off their short- and long-term learning with their estimates of relevance. Together, our results demonstrate that humans differentially combine information about stimulus uncertainty, relevance, and knowledge of short-term and long-term statistics, providing a new framework for understanding mechanisms of auditory processing in real life.

Results

Mapping out decision strategies for auditory categorization

Auditory categorization can depend on multiple factors, such as sensory uncertainty, stimulus relevance and temporal statistics of the sound stimulus. These factors can co-occur and therefore may interact with each other to affect decision-making. Here, we designed a categorization task to measure how these factors interacted. In a two-alternative forced-choice auditory task, participants were asked to report the category of a three-tone sequence (Fig. 1A). Each trial was randomly selected to be from the low or high category, and each tone of the trial could probabilistically either be signal or distractor. The frequency of signal tones was drawn at random from a category-specific Gaussian probability distribution, whereas the frequency of distractor tones was drawn from a uniform probability distribution (Fig. 1A-B). To test the effect of long-term stimulus statistics on categorization, on a subset of sessions, we biased the distribution of categories. Trials in an *unbiased* session were equally likely to be drawn from either category. In a *biased low* session, the low-category trials were overrepresented, and vice versa for a *biased high* session (Fig. 1C). This stimulus design allowed us to compute a measure of the participants' sensory uncertainty, stimulus relevance and their learning of short-term (effect of preceding trial) and long-term (biased versus unbiased sessions) stimulus statistics.

We developed a general framework to capture different strategies that a participant could use to classify the three-tone sequences[36,43–45]. These strategies depend on a participant's internal model of stimulus relevance of the three tone frequencies and can be formalized using the Bayesian approach (Fig. 1D). The full Bayesian model reflects the optimal decision-making strategy for a participant (Fig. 1E). This model estimates the probability of relevance assigned by the participant to each of the three tones and integrates information across the tones accordingly. It also captures the participant's prior beliefs of the trial categories. As alternate models, we considered two reduced (suboptimal) variants of this model. In the no-distractor model, the participant assumed that all three tones were 'signal'. Conversely, in the random-guess model, the participant assumed that all three tones were irrelevant 'distractors' and made their choice randomly akin to a coin flip based on only their learned prior beliefs. Across all three sessions, *unbiased, biased low* and *biased high,* we interpreted our experimental findings using this three-model Bayesian framework.



Trial Type	3 signal tones	2 signal + 1 distractor tones	1 signal + 2 distractor tones	3 distractor tones
Probability	0.343	0.441	0.189	0.027

Session	*Unbiased*	*Biased low*	*Biased high*
Probability of trial category	L: 0.5 H: 0.5	L: 0.7 H: 0.3	L: 0.3 H: 0.7

Figure 1. Schematic diagram of the auditory categorization task and the optimal decision-making strategy.
(A) Example trials from an *unbiased* session. The stimuli in each trial are three tones denoted by notes: signal tones from the low-frequency distribution (light blue), signal tones from the high-frequency distribution (dark green), and distractor tones from the uniform distribution (purple). L: low category trial, H: high category trial. (B) Underlying probability distributions for signal (low-frequency Gaussian: light blue; high-frequency Gaussian: dark green) and distractor (uniform: purple hatched) tones. (Table) All combinations of signal and distractor tones, and their corresponding probabilities. (C) Example stimuli from *unbiased, biased low,* and *biased high* trials. (Table) Respective probabilities of the two trial categories for each session. (D) For a given low- or high- category trial with tone frequencies t_1, t_2, t_3, a participant generates internal representations t_{p1}, t_{p2}, t_{p3}, which are corrupted by sensory noise. From these representations they compute their category choice $\widehat{Low}$ ($\hat{L}$) or $\widehat{High}$ ($\hat{H}$). (E) A schematic of the full Bayesian strategy: the participant considers that any tone can probabilistically either be 'signal' or 'distractor' (see main text for more information). Using a hypothetical trial, we illustrate that deciding that a tone is 'signal' versus 'distractor' shifts participants' category choice probabilities. $\hat{L}$: $\widehat{Low}$-, $\hat{H}$: $\widehat{High}$ - category decision.

Participants integrate information over tones weighing them by stimulus relevance

Before we analyze the effect of the different factors on categorization, we first tested whether the position of a tone within each trial (e.g., first, second or third) affected participants' decisions. We computed the normalized correlation (see Methods) of each of the three tone frequencies with category choice probability

$p(\widehat{\text{High}})$ for all trials in the *unbiased* session and found that these correlations were independent of tone positions (Fig. 2A; table 1, one-way rmANOVA). This analysis suggests that participants weighed all three tone positions equally within a trial and integrated information across all three tones when making their category choice, and therefore our analyses could treat all three tone positions the same.

We then tested whether when combining information across the three tones, participants took into account stimulus relevance. Compared to signal tones, distractors should have less of an impact on categorization. We computed the correlation of tone frequency with category choice probability for signal versus distractor tones. On average across the population, signal tones were more strongly correlated with participants' performance than distractor tones (Fig. 2B; two-sided paired t-test, trials with one distractor: p = 1.32e-22; trials with two distractors: p = 5.36e-18; extended data Figs. 1A-C and table 1). This is also corroborated by individual participant data. Category choice probability of nearly all participants was more strongly correlated with signal tones than with distractor tones for trials with one (Fig. 2C; N = 56) or two distractors (Fig. 2D; N = 55; extended data Figs. 1D-G).

We next tested whether the influence of distractor versus signal tones depended on the number of distractors. Indeed, distractor tones were more influential whereas signal tones were less influential with increase in number of distractors (Fig. 2B; extended data Figs. 1A-C; one-way rmANOVA, significant main effect for number of distractors on correlations with signal tone frequencies: $F(2,110)$ = 34.55, p = 2.28e-12, and on correlations with distractor tone frequencies: $F(1,55)$ = 24.36, p = 8e-6). These results indicate that participants weighed tones according to their relevance to the task.

Accordingly, we anticipated participants' accuracy to be low on trials with distractor tones. Indeed, average accuracy was inversely correlated with the number of distractors (Fig. 2E; N = 56, Spearman, ρ = -1, p = 0). An analysis as a function of trial category revealed similar effects (extended data Fig. 1H). Examination of individual participant accuracy further confirmed this result; nearly all participants (N = 54) were less accurate on trials with two distractors (Fig. 2F; Median (Mdn) = 66.24, Inter-quartile range (IQR) = 6.03) than on trials with one distractor (Mdn = 79.46, IQR = 11.26, two-sided Wilcoxon signed-rank test, Z = -6.39, p = 1.63e-10). These findings cumulatively demonstrate that distractors adversely impacted participants' categorization accuracy.

Next, we tested whether the effect of distractors depended on their tone frequency. A distractor tone whose frequency was similar to the trial category should impair participants' decisions less than a distractor tone whose frequency was different. We evaluated this prediction by computing the average participant accuracy for trials with one distractor tone. Category accuracy for these trials was significantly higher when the frequency of the distractor was similar to the trial category (Fig. 2G; Table 1 – Wilcoxon signed-rank test). For trials in which the distractor frequency was different than the trial category, we plotted separate psychometric curves corresponding to the signal or distractor tone frequencies (extended data Fig. 1I). Qualitative comparison of these curves to the psychometric curve for trials with only signal tones (extended data Fig. 1A) indicates that distractors have a substantial effect on participants' category choice. Combined, these data suggest that participants' ability to differentiate between signal and distractor tones depends on the frequency value of the distractors.

Finally, we explored whether the effect of distractors on categorization behavior varied significantly across the participants. This diversity may be due to participants' distinct internal priors, estimates of relevance or their sensory uncertainty. Indeed, there was significant heterogeneity in both the correlation of participants' category choice probability with different tone types (Figs. 2C, D) and in their accuracy (Fig. 2F). Accuracy for trials with one distractor tone ranged from 57% - 89% and for trials with two distractor tones ranged from 53% - 75%. Thus, in subsequent analyses, we not only quantified the effects of relevance, sensory uncertainty, and learning of stimulus statistics on the average behavior across the population, but also on inter-participant variability.

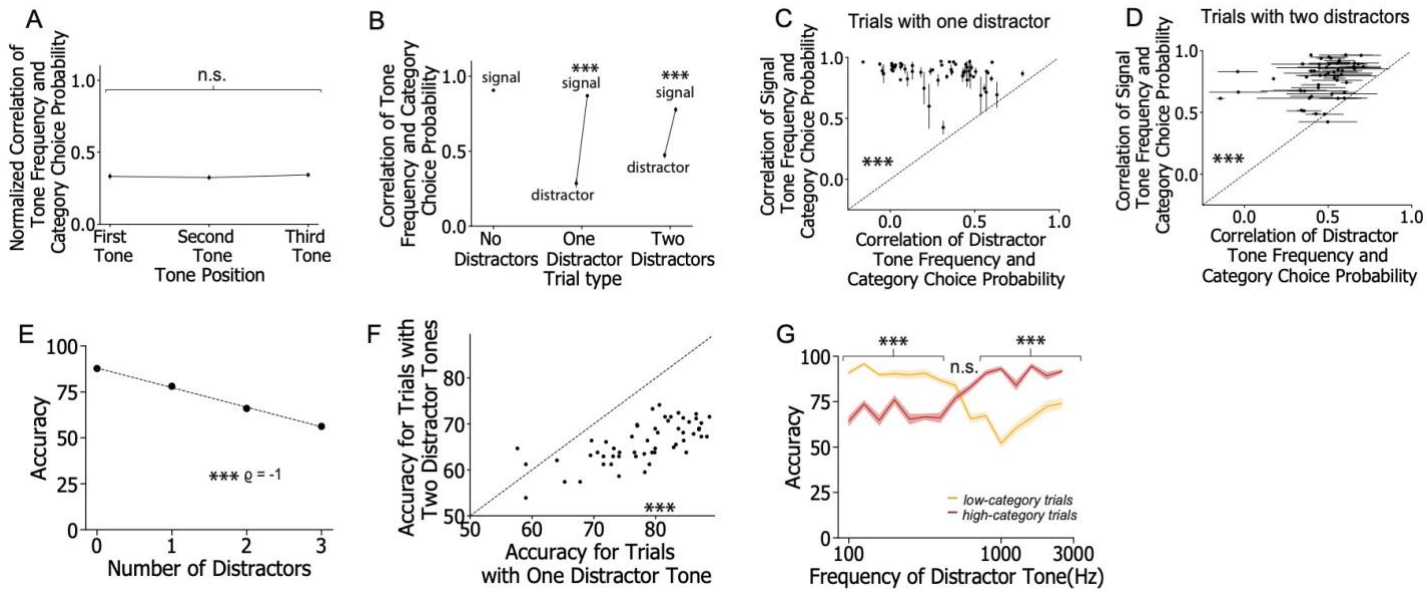


Figure 2. Participants integrated information over the three tone-sequence weighting tones according to their relevance. (A) In *unbiased* trials, on average, tone frequencies in each of the three temporal positions correlated equally with participants' category choice probability. (B) For trials with only signal tones, the frequencies of each tone were equally correlated with category choice probability. As the number of distractors increased, the correlation of signal tones decreased whereas that of distractor tones increased, but the signal tone frequencies were consistently more correlated with category choice probability. Error bars: standard error of the mean. *: two-sided paired t-test between correlations of signal and distractor tone frequencies with category choice probability. (C, D) Correlation of category choice probability with signal tone frequencies versus distractor tone frequencies. (C) Comparison of correlations for trials with one distractor. *: two-tailed Wilcoxon sign-rank test for the 56 participants with greater signal tone frequency correlation. (D) Comparison of correlations for trials with two distractors. *: two-tailed Wilcoxon sign-rank test for the 51 participants with greater signal tone frequency correlation. (E) Average accuracy across all 56 participants in *unbiased* trials decreased with the number of distractor tones. Error bars: standard error of the mean. *: Spearman correlation. (F) Categorization accuracy for the tone sequences with two versus one distractor tones. *: two-tailed Wilcoxon sign-rank test for the 54 participants with greater accuracy in trials with one distractor versus in trials with two distractors. Error bars: standard error of the mean obtained by assuming that a participant's accuracy follows a binomial distribution; not large enough to be distinguishable. (G) For trials with exactly one distractor tone, accuracy was higher when the distractor tone frequency was similar to the trial category. Binned frequencies of the distractor are shown on the x-axis. Error bars: standard error of the mean. *: two-tailed Wilcoxon sign-rank test between accuracy in high- and low-category trials computed at each distractor frequency. $^{\dagger}p<0.05$, $*p<0.01$, $**p<0.001$, $***p<0.0001$.

Sensory uncertainty and estimated task relevance underlie inter-participant variability

So far, we have demonstrated in a model-free way that when making decisions, participants across the population accounted for the relevance of the different tone frequencies as they integrated information from the three tones. Next, to investigate how these factors shaped *individual* participant's category decisions, we modeled each participant's behavior using the Bayesian approach.

On a standard 2AFC frequency categorization task with no distractor tones, tone frequencies at the intersection of the two gaussian distributions do not provide evidence for either category. Additionally, as the frequencies move further away from the decision boundary, tones become more informative about the trial's category (Fig. 3A, left) and consequently, a participant's accuracy increases the further the tone frequencies are from the decision boundary. Thus, in the absence of distractors, if a trial consisted of a single tone, a participant's choice probability would be modeled by a standard psychometric curve which is a monotonically increasing saturating sigmoidal response function. If the trial consisted of multiple tones, choice probability could be modeled as the product of the relevant sigmoidal single-tone psychometric curves. However, in the presence of distractor tones, participants' performance should be substantially different. Since tones whose frequencies are far from the category-specific Gaussians are more likely to be distractors than signal tones based on their

underlying probability distributions (Fig. 3A, right), participants should assign lower relevance to those tones. This effect should result in reduced choice performance for tones on the flanks of the signal distributions.

We first examined the performance of a representative participant with high accuracy on trials in the presence of distractors (Fig. 3B, one-distractor accuracy 80.59% and two-distractor accuracy 74.14%). Tones that were far from the relevant categories and thus more likely to be distractors were correlated less strongly with behavioral performance than tones closer to the middle of category distributions (Fig. 3B, black). Additionally, the optimal full Bayesian model (Fig. 3B, red) fit the participant's nonmonotonic behavior considerably better than the no-distractor or random-guess models (Fig. 3B, teal and salmon, respectively; Table 1 – two-sided Wilcoxon signed-rank test). This analysis demonstrated that this participant considers the relevance of the different tones when categorizing trials.

This finding is also supported by an alternative simpler model, a generalized linear model (GLM; see Methods). The GLM assigns weights to each of the tone frequencies (extended data Fig. 2A) based on a participant's category decisions. For this representative high-accuracy participant, weights for tones far from the means of the two categories were ~ 0 (extended data Fig. 2E), whereas those for tones that were near the category means had non-zero values.

We next studied the performance of two participants with intermediate and low accuracy. Their choice probability for tones drawn from the flanks (highest or lowest frequencies) of the signal distributions changed less in comparison to the participant with higher accuracy, resulting in psychometric curves that were more sigmoidal in nature (Figs. 3C, D). Furthermore, the difference in accuracy on trials with one and two distractors was lower for the low-accuracy than high-accuracy participants (Fig. 3C: one-distractor accuracy 78.24% and two-distractor accuracy 59.48%, Fig. 3D: one-distractor accuracy 59% and two-distractor accuracy 61.21%). This observation was substantiated by analysis from the corresponding GLMs (extended data Figs. 2F, G). For participants with lower accuracy, absolute values of the weights for tones far from the means of the two categories were more comparable to the weights for tones near the category means. However, even for these participants, although the effect was smaller, the full Bayesian model provided a better fit to the psychometric curves compared to the other two suboptimal models (red curve versus the teal and salmon curves; Table 1 – two-sided Wilcoxon signed-rank test).

These examples suggest that all participants, irrespective of individual accuracy, take into account stimulus relevance when choosing the stimulus category, albeit to varying degrees. Based on sampling these three individual participants (for more see extended data Fig. 3; Table 1 – two-sided Wilcoxon signed-rank tests), we tested how, across the entire population, discriminating relevance of different stimuli affected participants' category choice responses. Plotting the average psychometric curve and the average model fits across all participants (Fig. 3E), we found that there was a nonmonotonicity in the data that was well fit by the full Bayesian model, which qualitatively illustrates a population-wide effect of stimulus relevance on category choice decisions.

To quantify the extent to which stimulus relevance affects the population's choice responses, we compared the statistics for the model fits. Because the number of fitting parameters differs across the three models (6 in the full Bayesian, 5 in the no-distractor, and 2 in the random-guess model (see Methods and extended data Fig. 4)), we used Bayesian Information Criterion (BIC) scores to compare these different models. A lower BIC score indicates a better model fit. For all participants, the full Bayesian model was better than the random-guess model (Fig. 4A; population statistics: two-sided Wilcoxon signed-rank test on BIC, $Z = -6.51$, $p = 7.55e-11$) at describing categorization behavior. Additionally, we found that for most participants, except for two, it was also better than the no-distractor model (Fig. 4B; same test, $Z = -6.49$, $p = 8.41e-11$). For the two outlier participants (Fig. 4B, C; light blue), the full Bayesian and no-distractor model fits were similar (Table 1 – two-sided Wilcoxon sign-rank tests), indicating that they did not necessarily account for relevance when making decisions. However, for the rest, the full Bayesian model reproduced the main patterns in the data in an absolute sense because it accurately captured participants' responses across tone frequencies (Figs. 3B-E,

extended data Figs. 3A,D,G). These results confirm that nearly all participants took stimulus relevance into account when making category decisions.

Next, we correlated model fits to participants' task accuracy. Across the population, difference in BICs of the full Bayesian and the no-distractor models was inversely correlated with accuracy (Fig. 4C, Spearman, ρ = -0.85, p = 8.53e-17). For highly accurate participants, the full Bayesian model performed particularly well. These data, once again, suggest that participants' behavioral performance was determined by their ability to discriminate tones based on their decision relevance.

Participants' ability to discriminate between signal and distractor tones may depend both on their subjective estimate of task relevance (or more specifically, their estimate of the probability of distractors) as well as on their sensory uncertainty. For instance, some participants may have different prior convictions about the number of distractors (e.g., some may spend more time at noisy cocktail parties than others) which could bias how they integrate information over the three tones. Additionally, we expect that participants with low sensory uncertainty would distinguish better between signal and distractor tone frequencies than participants with high sensory uncertainty. We conjectured that both estimated priors for the probability of distractors and, sensory uncertainty would drive participants' category choice and task accuracy contributing to inter-individual variability.

To test the effect of the two factors on category choice, we defined a metric 'sigmoidicity' (Fig. 4D; see Methods) which quantitatively captured participant's category decisions across all trials. We found that this metric was significantly influenced by the model parameters $p_{distractor}$ (participant's estimated probability of distractors) and $\sigma_{sensory}$ (participant's sensory uncertainty) as well as by the interaction between the two parameters (Fig. 4E; Table 1 – two-factor model). Specifically, 'sigmoidicity' increased with $\sigma_{sensory}$ and decreased with $p_{distractor}$. We next tested the effect of $p_{distractor}$ and $\sigma_{sensory}$, on participants' accuracy. Although we did not find a relationship between accuracy and $p_{distractor}$ (Fig. 4F; Table 1 – two-factor model), we found a strong, roughly linear, relationship between accuracy and $\sigma_{sensory}$ (Spearman, ρ = -0.96, p = 3.15e-31). Taken together, these data demonstrate that participants' category decisions were shaped by both their estimate for the probability of distractors and their sensory uncertainty. However, much of the differences in accuracy across participants may be driven by differences in their ability to distinguish between tone frequencies.

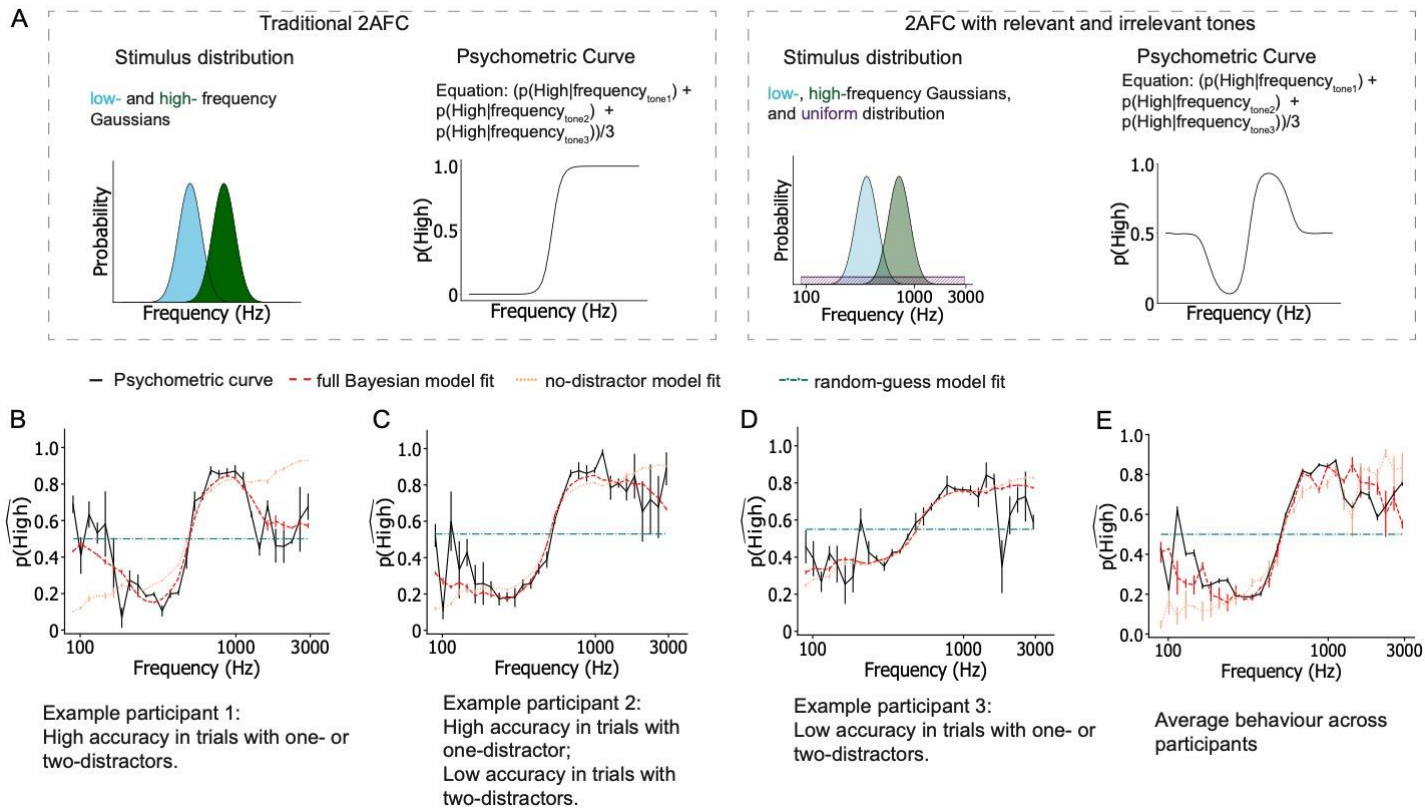


Figure 3. Psychometric curves for 3 representative participants and averaged across all participants.
(A) Cartoon models of different psychometric curves. If a trial contained only signal tones that were drawn from either the low- or high- frequency signal Gaussian distributions, the psychometric curve would be a sigmoid as in the traditional 2AFC task. However, if the tones can be either signal or distractor such that the stimulus distribution comprises of the two Gaussians and an additional uniform distribution, the psychometric curve becomes 'S'-shaped. Further, this curve plateaus at 0.5 because frequency values at the extremes are more likely to be distractors than signal tones. Because a given frequency can be played in any order (as the 1st, 2nd or 3rd tone) in a trial sequence and all three tone positions are weighed equally (Fig. 2A), we averaged the three resulting $p(\widehat{High})$ values. Error bars: standard error of the mean. (B, C, D) Psychometric curves (black lines) and respective model fits (full Bayesian model: red dashed lines, no-distractor model: salmon dashed lines, random-guess model: teal dashed lines) for 3 example participants (for posteriors of the full Bayesian model, and for GLM analysis, see extended data Fig. 2; for more participants see extended data Fig. 3). The x-axes of the psychometric curves span the frequency range of the stimulus tones, and the y-axes denote the mean probability of a $\widehat{High}$ category choice response. Error bars: standard error of the mean. (B) Example participant with high task accuracy on trials with either one or two distractor tones. (C) Example participant with high task accuracy on trials with one distractor, and poor accuracy on trials with two distractors. (D) Example participant with poor accuracy on trials with both one or two distractors. (E) Psychometric curve and model fits averaged across all participants.

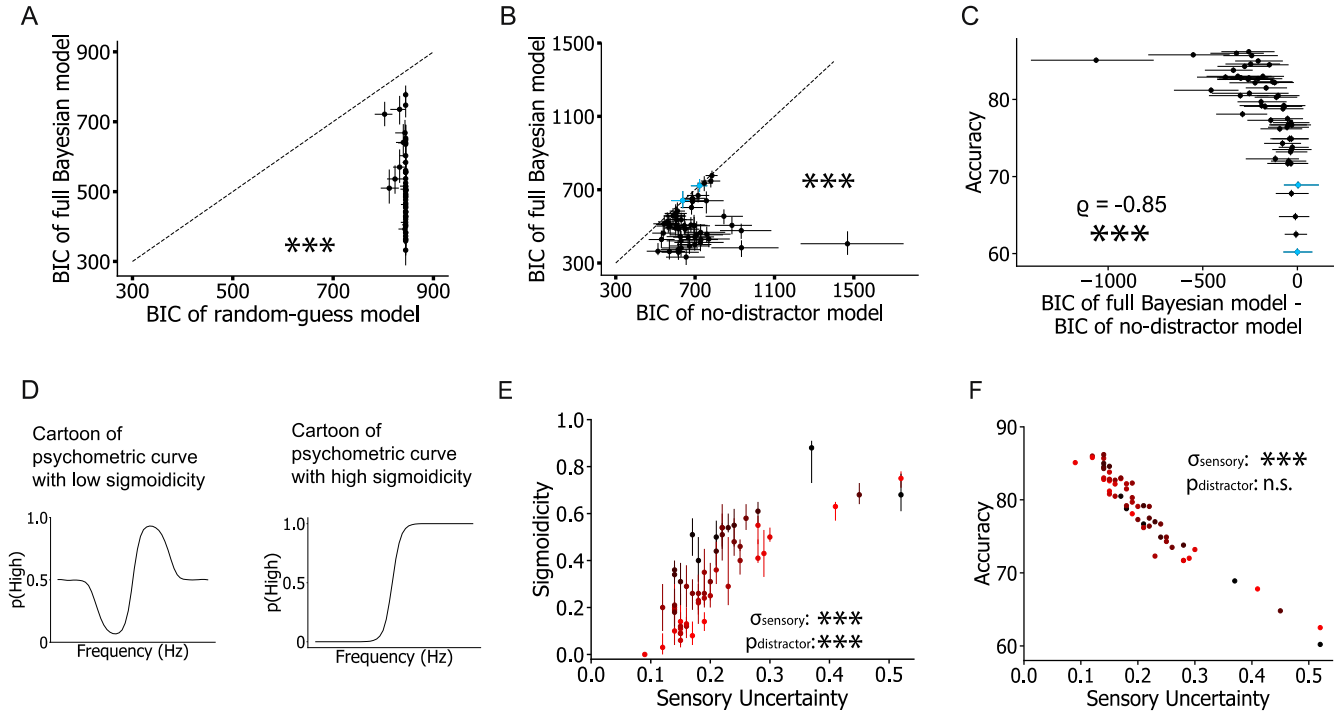


Figure 4. Nearly all participants were sensitive to stimulus relevance during decision-making; for an individual participant their category choice was driven by their sensory uncertainty and their subjective estimates of the probability of distractors but their accuracy was mostly driven by their sensory uncertainty.

On *unbiased* trials, BIC scores of the full Bayesian model were lower than that of the (A) random-guess model for all participants and (B) the no-distractor model for all except two participants. The two outliers are shown in blue. *: two-tailed Wilcoxon sign-rank test on the population. (C) Participants' accuracy was inversely correlated with difference in BIC scores between the full Bayesian and no-distractor models. *: Spearman correlation. (D) Cartoon models of different psychometric curves illustrate the qualitative range of sigmoidicty. (E) Each data point is a participant; its hue is given by $p_{distractor}$ which is the participant's estimated probability of distractors. It is a fitting parameter in the full Bayesian model and has value 0 in the no-distractor model. Black: $p_{distractor} \sim 0$; Red: $p_{distractor} \sim 1$. The model parameter $\sigma_{sensory}$ captures participant's sensory uncertainty. *: effect of $p_{distractor}$ and $\sigma_{sensory}$ on sigmoidicity using a two-factor model. (F) Participants' accuracy did not significantly depend on $p_{distractor}$ but is inversely correlated with their sensory uncertainty. *: two-factor model. In all panels, data point: one participant; $^{\dagger}p<0.05$; $*p<0.01$, $**p<0.001$, $***p<0.0001$. Error bars in (A, B, E) and horizontal error bars in (C, F): 95% confidence intervals derived from respective model fits to 100 bootstrapped samples. Vertical error bars in (C, F): standard error of the mean, obtained by assuming that accuracy follows a binomial distribution; vertical error bars not large enough to be distinguishable.

Participants rely on past statistics of category probabilities

Having established how stimulus relevance and sensory uncertainty shape auditory categorization, we next asked whether and how participants' learning of the past statistics of the category probabilities informed their current decision. We first analyzed the influence of short-term statistics by computing for *unbiased* sessions the mean probability of a $\widehat{\text{High}}$ category choice response for a given trial t conditioned on the previous trial $t-1$, $\mathrm{p}(\widehat{\mathrm{H}_t}|\mathrm{H}_{t-1})$ and $\mathrm{p}(\widehat{\mathrm{H}_t}|\mathrm{L}_{t-1})$ (Figs. 5A-C). We found that the difference between the two curves was quite variable across participants, suggesting that although some participants' decisions were substantially affected by the short-term history of category probabilities (Figs. 5B, C), others' decisions were not (Fig. 5A). However, on average, the previous trial (i.e., short-term stimulus statistics) significantly influenced participants' choices on the current trial (Fig. 5D; $\Delta_{\text{short-term}}$: difference between $\mathrm{p}(\widehat{\mathrm{H}_t}|\mathrm{H}_{t-1})$ and $\mathrm{p}(\widehat{\mathrm{H}_t}|\mathrm{L}_{t-1})$ at central frequency; two-sided Wilcoxon signed-rank test on $\Delta_{\text{short-term}}$, Z = 6.04, p = 1.58e-9). This confirms the presence of short-term learning in category decision-making.

In conjunction with short-term learning, participants may also use long-term learning when there is a bias in the trial categories toward low or high choices over a longer time scale. To test whether participants were affected by only one of the two types of learning or by a combination of both, we changed the probability of the low category trials, $p_L = 0.5$ to $p_L = 0.7$ in *biased low* sessions and vice versa, to $p_L = 0.3$ in *biased high* sessions (Participants: N = 53 in *biased low* and N = 48 jointly in *biased low* and *biased high*). Anecdotally, individual participants with high relevance discrimination did not use either type of learning when making decisions (Fig. 5E; extended data Fig. 5A). However, participants with lower relevance discrimination seemed to use previous trial statistics to varying degrees (Figs. 5F, G). Some participants used both short- and long-term learning (Fig. 5F; extended data Fig. 5B), whereas others primarily used long-term learning (Fig. 5G; extended data Fig. 5C). Overall, in the population, there was a clear effect of both the session type and the last trial on category choice (Fig. 5H, extended data Fig. 5; $\Delta_{\text{long-term}}$: difference between psychometric curves from *biased low* and *biased high* sessions at central frequency; population statistics: two-sided Wilcoxon signed-rank test on $\Delta_{\text{long-term}}$, Z = 8.63, p = 6.34e-18; two-sided Wilcoxon signed-rank test on $\Delta_{\text{short-term}}$, *biased low*: Z = 5.94, p = 2.91e-9; *biased high*: Z = 5.94, p = 2.91e-9). Thus, across the three sessions, short- and long-term learning influenced participants' decision-making presumably depending on their relevance estimate.

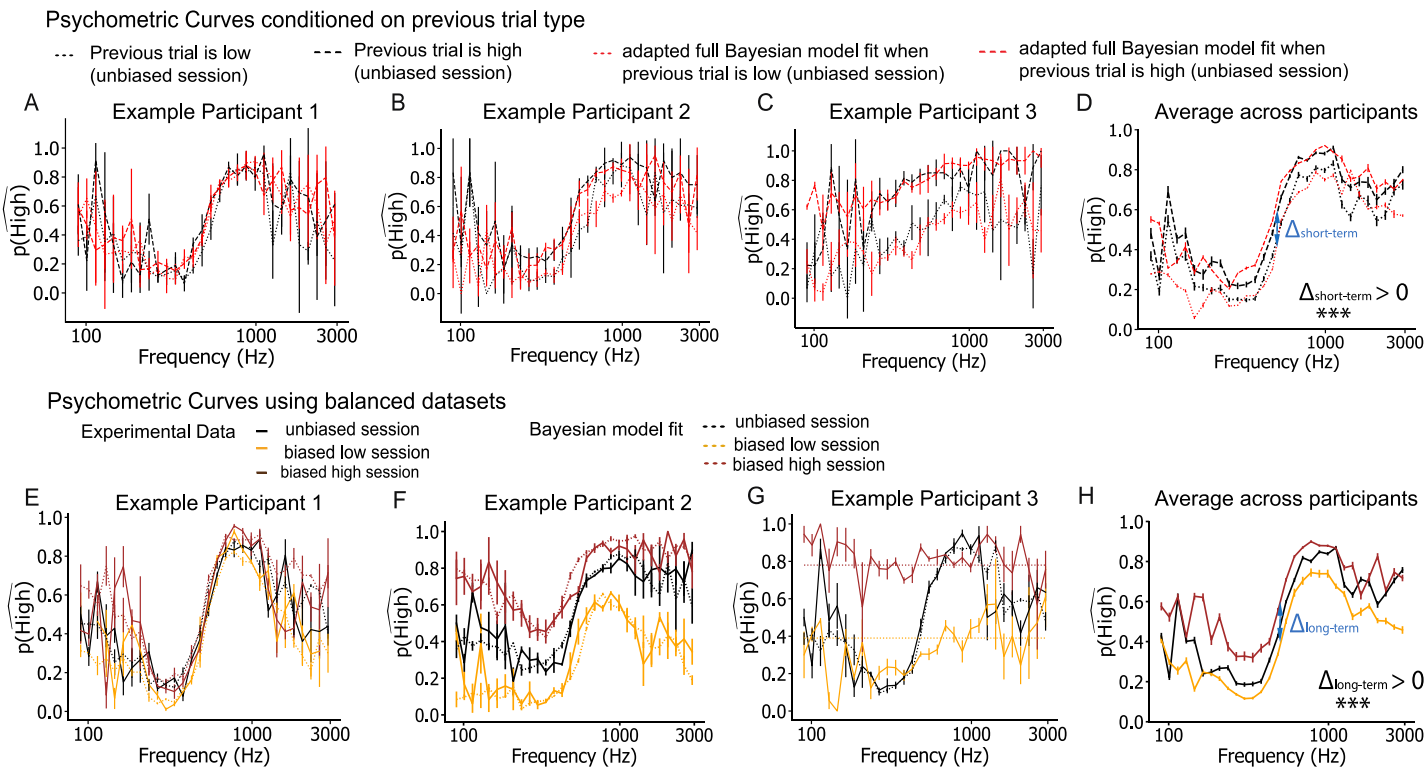


Figure 5: To a varying degree, participants incorporated past statistics of category probabilities in their decisions. (A-C) Average psychometric curves (black) and full Bayesian model fits (red) conditioned on the previous trial type for three participants in the *unbiased* session. Error bars: standard error of the mean. Difference between $p(\widehat{High})$ computed at the central tone frequency 511 Hz ($\Delta_{short-term}$): (A) ~0.1, (B) ~0.22 and (C) ~0.32. (D) Corresponding psychometric curves averaged across all participants; $\Delta_{short-term}$: ~0.12. *: two-tailed Wilcoxon sign-rank test. (E-G) Average psychometric curves for three participants in the *unbiased* (black), *biased low* (yellow) and *biased high* (brown) sessions. Respective model fits also shown using dotted lines; (E, F) fit using the full Bayesian model and (G) fit using the random-guess model. Error bars: standard error of the mean. Curves for the *biased* sessions are evaluated using balanced datasets (see Methods). Curves for the *unbiased* trials are similar to black lines in Figs. 3(B,C,D). Difference between $p(\widehat{High})$ for *biased low* and *biased high* trials computed at the central tone frequency 511 Hz ($\Delta_{long-term}$): (E) ~ 0.06, (F) ~ 0.35 and (G) ~ 0.51. (H) Psychometric curves averaged across all participants; $\Delta_{long-term}$: ~0.2. *:

two-tailed Wilcoxon sign-rank test. Panels (A) and (E) are from the same participant. (B,C,F,G) represent behaviors of 4 other different participants. In panels (D,H), †p<0.05; *p<0.01, **p<0.001, ***p<0.0001.

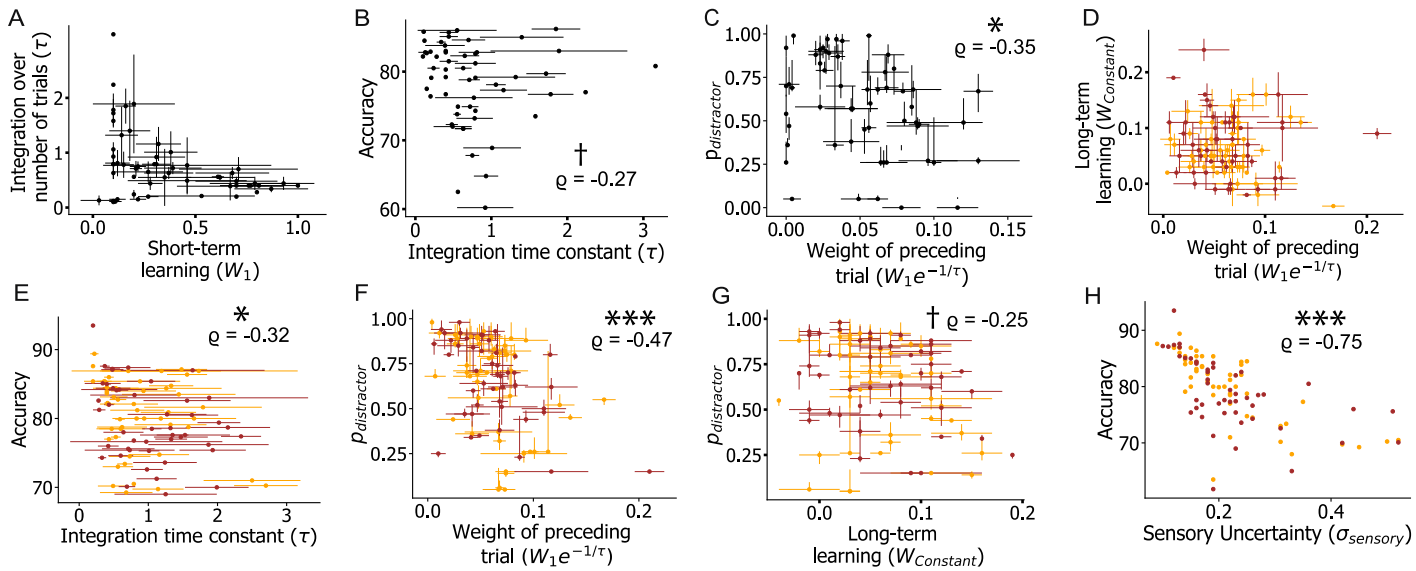


Figure 6: Participants trade-off their subjective estimate of distractor probability with learning of stimulus statistics when making category decisions. However their categorization accuracy is largely determined by their sensory uncertainty. (A) On *unbiased* trials, nearly all participants (N=55) exhibited short-term learning (W_1 > 0). (B) Their accuracy was inversely correlated with the time constant (τ) for such learning. Additionally, the median number of previous trials over which participants' retained information was 0.63; effectively participants retained information over the immediately preceding trial. (C) The influence of the preceding trial ($W_1e^{-1/\tau}$) was inversely correlated with participants' estimated probability of distractors ($p_{distractor}$). *: Spearman correlation. (D) In *biased* trials (*biased low*: orange, N=53; *biased high*: brown, N=48), the effect of short- and long-term learning varied substantially across participants. For participants who were fit using the adapted full Bayesian model (*biased low*: orange, N=51; *biased high*: brown, N=46), similar to *unbiased* trials, (E) their short-term learning time constant was inversely correlated with their accuracy. Moreover, $p_{distractor}$ was inversely correlated with both (F) short-term learning given by $W_1e^{-1/\tau}$ and (G) long-term learning captured using $W_{constant}$. (H) For all participants, similar to Fig. 4(F), category accuracy was impaired by sensory uncertainty. All panels, data point: one participant; †p<0.05; *p<0.01, **p<0.001, ***p<0.0001; error bars: standard error of the mean.

To quantify how learning modulates category decision-making, we adapted the full Bayesian model to incorporate both short- (exponential weighted: $W_1e^{-t/\tau}$) and long-term (constant: $W_{constant}$) components (Values of W_1 and $W_{constant}$ ~ 0 implies that there was not any learning, whereas values ~1 implies learning). This adapted model captured the combined effects of temporal integration over tones, stimulus relevance, sensory uncertainty, as well as learning of temporal stimulus statistics. Using this framework, we first tested if participants continued to account for stimulus relevance when categorizing trials during *biased low* and *high* sessions. We observed that the performance of most participants, except for two, was best fit by the full Bayesian model (extended data Fig. 6). These two outlier participants primarily used their learned priors and their performance was thus best fit by the random-guess model (extended data Figs. 6A-D). These data reinforce our previous finding that stimulus relevance is central to categorization.

Next, we tested whether participants learned the short-term statistics of trials and identified the nature of this learning process. Fitting the adapted version of the full Bayesian model to data from the *unbiased* session, we found that nearly all participants demonstrated significant short-term learning (Fig. 6A; $W_1 > 0$ for N = 55). However, there was substantial variability across participants such that for some W_1~0 whereas for some others it was ~1. Similar heterogeneity was also observed in the time constant for learning such that the time constant was inversely correlated with participants' accuracy (Fig. 6B; Spearman, ρ = -0.27, p = 0.042). Overall, the average time constant of short-term learning (τ) over the 56 participants in the *unbiased* session was ~1 trial

(Mdn: 0.63 trial, IQR: 0.44; Figs. 6A, B). The statistical nature of W_1 (Fig. 6D) and τ (Fig. 6E; Table 1 – Spearman correlation with accuracy) was similar in the *biased* sessions. These results jointly suggest that participants used short-term learning to a varying degree when making category decisions but despite the variability this learning was mostly restricted to only the immediately preceding trial.

Next, we explored whether participants exhibited effects of long-term learning and how it might be related to their short-term learning. We found that the vast majority of participants learned long-term trial statistics (Fig. 6D), but this learning was not correlated to their short-term learning (Spearman; Table 1). Thus, participants appear to use both the bias in the session and their knowledge of previous trials albeit to different degrees when making decisions. However, when we correlated these two learning mechanisms to participants' estimated probability of distractors, the weight of the preceding trial ($W_1 e^{-1/\tau}$) was inversely correlated with $p_{distractor}$ for all three sessions (Figs. 6C, F; Spearman, *unbiased*: ρ = -0.36, p = 0.0069, *biased low and high*: ρ = -0.47, p = 1e-6). Similarly, long-term bias ($W_{constant}$) was also inversely correlated with $p_{distractor}$ (Fig. 6G; Spearman, *biased low and high*: ρ = -0.25, p = 0.013). This indicates that participants traded off their estimated probability of decision-irrelevant tones with their stimulus-category knowledge derived over both short- and long-time scales.

Finally, we asked how the four parameters that drive performance – two types of learning, stimulus relevance and sensory uncertainty relate to task accuracy. We found that sensory uncertainty was a strong predictor of accuracy, accounting for much of the variance (Fig. 4F, 6H; Spearman, *biased low and high*: ρ = -0.75, p = 2.38e-19). The main factor in deciding how proficiently people categorized sounds could be interpreted as how well they perceived the different tone frequencies. These results support to the basic assumption of Bayesian models that sensory uncertainty is a fundamental factor in decision making.

Discussion

Here, we tested whether and how the interplay of stimulus relevance, sensory uncertainty, and learning over short- and long-time durations shapes human auditory categorization. We found that when categorizing tone sequences consisting of both category-irrelevant (distractors) and relevant tones, participants weighed individual tones by their task relevance (Figs. 2,3; extended data Figs. 1,2,3,6). Additionally, participants' category choices were jointly determined by their sensory uncertainty and their subjective estimate of stimulus relevance, whereas their accuracy was mostly driven by their sensory uncertainty (Fig. 4). We also observed that participants' prior expectations of stimulus category were affected by their knowledge of stimulus statistics over both short and long timescales and this effect was inversely correlated with participants' estimated probability of distractors (Figs. 5,6; extended data Fig. 5). Importantly, our Bayesian framework (Fig. 1) provided good fits to the data from both individual participants and the overall population (Figs. 3,5; extended data Figs. 2,3,5,6). Taken together, we found that participants differentially combined information about stimulus relevance, sensory uncertainty and learning when making categorization decisions.

It has long been known that humans and other animals can isolate stimuli of interest in noisy, crowded environments. For example, several studies have found that when presented with two competing auditory streams, participants' neural responses to the task-relevant stimuli were enhanced[46,47], but, their processing of the irrelevant distractor stimuli was mediated by attention[48,49]. Similar findings in the visual object-processing literature suggest that spatial and feature attention affect the neural representation of task-irrelevant sensory information[50,51]. Furthermore, because such information is often incompletely suppressed, the presence of distractors has been shown to impair human decision-making behavior[15]. However, much of this work has studied the effects of relevance in isolation. When making decisions in noisy, uncertain environments, humans often incorporate multiple sources of information. One such source is the long-term sensory statistics of the stimuli of interest[26,52–54] – studies have found that sensitivity to these statistics introduces observable biases in perception. However, in addition to tuning to long-term regularities, our sensory representations often need to rapidly adapt to short-term changes. Indeed, such adaptation underscores context-dependent speech perception and categorization[23]. In our work we investigated this complexity in perceptual decision-making by testing how participants' sensory uncertainty, measure of stimulus relevance and learning interacted to shape their auditory categorization behavior.

Compared to previous work[26,55], our experimental design is unique in that the stimulus was a three-tone sequence and any number of these tones could be task-irrelevant distractors (Fig. 1). This allowed us to examine not only how participants accounted for relevance during categorization but also how their sensory uncertainty might have affected their associated measures of relevance (Figs. 2,3,4; extended data Fig. 1). Additionally, though humans use both short- and long-term knowledge to efficiently make decisions, most studies probing the influence of learned expectations on categorization have largely examined either short- or long-term effects without testing for their interaction. Because in part of our experiment, we biased the tone sequences towards either the low or the high category, we could investigate how participants' reliance on long-term stimulus statistics interacted with their learning of short-term statistics i.e., the effect of the previous trial's category (Fig. 6). We found that the effects of the two time scales were independent of each other which is broadly consistent with previous work[53] demonstrating that long-term representations are robust in the face of passive exposure to short-term change in input statistics.

However, integral to our experimental paradigm was assessing how the interplay between all four factors shaped participants' category decisions. Here, we leveraged a Bayesian model to interrogate this question (Figs. 1,4,6). Bayesian models are a powerful tool that has been successfully used to study human behavior in sensorimotor learning[12,56], sensory perception[22,57], categorization[26,37] etc. Our Bayesian analysis allowed us to examine the distribution of model parameters, in particular the relevance parameter $p_{distractor}$ and the learning parameters W_0, W_1 and τ. Interestingly, from the model fits we found an average $p_{distractor}$ higher than the true experimental value (inferred $p_{distractor} \sim 0.58$ (unbiased) and ~ 0.65 (biased)vs. experimental $p_{distractor} =$

0.3). Participants had a tendency to ignore tones more than the statistics of the task would require, a potential explanation for which is that humans consider most sounds in their environment to be distracting. Furthermore, we found that the average values of W_0 (i.e. long-term learning) and $W_1 e^{-\frac{1}{\tau}}$ (i.e. short-term learning) were greater than 0 (Fig. 6) and across the population both were inversely correlated with $p_{distractor}$. This demonstrates that when making decisions, participants combined their uncertainty about stimulus relevance with other sources of information, a finding that supports similar observations from previous studies[58].

A longstanding trend in sensory psychophysics has been to investigate and report average human behavior leaving inter-participant variability, except for that in sensory uncertainty and auditory thresholds, relatively underexplored[55,59,60]. Here, using different possible decision-making models we characterized how individual participants may vary in their categorization behaviors. Across all three sessions, performance of nearly all participants was best captured by the full Bayesian model (Figs. 3-5; extended data Fig. 3). However, in the *unbiased* session, few participants were found to be equally well fit by both the full Bayesian and the no-distractor models, suggesting that they did not take relevance into account when making category decisions. In the *biased* sessions, two participants who relied only on long-term stimulus statistics to make their decisions were best captured by the suboptimal random-guesses model (extended data Fig. 6). In addition to this heterogeneity across strategies, we found that even among the participants using the optimal strategy there was substantial variability (Figs. 3-6; extended data Figs. 2-5) resulting from participants' subjective measure of relevance and their degree of short- and long-term learning. This variability was evident not only in participants' psychometric curves which were shaped by both their sensory uncertainty and measure of relevance (Fig. 3; extended data Fig. 3), but also in participants' trade-off between learning and relevance (Fig. 6). In the latter, we observed that not all individuals used their prior information. Overall, these results illustrate diversity in the human perceptual decision-making process and further work could test whether participants use the same strategy throughout the experiment or switch between different strategies based on the complexity of the trial.

The interaction of the four factors studied here will also have implications for the understanding of how humans navigate crowded noisy environments, such as those at a cocktail party. For example, spatial cues[61,62] and attention[63–65] are integral to parsing multiple speakers, but it is not yet well understood how humans can effortlessly identify relevant sounds in these noisy conditions. In our study, participants categorized three-tone sequences into high or low categories such that their category decisions were not affected by tone positions. However, real-life stimuli such as speech[66] and musical chords have rich information both within and across sequences. Whether such structure can aid relevance discrimination remains untested. Furthermore, our results indicate that participants traded off their estimates of stimulus relevance with their learning of stimulus statistics over different time scales. Future work should explore the role of relevance in the cocktail party problem and whether prior knowledge of one's environment can detract from identification of relevant sounds.

By covering many aspects of the complex process of auditory category learning, our approach may also open new ways of studying the neural basis of audition. Short- and long-term learning[67–71], uncertainty in relevance of stimuli[50], uncertainty in category of trials[72] and sensory uncertainty may all be captured by distinct circuit elements and brain regions[73,74]. As such, experiments that vary these factors in a model-driven framework can enable new approaches towards understanding neural computation in the auditory system.

Methods

Ethics Statement

The protocol was reviewed by the Institutional Review Board of the University of Pennsylvania and determined to be exempt as authorized by 45 CFR 46.104, category 2. All participants provided informed consent prior to participating.

Human Psychophysics

Apparatus. We used the crowdsourcing online platform provided by Prolific to recruit participants. We collected participants' consent and demographic data using Qualtrics and conducted all experiments using Pavlovia. The experiment could only be conducted on a laptop or desktop computer but not on a tablet or mobile phone. Each participant had to pass a 'headphone-check' experiment that helped ensure that they were wearing headphones or earphones as instructed[75] before continuing with the experiment.

Stimulus creation and task design. Human participants performed an auditory categorization task in which they reported the frequency category ($\widehat{\text{High}}$ or $\widehat{\text{Low}}$) of a three-tone sequence (tone duration: 300 ms during training, 280 ms during testing; inter-tone interval: 175 ms during training, 165 ms during testing; 44100 Hz sampling rate; 20 ms on- and off-ramps). Each of the three tones probabilistically could be either signal or distractor; that is a signal tone was relevant to the category membership of a sequence, whereas a distractor tone was irrelevant. The signal frequency probability distributions were Gaussian (p_G; Eqs. 1a,b) in log space: the mean of the low-frequency distribution (μ_L) was 2.55, the mean of the high-frequency distribution (μ_H) was 2.85, and standard deviation of both (σ_{expt}) was 0.1. The distractor frequency probability distribution was uniform (p_U; Eq. 1c) and overlapped with both Gaussian distributions.

$$p_{GL} = \frac{1}{\sqrt{2\pi}\sigma_{expt}} \exp\left(-\frac{(t-\mu_L)}{2\sigma_{expt}^2}\right) \qquad (1a)$$

$$p_{GH} = \frac{1}{\sqrt{2\pi}\sigma_{expt}} \exp\left(-\frac{(t-\mu_H)}{2\sigma_{expt}^2}\right). \qquad (1b)$$

$$p_U = \begin{cases} \frac{1}{b-a} & \text{for a} \leq \text{t} \leq \text{b} \\ 0 \text{ for t} < \text{a and t} > \text{b.} \end{cases} \qquad (1c)$$

Overall, a low (high) category sequence was a combination of low-frequency (high-frequency) signal tones and distractor tones. The complete frequency range of the tones was between 90 – 3000 Hz and was uniformly sampled in log space such that the frequency of each tone was one of 30 possible values.

Participants completed three sessions of the experiment: *unbiased, biased low,* and *biased high*. Tone sequences in the *unbiased* sessions were equally likely to be drawn from either the high or low categories, i.e., $p_L = 0.5$. Conversely, in the *biased low* (*biased high*) sessions, we overrepresented the corresponding distribution such that $p_L(p_H) = 0.7$. Across all three sessions, the probability that a tone was signal was $p_S = 0.7$ and of it being a distractor was $p_D = 0.3$. The generative process of the experimental stimuli can be summarized as,

$$p(\log(t_1, t_2, t_3)\,|C) = \sum_{r_1=R,I} \sum_{r_2=R,I} \sum_{r_3=R,I} p(\log(t_1, t_2, t_3)|C, r_1, r_2, r_3)p(r_1, r_2, r_3|C)\,. \quad (2)$$

t_i is an experimental tone. r_i is the relevance of the tone such that r_i = R implies that the tone t_i is signal (relevant) and r_i = I implies that it is a distractor (irrelevant) ($i \in \{1,2,3\}$). C is the category of the sequence (high (H) or low (L)). Because the tones and their relevance are conditionally independent, given a trial's category choice, we can rewrite Eq. 2 as:

$$p(\log(t_1, t_2, t_3)\,|C) = \sum_{r_1=R,I}\sum_{r_2=R,I}\sum_{r_3=R,I} p(\log(t_1)\,|C, r_1)p(\log(t_2)\,|C, r_2)p(\log(t_3)\,|C, r_3)p(r_1|C)p(r_2|C)p(r_3|C). \quad (3)$$

For example, in a high-category sequence for a given tone t_i,

$$\sum_{r_i=R,I} p(\log(t_i)\,|H, r_i)p(r_i|H) = p(\log(t_i)\,|H, r_i = R)p(R|H) + p(\log(t_i)\,|H, r_i = I)p(I|H) \quad (4)$$

$$= \mathrm{p}_{GH}\big(\log(t_i)\,|\mu_{\mathrm{H}}, \sigma_{expt}\big)(1 - \mathrm{p}_D) + \mathrm{p}_U(\log(t_i))\mathrm{p}_D. \quad (5)$$

The priors in the generative process are $p(C = L) = p_L$ and $p(C = H) = p_H = 1 - p_L$.

Participants. In total, 48 healthy adults (22 female; mean ± SD age of the 48 adults: 26.32 ± 7.17) successfully completed the full study. Of the 169 adults who signed up for the study, 32 declined to participate after signing the consent forms, 72 failed the 'headphone check', and 9 others either only completed one of the *biased* sessions or did not exceed 70% accuracy on those trials that did not have distractors. We tested for this accuracy threshold during a participant's first session, which ensured that the participants understood the basic task of categorizing three-tone sequences into two categories. We had 3 study variations to randomly counterbalance the order of the session type across participants: (1) *unbiased, biased low,* and *biased high* (N=16); (2) *biased high, biased low,* and *unbiased* (N=17); and (3) *biased low, unbiased,* and *biased high* (N=15). 8 other adults (3 female; mean ± SD age of the 8 adults: 32 ± 13.95) only partially completed the study, which supplemented data in the *unbiased* experiment. Data from 5 of these participants supplemented data in the *biased low* experiments.

All participants had normal or corrected-to-normal hearing. A power analysis of pilot data indicated that our sample size of 56 participants in the *unbiased* session and 48 participants across both the *biased* sessions was triple the necessary size, which helped to ensure that we could capture variability across real-world participants. Additionally, the cohort size is similar to other comparable studies[26,36,76–78].

Experimental setup. During the training part of the task, participants were serially presented with 16 easy trials in which the frequencies of all three tones were signal and drawn from the mean value $\pm\, 1\sigma$ of either the low- or high-frequency Gaussian (signal) distributions. They were then presented with 10 more difficult trials in which two of the three tones were signal: two tones were drawn from the mean $\pm 1\sigma$ of either the low- or high-frequency Gaussians, whereas the third was distractor and drawn from the extremes of the uniform distractor distribution ($< \mu_L - 4\sigma$ or $> \mu_H + 4\sigma$). Finally, participants performed 9 very difficult trials in which only one of the three tones was signal and the other two were distractors. Additionally, participants were exposed to *unbiased*, *biased low,* and *biased high* sessions using the same training paradigm. We did not inform them about the potentially biased nature of a session.

In the testing phase, we had 600 trials in the *unbiased* session and 800 trials in each of the *biased* sessions. We randomized the trials and divided them into 4 blocks of 150 trials each for the *unbiased* session and 5 blocks of 160 trials for the *biased* sessions. The participants could take a break (maximum: 3 min) between each block. Participants had to respond within 1.8 seconds after offset of the last tone and were given feedback about their report after each trial. In the *unbiased* session, the median number of no-response trials was 1, and the 90th percentile was 3.8. The number of no-response trials was also low during *biased low* (median: 1; 90th percentile: 7) and *biased high* (median:1; 90th percentile: 11.4) sessions. We excluded the no-response trials when we computed participant accuracy.

At the end of each experimental session, participants were asked for their consent to be recalled for the next session. The dates of the data collection were staggered and we conducted each subsequent session after a variable number of days ranging from (1-30 days). We could not identify any relationship between the days between sessions and a participant's behavior.

Raw Data Analysis

Fig. 2A: High-category choice was coded as 1 and low-category choice was coded as 0; correlations were normalized to add to 1.
Fig. 2B: Signal and distractor refer to the tone's ground truth stimulus type.

Computational models

Bayesian models. We assumed that participants performed Bayesian inference over a generative model when solving the categorization task: given their sensory evidence of the tones t_{p_i} where $i \in \{1,2,3\}$, participants computed the posterior probability,

$$p\left(C \middle| \log\left(t_{p_1}, t_{p_2}, t_{p_3}\right)\right) = \frac{p\left(\log\left(t_{p_1}, t_{p_2}, t_{p_3}\right) \middle| C\right) p(C)}{p\left(\log\left(t_{p_1}, t_{p_2}, t_{p_3}\right)\right)} \qquad (6)$$

where C is the category choice ($\widehat{\text{High}}$ ($\widehat{\text{H}}$) or $\widehat{\text{Low}}$ ($\widehat{\text{L}}$)). In this equation, the likelihood, $p\left(\log\left(t_{p_1}, t_{p_2}, t_{p_3}\right) \middle| C\right)$ that the sequence of sensory evidence t_{p_i} belonged to a particular category was calculated by marginalizing over a broad range of hypothesized tone frequencies t_{H_i}, $i \in \{1,2,3\}$. To account for the entire frequency range of human hearing[79,80], both t_{H_i} and $t_{p_i} \in \{3.98,\ 39{,}810\}$ Hz. In addition, we introduced a dummy variable r_{p_i} that explicitly characterized whether a participant deemed each tone (t_{p_i}) as 'relevant' (denoted by R) or 'irrelevant' (given by I) to their category decision. Thus, the likelihood of a high-category three-tone trial is,

$$p\left(\log\left(t_{p_1}, t_{p_2}, t_{p_3}\right) \middle| H\right) = \sum_{r_{p_1}=R,I} \sum_{r_{p_2}=R,I} \sum_{r_{p_3}=R,I} p\left(\log\left(t_{p_1}, t_{p_2}, t_{p_3}\right) \middle| H, r_{p_1}, r_{p_2}, r_{p_3}\right) p\left(r_{p_1}, r_{p_2}, r_{p_3} \middle| H\right) \qquad (7)$$

$$= \sum_{r_{p_1}=R,I} \sum_{r_{p_2}=R,I} \sum_{r_{p_3}=R,I} \sum_{t_{H_1}} \sum_{t_{H_2}} \sum_{t_{H_3}} p\left(\log\left(t_{p_1}, t_{p_2}, t_{p_3}\right)\right) \middle| H, r_{p_1}, r_{p_2}, r_{p_3}, \log\left(t_{H_1}, t_{H_2}, t_{H_3}\right)\right)$$
$$* \, p\left(\log\left(t_{H_1}, t_{H_2}, t_{H_3}\right) \middle| H, r_{p_1}, r_{p_2}, r_{p_3}\right) * p\left(r_{p_1}, r_{p_2}, r_{p_3} \middle| H\right). \qquad (8)$$

Each hypothesized tone t_{H_i} generated sensory evidence t_{p_i} according to the probability density $p\left(\log\left(t_{p_i}\right) \middle| \log\left(t_{H_i}\right)\right)$, which is a Gaussian characterized by the participant's sensory uncertainty and noise in the auditory pathway (denoted using parameter $\sigma_{sensory}$). Thus,

$$p\left(\log\left(t_{p_1}, t_{p_2}, t_{p_3}\right) \middle| H, r_{p_1}, r_{p_2}, r_{p_3}, \log\left(t_{H_1}, t_{H_2}, t_{H_3}\right)\right)$$
$$= p\left(\log\left(t_{p_1}\right) \middle| \log\left(t_{H_1}\right)\right) p\left(\log\left(t_{p_2}\right) \middle| \log\left(t_{H_2}\right)\right) p\left(\log\left(t_{p_3}\right) \middle| \log\left(t_{H_3}\right)\right). \quad (9)$$

The details of how we computed the value of $\sigma_{sensory}$ are given below in Methods - model fit and predictions.

Next, because the hypothesized tones and participant's estimated relevance are conditionally independent given trial category, we can rewrite:

$$\sum_{r_{p_1}=R,I} \sum_{r_{p_2}=R,I} \sum_{r_{p_3}=R,I} p\left(\log\left(t_{H_1}, t_{H_2}, t_{H_3}\right) \middle| H, r_{p_1}, r_{p_2}, r_{p_3}\right) p\left(r_{p_1}, r_{p_2}, r_{p_3} \middle| H\right)$$
$$= \sum_{r_{p_1}=R,I} \sum_{r_{p_2}=R,I} \sum_{r_{p_3}=R,I} p\left(\log\left(t_{H_1}\right) \middle| H, r_{p_1}\right) p\left(\log\left(t_{H_2}\right) \middle| H, r_{p_2}\right) p\left(\log\left(t_{H_3}\right) \middle| H, r_{p_3}\right) p\left(r_{p_1}, r_{p_2}, r_{p_3} \middle| H\right). \quad (10)$$

Additionally, because we observed that frequencies in each tone position are equally correlated with behavior (Fig. 2A), the relevance of a tone is agnostic of its order in the three-tone sequence. Using the generative model (see Eqs. 4 and 5), we can then expand Eq. 10 as follows,

$$\sum_{r_{p_1}=R,I} \sum_{r_{p_2}=R,I} \sum_{r_{p_3}=R,I} p\left(\log\left(t_{H_1}\right) \middle| H, r_{p_1}\right) p\left(\log\left(t_{H_2}\right) \middle| H, r_{p_2}\right) p\left(\log\left(t_{H_3}\right) \middle| H, r_{p_3}\right) p\left(r_{p_1}, r_{p_2}, r_{p_3} \middle| H\right)$$
$$= \prod_{i=1,2,3} p\left(\log\left(t_{H_i}\right) \middle| H, r_{p_i} = R\right) p(R|H) + p\left(\log\left(t_{H_i}\right) \middle| H, r_{p_i} = I\right) p(I|H) \qquad (11)$$

$$= \prod_{i=1,2,3} p_{GH}\left(\log\left(t_{H_i}\right) \middle| \mu_{high}, \sigma\right) * (1 - p_{distractor}) + p_U(\log(t_{H_i})) * p_{distractor}. \quad (12)$$

When computing the posterior in Eq. 6, $p(C) = p_{low}$ for category choice $\widehat{\text{Low}}$ and p_{high} for category choice $\widehat{\text{High}}$. Here, $\mu_{low}, \mu_{high}, \sigma, p_{distractor}$, and p_{low} are the participants' estimated values of the true experimental parameters $\mu_L, \mu_H, \sigma_{expt}, p_D$, and p_L and are fitting parameters in the Bayesian models. When we fit these parameters to the participants' data, we minimized the negative likelihood (-log(posterior)), which is the cost function.

We compared three decision-making models in the analysis: (1) A 'full Bayesian' model, which assumed that participants probabilistically determined if a tone was 'signal' or 'distractor'; (2) a 'no-distractor' model, which assumed that participants considered all of the tones to be 'signal'; and (3) a 'random-guess' model, which assumed that participants considered all of the tones to be 'distractors' and thus responded randomly.

The full Bayesian model has 6 fitting parameters – $\mu_{low}, \mu_{high}, \sigma, \sigma_{sensory}, p_{distractor}$, and p_{low} – such that $0 < p_{distractor} < 1$. On the other hand, in the no-distractor model, $r_{p_i} = R$ for $i \in 1,2,3$ in Eq. 11. Because this model assumes that all of the tones are 'signal', we can simplify the likelihood $p\left(\log\left(t_{H_1}, t_{H_2}, t_{H_3}\right) \middle| H\right)$ into a product of three Gaussians. As a consequence, the no-distractor model has 5 fitting parameters – $\mu_{high}, \mu_{low}, \sigma, p_{low}$, and $\sigma_{sensory}$. In the random-guess model, participants categorized trials using random guesses proportional to the probability of the prior, and the 2 fitting parameters are p_{low} and $\sigma_{sensory}$. More specifically, in this model the posterior in Eq. 6 simplifies to,

$$p\left(C \middle| \log\left(t_{p_1}, t_{p_2}, t_{p_3}\right)\right) = \sum_{t_{H_1}} \sum_{t_{H_2}} \sum_{t_{H_3}} p\left(\log\left(t_{p_1}, t_{p_2}, t_{p_3}\right) \middle| \log\left(t_{H_1}, t_{H_2}, t_{H_3}\right)\right) p(C).$$

$$= p\left(\log\left(t_{p_1}\right) \middle| \log\left(t_{H_1}\right)\right) p\left(\log\left(t_{p_2}\right) \middle| \log\left(t_{H_2}\right)\right) p\left(\log\left(t_{p_3}\right) \middle| \log\left(t_{H_3}\right)\right) p(C). \quad (13)$$

To capture change in expectations resulting from short- and long-term learning (Fig. 6), we fit an adapted full Bayesian model with a time-varying prior, such that,

$$p(L)_n = W_0 + W_1 \sum \mathbb{1}_{t_C} e^{-t/\tau}. \quad (14)$$

Here, $p(L)_n$ denotes a participant's prior for the low-category choice decision in the nth trial. The indicator function $\mathbb{1}_{t_C}$ takes the value 1 if the ground truth of $(n-t)^{th}$ trial was low (L) and it takes the value -1 if it was high (H). $W_{constant} = |W_0 - 0.5| * 2$ captures the long-term effect on participant's expectations, W_1 indicates the short-term effect and τ is the time constant governing the short-term effect.

Lastly, to fit each of the above models to the participants' psychometric curves (see Figs. 3,5; extended data Figs. 4-6), we applied a decision threshold to the Bayesian model posterior such that,

$$p(C \mid \log(t_1, t_2, t_3)) = \sum_{t_{p_1}} \sum_{t_{p_2}} \sum_{t_{p_3}} p\left(\log\left(t_{p_1}, t_{p_2}, t_{p_3}\right) \middle| \log\left(t_1, t_2, t_3\right)\right) \left[p\left(C \middle| \log\left(t_{p_1}, t_{p_2}, t_{p_3}\right)\right) > 0.5\right]. \quad (15)$$

Similar to Eq. 9, we considered that each tone t_i generated sensory evidence t_{p_i} according to the probability density $p\left(\log\left(t_{p_i}\right) \middle| \log(t_i)\right)$, which is a Gaussian characterized by the participant's sensory uncertainty (denoted by the parameter $\sigma_{sensory}$).

Generalized Linear model. As a simpler alternative model to the Bayesian model, we constructed an elastic net (Lasso + Ridge) regularized binary logistic regression model[81] for each participant such that the model predictors are the indicator functions for the three tones in a trial. An indicator function $\mathbb{1}_{t_i}$ takes the value of 1 if tone t_i is present in the trial and 0 otherwise. There are 30 predictors because a tone frequency can have 30 possible values. The model can be represented as,

$$y = \begin{cases} 1 & \text{expit}(\beta_0 + \beta_1 \mathbb{I}_1 \dots \beta_{30} \mathbb{I}_{30}) > 0.5 \\ 0 & \text{otherwise.} \end{cases} \quad (16)$$

Here, y is the participant's category report. The βs capture the relevance (weight) of the different experimental tone frequencies in a participant's decision-making process. The function $\text{expit}(x)$ is computed as $\frac{1}{1+\exp(-x)}$.

Model Fit and Predictions

Bayesian models: full Bayesian and no-distractor models. We used two approaches to fit the full Bayesian and no-distractor models to the *unbiased* data. First, we fit all the parameters to the participant data. However, we suspected that we may be overfitting $\mu_{low}, \mu_{high}, \sigma$ and $p_{distractor}$, because the parameters are redundant, and there are multiple configurations that lead to the same likelihood function (Eq. 12) for the Bayesian model. Thus, in the second approach, we considered participants to be veridical about the parameters of the Gaussian distributions ($\mu_{low} = 2.55$, $\mu_{high} = 2.85$ and $\sigma = 0.1$) and only fit the remaining parameters – $\sigma_{sensory}, p_{low}$ and for the full Bayesian model $p_{distractor}$. Because both approaches had similar fits (see extended data Figs. 3F, G) and to prevent overfitting, we followed the second approach.

In both the approaches, we initially fit these models to a participant's performance using a coarse parameter grid search, which allows for systematic multi-start optimization. When applicable, we used the following grid values: $\mu_{low} \in \{2.1, 2.25, 2.4, 2.55, 2.7\}$, $\mu_{high} \in \{2.7, 2.85, 3, 3.15, 3.3\}$, $\sigma \in \{0.05, 0.25, 0.45, 0.65, 0.85\}$, $p_{distractor} \in \{0.05, 0.25, 0.45, 0.65, 0.85\}$ and p_{low} was constrained by each participant's correct performance. We also assumed that $\sigma_{sensory}$ is participant and task dependent but not model dependent. Moreover, because the essential experimental goal was the same in both the *unbiased* and the *biased* sessions, we assumed that $\sigma_{sensory}$ was constant across the three session types. Thus, instead of running a separate psychophysics experiment to compute $\sigma_{sensory}$, we used data from the *unbiased* session for trials with all three tones drawn from the signal distributions and no distractor tones. We fit this data using a smaller number of model parameters – $\mu_{low}, \mu_{high}, \sigma, \sigma_{sensory}$, and p_{low}. When the remaining 4 parameters were systematically varied, we found that a participant's $\sigma_{sensory}$ was essentially constant; across the population: median – 0.19; 10th percentile – 0.14; and 90th percentile – 0.29.

Similar to the *unbiased* session, we also fit data from the two *biased* sessions using the second fitting approach. To generate confidence intervals on our fits, we created either 100 bootstrapped (with replacement) datasets comprising of 600 trials each for the *unbiased* session (refer to Fig. 4, extended data Fig. 4) or 100 balanced datasets for the *biased* tasks (Fig. 6, extended data Fig. 6). A balanced dataset was constructed using all trials from the underrepresented category and subsampling an equal number of trials from the overrepresented category. This balanced dataset separated the external bias in the experiment (captured by $p_L = 0.7$ or $p_H = 0.7$) from the internalized expectation of the participants, as the latter can vary from 0 to 1. For example, participant in Fig. 5E has negligible internalized expectation of bias, whereas participants in extended data Figs. 6C, D have high internalized expectations. These data were fit with "post-grid search", which is a finer-resolution optimization routine using the Nelder-Mead method from scipy.optimize.minimize (Python). During this optimization, we allowed $\sigma_{sensory}$ to vary within $\pm$ 0.02.

When fitting the adapted full Bayesian model to participant performance, we reduced the number of fitting parameters from 5 ($p_{distractor}, \sigma_{sensory}, W_0, W_1$ and τ) to 3 by using the median values of $p_{distractor}$ and $\sigma_{sensory}$. The parameter median values were calculated from the previous full Bayesian model fits to the bootstrapped (*unbiased*) or the balanced (*biased low* and *biased high*) datasets. To generate confidence intervals for W_0, W_1 and τ, for the *unbiased* data, we used 10 subsets of 500 contiguous trials and for the *biased* data, 10

subsets of 600 contiguous trials. Additionally, we fit these parameters using the following grid values: $W_0 \in \{0.35, 0.38 \ldots 0.66\}, W_1 \in \{0, 0.1 \ldots 1\}$ and $\tau \in 10^{\{-1, -0.85 \ldots 0.7\}}$.

Random-guess model. Because this model has only two free parameters, $\sigma_{sensory}$ and p_{low}, we simplified the fitting procedure to directly fit either the 100 bootstrapped datasets (*unbiased* session) or the 100 balanced datasets (*biased* sessions). We fit p_{low} with a grid search over {0, 0.1, 0.2, … ,1}.

Metrics. Across all three experimental sessions, we used the final model parameter values of the Full Bayesian model to compute a participant's psychometric-curve fits and their Bayesian model posteriors (Figs. 3, 5; extended data Figs. 2, 3, 6). We tested the goodness of fit of the different models by comparing their Bayesian Information Criterion (BIC) values, which penalizes model complexity and encourages fits of simpler models (Figs. 4A-C; extended data Fig. 6). We also tested the results using the Akaike Information Criterion and the corrected Akaike Information Criterion[82], which were consistent with the BIC analysis.

From the full Bayesian model fits, we derived 'sigmoidicity' for each participant. This metric encapsulates the nature of the psychometric curve, it is ~ 0 for participants who can accurately identify distractor tones as irrelevant to their category decisions (example subject in Fig. 3B) and is ~ 1 for participants whose psychometric curve resembles a step function. Specifically, for each participant sigmoidicity is computed using their respective posterior curve as follows:

$$\frac{\Sigma\left|p(\widehat{\text{High}})_{\text{distractor sensory percepts}} - \frac{\left(p(\widehat{\text{High}})_{\text{sensory percept=3.98 Hz}} + p(\widehat{\text{High}})_{\text{sensory percept=39810.72 Hz}}\right)}{2}\right|}{\Sigma\left|p(\widehat{\text{High}})_{\text{all sensory percepts}} - \frac{\left(p(\widehat{\text{High}})_{\text{sensory percept=3.98 Hz}} + p(\widehat{\text{High}})_{\text{sensory percept=39810.72 Hz}}\right)}{2}\right|},$$

where the distractor sensory percept is more likely to be generated by the distractor tones than by the signal tones. Thus, in this formula, distractor sensory percept $\in \{3.98,\ \mu_{\text{low}} - 2.1\sigma - \text{median}(\sigma_{\text{sensory}})\}$Hz and $\{\mu_{\text{high}} + 2.1\sigma + \text{median}(\sigma_{\text{sensory}}),\ 39810.72\}$Hz.

Generalized Linear model (GLM). We fit a GLM to each participant's psychophysical performance by choosing a weighting parameter α using ten-fold cross-validation. α is a parameter controlling a convex combination of Lasso and Tikhonov regularization. $\alpha = 0$ implies only Lasso regularization while $\alpha = 1$ implies only Tikohonov regularization. For example, α for the curve fit in extended data Fig. 2(E) is 0.3 whereas it is 0 for those in extended data Figs. 2(F) and (G). The Tikhonov matrix (τ) is designed such that $\tau_{ii} = 2, \tau_{i,i-1} = -1$ and $\tau_{i,i+1} = -1$.

1. Heekeren, H. R., Marrett, S. & Ungerleider, L. G. The neural systems that mediate human perceptual decision making. *Nat. Rev. Neurosci.* **9**, 467–479 (2008).
2. Chen, S. Y., Ross, B. H. & Murphy, G. L. Decision making under uncertain categorization. *Front. Psychol.* **5**, (2014).
3. Niwa, M. & Ditterich, J. Perceptual Decisions between Multiple Directions of Visual Motion. *J. Neurosci.* **28**, 4435–4445 (2008).
4. Bushdid, C., Magnasco, M. O., Vosshall, L. B. & Keller, A. Humans Can Discriminate More than 1 Trillion Olfactory Stimuli. *Science* (2014) doi:10.1126/science.1249168.
5. Garcia, S. E., Jones, P. R., Rubin, G. S. & Nardini, M. Auditory Localisation Biases Increase with Sensory Uncertainty. *Sci. Rep.* **7**, 40567 (2017).
6. Heron, J., Whitaker, D. & McGraw, P. V. Sensory uncertainty governs the extent of audio-visual interaction. *Vision Res.* **44**, 2875–2884 (2004).
7. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16**, 1170–1178 (2013).
8. Qamar, A. T. *et al.* Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proc. Natl. Acad. Sci.* **110**, 20332–20337 (2013).
9. Beierholm, U., Rohe, T., Ferrari, A., Stegle, O. & Noppeney, U. Using the past to estimate sensory uncertainty. *eLife* **9**, e54172 (2020).
10. Barthelmé, S. & Mamassian, P. Evaluation of Objective Uncertainty in the Visual System. *PLOS Comput. Biol.* **5**, e1000504 (2009).
11. Zhou, Y., Acerbi, L. & Ma, W. J. The role of sensory uncertainty in simple contour integration. *PLOS Comput. Biol.* **16**, e1006308 (2020).
12. Körding, K. P. & Wolpert, D. M. Bayesian integration in sensorimotor learning. *Nature* **427**, 244–247 (2004).
13. Wei, K. & Körding, K. Relevance of Error: What Drives Motor Adaptation? *J. Neurophysiol.* **101**, 655–664 (2009).
14. Daliri, A. & Dittman, J. Successful auditory motor adaptation requires task-relevant auditory errors. *J. Neurophysiol.* **122**, 552–562 (2019).

15. Anders, U. M., McLean, C. S., Ouyang, B. & Ditterich, J. Perceptual Decisions in the Presence of Relevant and Irrelevant Sensory Evidence. *Front. Neurosci.* **11**, (2017).

16. Mirza, M. B., Adams, R. A., Friston, K. & Parr, T. Introducing a Bayesian model of selective attention based on active inference. *Sci. Rep.* **9**, 13915 (2019).

17. Hansen, K., Hillenbrand, S. & Ungerleider, L. Effects of Prior Knowledge on Decisions Made Under Perceptual vs. Categorical Uncertainty. *Front. Neurosci.* **6**, (2012).

18. Kok, P., Brouwer, G. J., Gerven, M. A. J. van & Lange, F. P. de. Prior Expectations Bias Sensory Representations in Visual Cortex. *J. Neurosci.* **33**, 16275–16284 (2013).

19. Rahnev, D., Lau, H. & Lange, F. P. de. Prior Expectation Modulates the Interaction between Sensory and Prefrontal Regions in the Human Brain. *J. Neurosci.* **31**, 10741–10748 (2011).

20. Kok, P., Mostert, P. & de Lange, F. P. Prior expectations induce prestimulus sensory templates. *Proc. Natl. Acad. Sci.* **114**, 10473–10478 (2017).

21. Rohenkohl, G., Cravo, A. M., Wyart, V. & Nobre, A. C. Temporal Expectation Improves the Quality of Sensory Information. *J. Neurosci.* **32**, 8424–8428 (2012).

22. Stocker, A. A. & Simoncelli, E. P. Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* **9**, 578–585 (2006).

23. Holt, L. L. Speech categorization in context: Joint effects of nonspeech and speech precursors. *J. Acoust. Soc. Am.* **119**, 4016–4026 (2006).

24. Kluender, K. R., Coady, J. A. & Kiefte, M. Sensitivity to change in perception of speech. *Speech Commun.* **41**, 59–69 (2003).

25. Holt, L. L., Lotto, A. J. & Kluender, K. R. Neighboring spectral content influences vowel identification. *J. Acoust. Soc. Am.* **108**, 710–722 (2000).

26. Gifford, A. M., Cohen, Y. E. & Stocker, A. A. Characterizing the Impact of Category Uncertainty on Human Auditory Categorization Behavior. *PLOS Comput. Biol.* **10**, e1003715 (2014).

27. Berniker, M., Voss, M. & Kording, K. Learning Priors for Bayesian Computations in the Nervous System. *PLOS ONE* **5**, e12686 (2010).

28. Simoncelli, E. P. & Olshausen, B. A. Natural Image Statistics and Neural Representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).

29. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).

30. Mendonça, A. G. *et al.* The impact of learning on perceptual decisions and its implication for speed-accuracy tradeoffs. *Nat. Commun.* **11**, 2757 (2020).

31. Russ, B. E., Lee, Y.-S. & Cohen, Y. E. Neural and behavioral correlates of auditory categorization. *Hear. Res.* **229**, 204–212 (2007).

32. Banno, T., Lestang, J.-H. & Cohen, Y. E. Computational and neurophysiological principles underlying auditory perceptual decisions. *Curr. Opin. Physiol.* **18**, 20–24 (2020).

33. Ley, A. *et al.* Learning of New Sound Categories Shapes Neural Response Patterns in Human Auditory Cortex. *J. Neurosci.* **32**, 13273–13280 (2012).

34. Tsunada, J. & Cohen, Y. E. Neural mechanisms of auditory categorization: from across brain areas to within local microcircuits. *Front. Neurosci.* **8**, (2014).

35. Cao, Y., Summerfield, C., Park, H., Giordano, B. L. & Kayser, C. Causal Inference in the Multisensory Brain. *Neuron* **102**, 1076-1087.e8 (2019).

36. Acerbi, L., Dokka, K., Angelaki, D. E. & Ma, W. J. Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *PLOS Comput. Biol.* **14**, e1006110 (2018).

37. Rigoli, F., Pezzulo, G., Dolan, R. & Friston, K. A Goal-Directed Bayesian Framework for Categorization. *Front. Psychol.* **8**, (2017).

38. Adler, W. T. & Ma, W. J. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLOS Comput. Biol.* **14**, e1006572 (2018).

39. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).

40. Kording, K. P. Bayesian statistics: relevant for the brain? *Curr. Opin. Neurobiol.* **25**, 130–133 (2014).

41. Piasini, E., Balasubramanian, V. & Gold, J. I. Effect of Geometric Complexity on Intuitive Model Selection. in *Machine Learning, Optimization, and Data Science* (eds. Nicosia, G. et al.) 1–24 (Springer International Publishing, 2022). doi:10.1007/978-3-030-95467-3_1.

42. Weiss, Y., Simoncelli, E. P. & Adelson, E. H. Motion illusions as optimal percepts. *Nat. Neurosci.* **5**, 598–604 (2002).

43. Winkel, K. N. de, Katliar, M. & Bülthoff, H. H. Forced Fusion in Multisensory Heading Estimation. *PLOS ONE* **10**, e0127104 (2015).

44. Wozny, D. R., Beierholm, U. R. & Shams, L. Probability Matching as a Computational Strategy Used in Perception. *PLOS Comput. Biol.* **6**, e1000871 (2010).

45. Rahnev, D. & Denison, R. N. Suboptimality in perceptual decision making. *Behav. Brain Sci.* **41**, e223 (2018).

46. Ding, N. & Simon, J. Z. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci.* **109**, 11854–11859 (2012).

47. Mesgarani, N. & Chang, E. F. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485**, 233–236 (2012).

48. Schwartz, Z. P. & David, S. V. Focal Suppression of Distractor Sounds by Selective Attention in Auditory Cortex. *Cereb. Cortex* **28**, 323–339 (2018).

49. Jensen, A., Merz, S., Spence, C. & Frings, C. Interference of irrelevant information in multisensory selection depends on attentional set. *Atten. Percept. Psychophys.* **82**, 1176–1195 (2020).

50. Chen, J., Scotti, P. S., Dowd, E. W. & Golomb, J. D. Neural Representations of Task-relevant and Task-irrelevant Features of Attended Objects. 2021.05.21.445168 Preprint at https://doi.org/10.1101/2021.05.21.445168 (2021).

51. Xu, Y. The Neural Fate of Task-Irrelevant Features in Object-Based Processing. *J. Neurosci.* **30**, 14020–14028 (2010).

52. Hansen, K., Hillenbrand, S. & Ungerleider, L. Persistency of Priors-Induced Bias in Decision Behavior and the fMRI Signal. *Front. Neurosci.* **5**, (2011).

53. Roark, C. L. & Holt, L. L. Long-term priors constrain category learning in the context of short-term statistical regularities. *Psychon. Bull. Rev.* (2022) doi:10.3758/s13423-022-02114-z.

54. Hansen, K. A., Hillenbrand, S. F. & Ungerleider, L. G. Human brain activity predicts individual differences in prior knowledge use during decisions. *J. Cogn. Neurosci.* **24**, 1462–1475 (2012).

55. Tardiff, N., Suriya-Arunroj, L., Cohen, Y. E. & Gold, J. I. Rule-based and stimulus-based cues bias auditory decisions via different computational and physiological mechanisms. 2021.12.09.471952 Preprint at https://doi.org/10.1101/2021.12.09.471952 (2021).

56. Heald, J. B., Lengyel, M. & Wolpert, D. M. Contextual inference underlies the learning of sensorimotor repertoires. *Nature* **600**, 489–493 (2021).

57. Knill, D. C. & Richards, W. *Perception as Bayesian Inference*. (Cambridge University Press, 1996).

58. Honig, M., Ma, W. J. & Fougnie, D. Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. *Proc. Natl. Acad. Sci.* **117**, 8391–8397 (2020).

59. Filipowicz, A. L., Glaze, C. M., Kable, J. W. & Gold, J. I. Pupil diameter encodes the idiosyncratic, cognitive complexity of belief updating. *eLife* **9**, e57872 (2020).

60. Ruesseler, M., Weber, L. A., Marshall, T. R., O'Reilly, J. X. & Hunt, L. T. Decision-making in dynamic, continuously evolving environments: quantifying the flexibility of human choice. 2022.08.18.504278 Preprint at https://doi.org/10.1101/2022.08.18.504278 (2022).

61. Ihlefeld, A. & Shinn-Cunningham, B. Spatial release from energetic and informational masking in a selective speech identification task. *J. Acoust. Soc. Am.* **123**, 4369–4379 (2008).

62. Middlebrooks, J. C. & Waters, M. F. Spatial Mechanisms for Segregation of Competing Sounds, and a Breakdown in Spatial Hearing. *Front. Neurosci.* **14**, (2020).

63. Bronkhorst, A. W. The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Atten. Percept. Psychophys.* **77**, 1465–1487 (2015).

64. Shinn-Cunningham, B. G. Object-based auditory and visual attention. *Trends Cogn. Sci.* **12**, 182–186 (2008).

65. Wöstmann, M., Herrmann, B., Maess, B. & Obleser, J. Spatiotemporal dynamics of auditory attention synchronize with speech. *Proc. Natl. Acad. Sci.* **113**, 3873–3878 (2016).

66. Holt, L. L. & Lotto, A. J. Speech perception as categorization. *Atten. Percept. Psychophys.* **72**, 1218–1227 (2010).

67. Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. A Comparison of Primate Prefrontal and Inferior Temporal Cortices during Visual Categorization. *J. Neurosci.* **23**, 5235–5246 (2003).

68. Zhong, L., Zhang, Y., Duan, C. A., Pan, J. & Xu, N. Dynamic and causal contribution of parietal circuits to perceptual decisions during category learning. 304071 Preprint at https://doi.org/10.1101/304071 (2018).

69. Runyan, C. A., Piasini, E., Panzeri, S. & Harvey, C. D. Distinct timescales of population coding across cortex. *Nature* **548**, 92–96 (2017).

70. Akrami, A., Kopec, C. D., Diamond, M. E. & Brody, C. D. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature* **554**, 368–372 (2018).

71. Bianco, R. *et al.* Long-term implicit memory for sequential auditory patterns in humans. *eLife* **9**, e56073 (2020).

72. Grinband, J., Hirsch, J. & Ferrera, V. P. A neural representation of categorization uncertainty in the human brain. *Neuron* **49**, 757–763 (2006).

73. van Bergen, R. S., Ji Ma, W., Pratte, M. S. & Jehee, J. F. M. Sensory uncertainty decoded from visual cortex predicts behavior. *Nat. Neurosci.* **18**, 1728–1730 (2015).

74. Fetsch, C. R., Pouget, A., DeAngelis, G. C. & Angelaki, D. E. Neural correlates of reliability-based cue weighting during multisensory integration. *Nat. Neurosci.* **15**, 146–154 (2012).

75. Woods, K. J. P., Siegel, M. H., Traer, J. & McDermott, J. H. Headphone screening to facilitate web-based auditory experiments. *Atten. Percept. Psychophys.* **79**, 2064–2072 (2017).

76. Bankieris, K. R., Bejjanki, V. R. & Aslin, R. N. Sensory cue-combination in the context of newly learned categories. *Sci. Rep.* **7**, 10890 (2017).

77. Bejjanki, V. R., Clayards, M., Knill, D. C. & Aslin, R. N. Cue Integration in Categorical Tasks: Insights from Audio-Visual Speech Perception. *PLOS ONE* **6**, e19812 (2011).

78. McPherson, M. J. & McDermott, J. H. Diversity in pitch perception revealed by task dependence. *Nat. Hum. Behav.* **2**, 52–66 (2018).

79. Moller, H. & Pedersen, C. S. Hearing at low and infrasonic frequencies. *Noise Health* **6**, 37 (2004).

80. Ashihara, K. Hearing thresholds for pure tones above 16kHz. *J. Acoust. Soc. Am.* **122**, EL52–EL57 (2007).

81. Jas, M. *et al.* Pyglmnet: Python implementation of elastic-net regularized generalized linear models. *J. Open Source Softw.* **5**, 1959 (2020).

82. Cavanaugh, J. E. Unifying the derivations for the Akaike and corrected Akaike information criteria. *Stat. Probab. Lett.* **33**, 201–208 (1997).

Acknowledgements
We thank the members of the Geffen Lab for advice and feedback on the project. We also thank Doris Dijksterhuis and Sandra Reinert for their ideas regarding data analysis. This work was supported by the NIH grant (RFA-NS-18-030) to MNG, YC, and KK. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author information

Authors and Affiliations
Department of Otorhinolaryngology, University of Pennsylvania, Philadelphia, USA

Janaki Sheth, Jared Collina, Yale Cohen, Maria Geffen

Department of Neuroscience, International School for Advanced Studies, Trieste, Italy

Eugenio Piasini

Department of Neuroscience, University of Pennsylvania, Philadelphia, PA

Konrad Kording, Yale Cohen, Maria Geffen

Department of Bioengineering, University of Pennsylvania, Philadelphia, PA

Konrad Kording, Yale Cohen

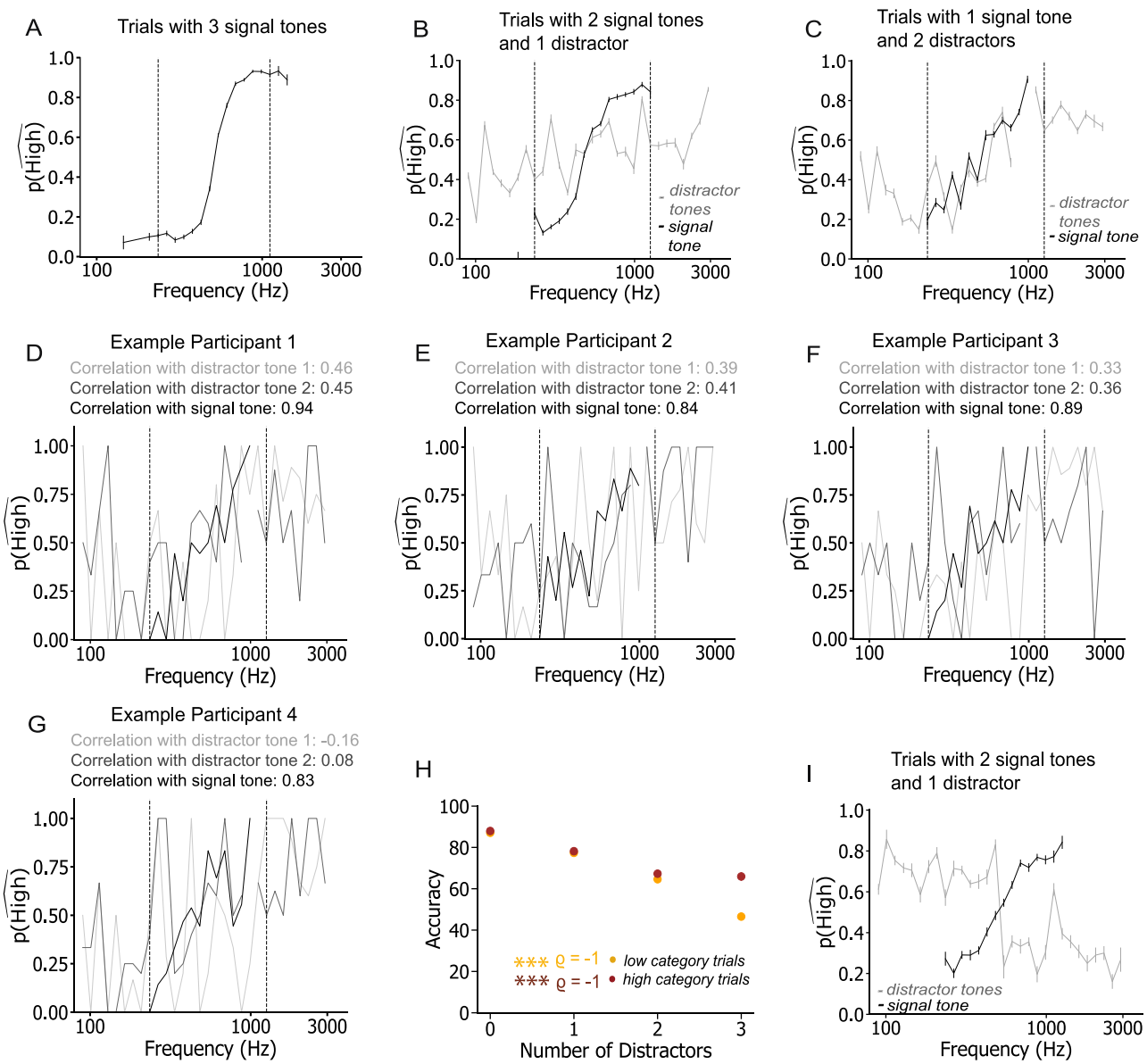
Department of Neurology, University of Pennsylvania, Philadelphia, PA

Maria Geffen

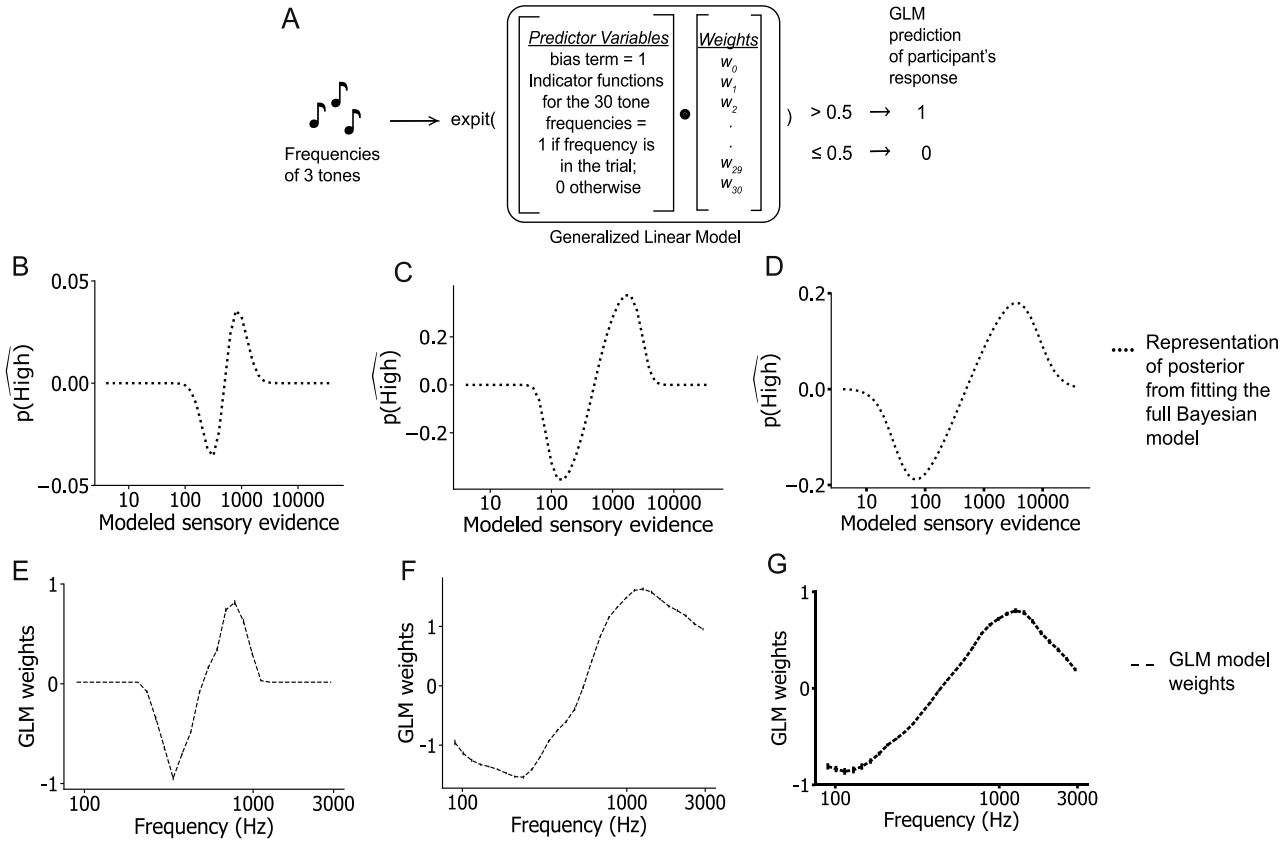
Contributions
J.S., J.C, K.K., E.P., Y.C. and M.N.G. designed the study. J.S. collected and analysed the data. J.J., K.K., Y.C. and M.N.G. wrote the paper.

Corresponding authors
Correspondence to Janaki Sheth or Maria Neimark Geffen.

Extended Data

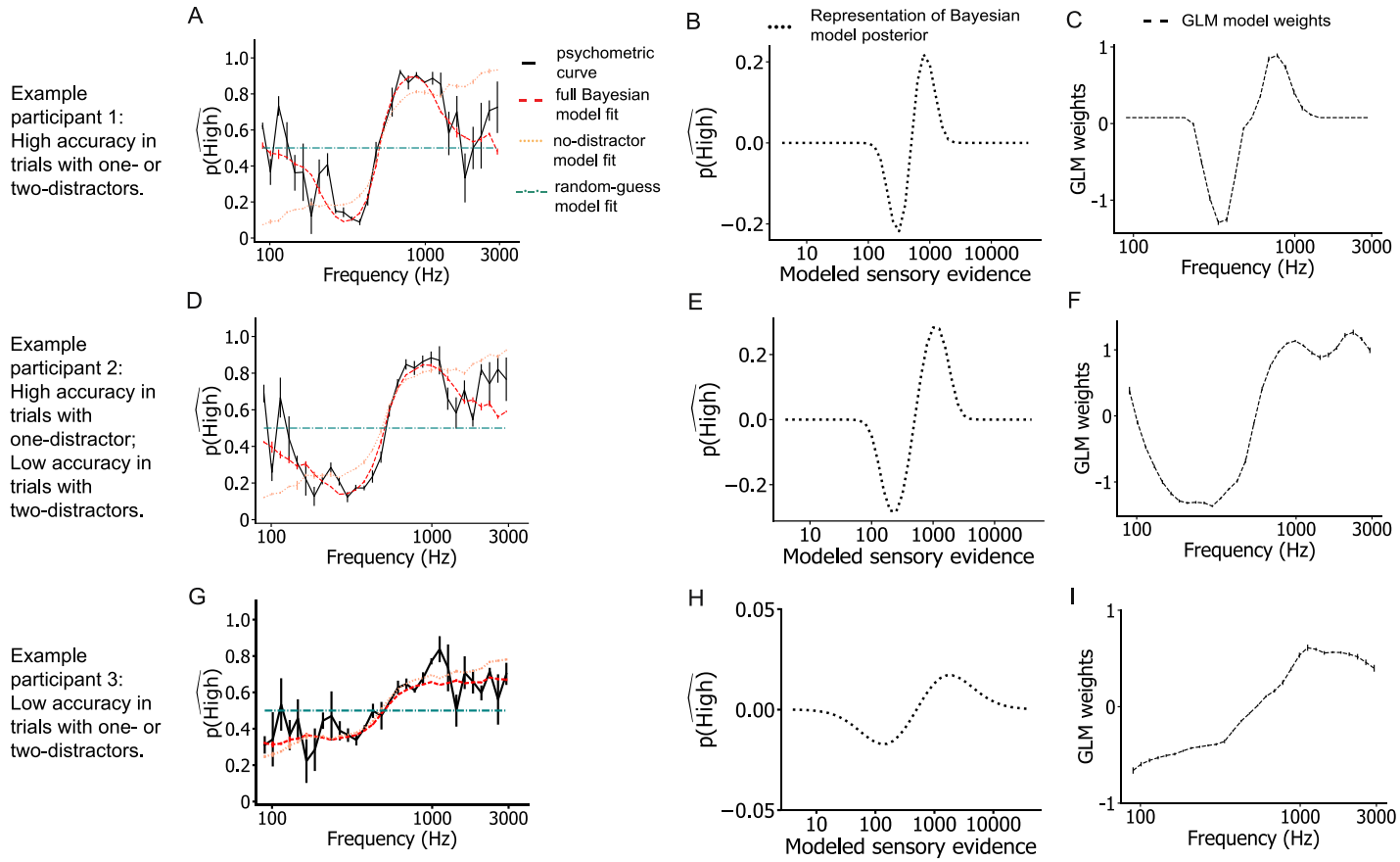


Extended data figure 1: Participants weigh tones according to their relevance; distractors adversely affect accuracy. (A) Psychometric curve averaged across all participants for trials with three signal tones. Because a given signal tone frequency can be played in any order (as the 1st, 2nd or 3rd tone) in a trial sequence, here, we plot the mean of the three resulting $p(\widehat{High})$ values. Average psychometric curves plotted separately for signal and distractor tones (B) for trials with one distractor tone and (C) for trials with two distractor tones. Also, to compute the correlations in Fig. 2B, we only consider the frequency range – marked using black dashed lines – where the category Gaussian probability distributions are greater than the distractor uniform distribution. This allows us to compare signal and distractor correlations across the different trial types. (D-G) Correlations for trials with two distractor tones for 4 example participants. (H) Accuracy for trials from both categories is inversely correlated with number of distractors (Spearman; Table 1). (I) Average psychometric curve for signal and distractor tones when the distractor tone frequency is incongruent with the trial category, plotted for trials with only one distractor. In panels (A-C, H, I), mean and SEM over N=56 participants.

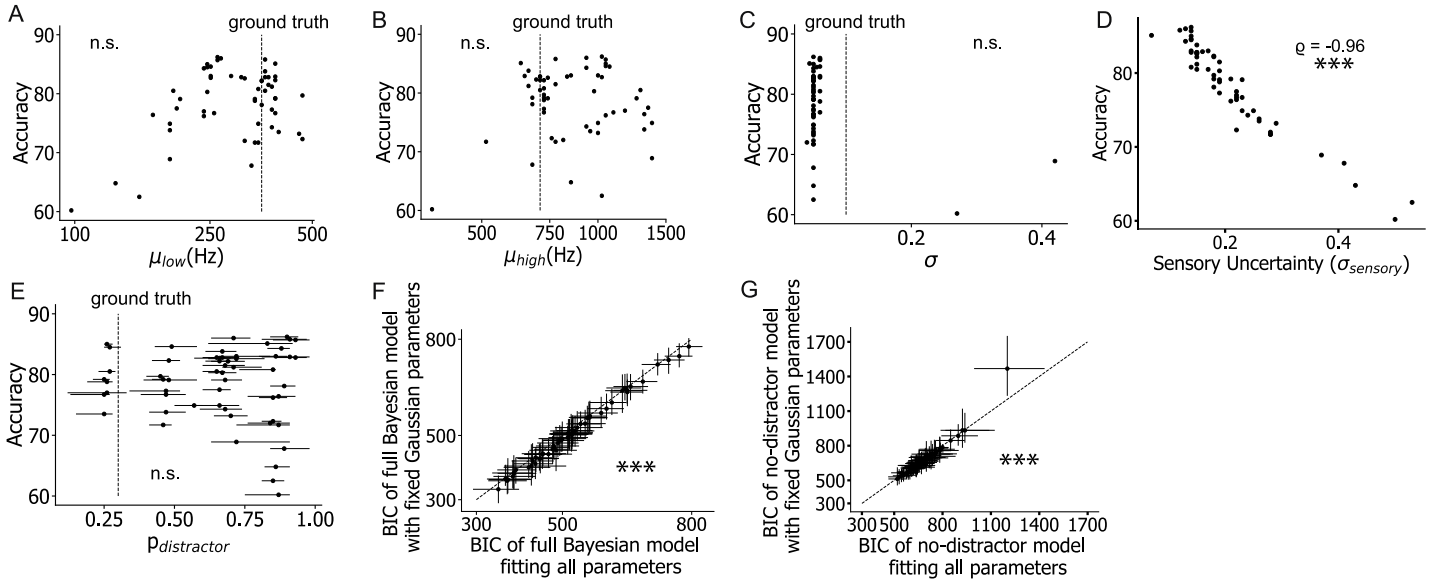


Extended data figure 2: For 3 representative participants, their posteriors derived from the full Bayesian model fits and weights from a fitted generalized linear model.

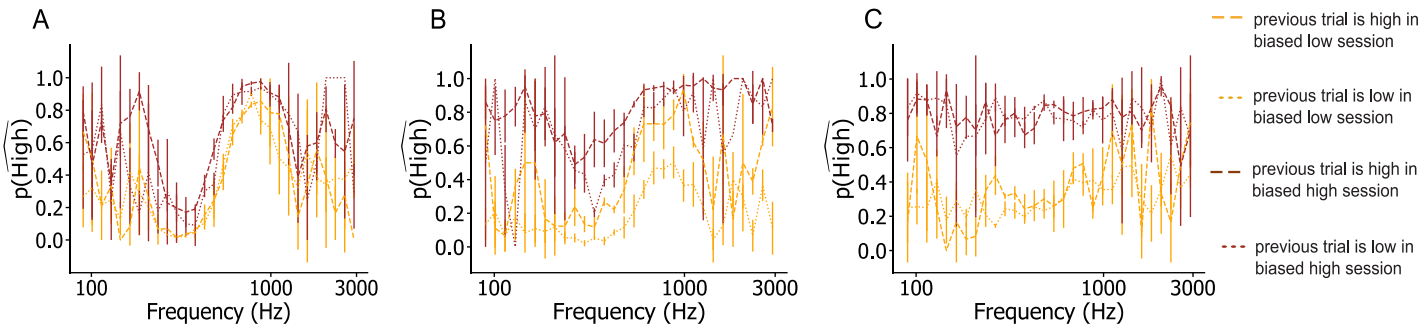
(A) Schematic of the generalized linear model (GLM) that has the frequencies of the three tones in a trial as input and uses the corresponding indicator functions as its predictors to predict the participant's category choice response (see Methods). (B, C, D) Posteriors generated from the full Bayesian model fits to the representative participants' data in Fig. 3. The x-axis is a participant's sensory evidence i.e., their internal representation of the different tone frequencies. The average posterior for each participant is generated in a manner similar to their average psychometric curve. Error bars: standard error of the mean; not large enough to be distinguishable. (E, F, G) GLM weights of tone frequencies in the respective participant's category decision. Mean and error bars of the curve: computed using cross-validation (n=10, see Methods).



Extended data figure 3: Bayesian and GLM analysis of 3 additional participants.
(A, D, G) Similar to Fig. 3 psychometric curves for 3 additional participants; black lines: raw data, red dashed lines: full Bayesian model fit, salmon lines: no-distractor model fit, teal lines: random-guess model fit. (B, E, H) Posteriors generated from the full Bayesian model fits to performance in (A, D, G) respectively. (C, F, I) Relative weights of the tone frequencies determined using the GLM. For all subplots the average curve and its error bars are obtained as described in the captions for Fig. 3 and extended data Fig. 2. (A, B, C): example participant with high task accuracy in trials with either one or two distractors; one distractor: 86.55, two distractors: 72.17. (D, E, F): example participant with high task accuracy in trials with one distractor, and poor accuracy in trials with two distractors; one distractor: 82.91, two distractors: 64.91. (G, H, I): example participant with poor accuracy in both trials with one or two distractors; one distractor: 57.63, two distractors: 64.66.

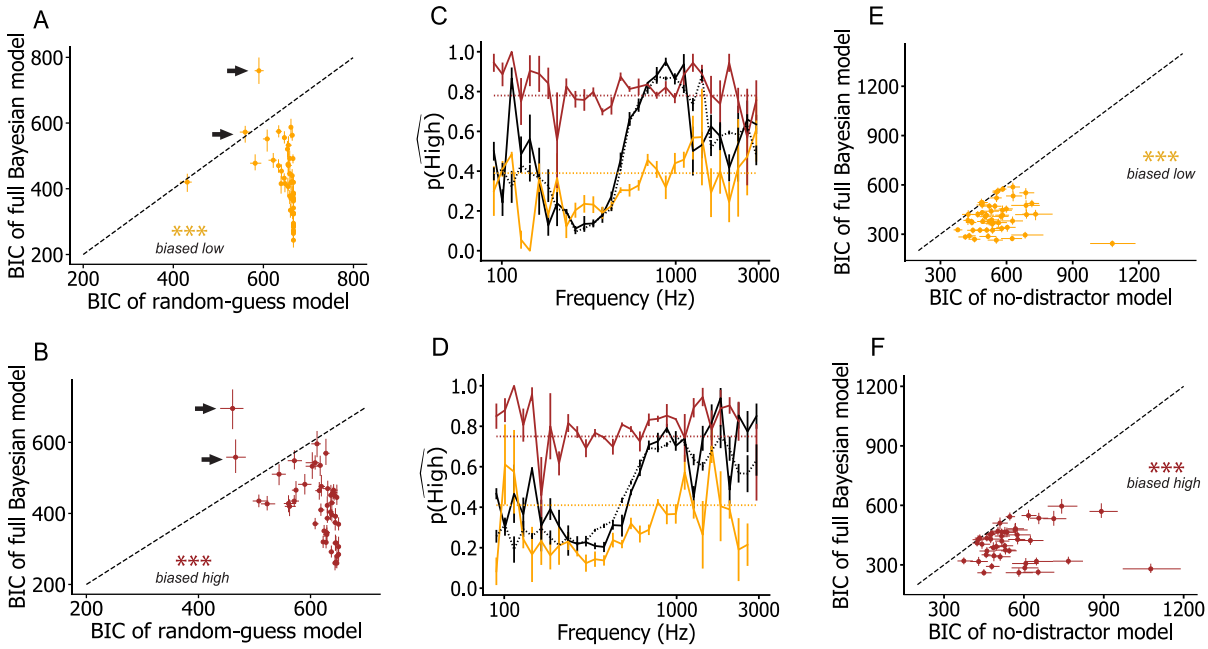


Extended data figure 4: Comparison of the two model fitting procedures.
(A-E) Fitting all parameters of the full Bayesian model to participants' data. Parameters (A) μ_{low}, (B) μ_{high}, (C) σ of the two Gaussians, and (E) $p_{distractor}$ are uncorrelated with accuracy. (D) $\sigma_{sensory}$ i.e., sensory uncertainty is inversely correlated with accuracy. Vertical lines in (A, B, C, E) denote ground truth values which are the parameter values used in the generative process. Panels (A-E) *: Spearman Correlation (Table 1). (F, G) Comparison of the two fitting approaches: for most participants BIC values of the full Bayesian and no-distractor models from the two approaches are roughly similar although fitting all parameters results in slightly higher BIC values. (full Bayesian model, fitting all parameters: Mdn: 514.99, IQR: 115.13, fixing Gaussian Parameters: Mdn: 504.21, IQR: 127.23; no-distractor model, fitting all parameters: Mdn: 660.13, IQR: 111, fixing Gaussian Parameters: Mdn: 659.66, IQR: 117.39) *: two-tailed Wilcoxon signed-rank test (Table 1). In all panels, data point: one participant; $^{\dagger}p<0.05$; $^{*}p<0.01$, $^{**}p<0.001$, $^{***}p<0.0001$.



Extended data figure 5: Participants are variable in their usage of short-term and long-term learning in the *biased low* and *high* sessions.

(A-C) Similar to Figs. 5A-C average psychometric curves conditioned on the previous trial type for three participants in the *biased low* (yellow) and *biased high* (brown) sessions. These curves are from the same participants as Figs. 5E-G. (A) Example participant who anecdotally does not demonstrate any reliance on the previous trial category nor on session-based bias. (B) Example participant who relies on both the previous trial category and on session-based bias. (C) Example participant who relies on session-based bias but not on the previous trial category. $\Delta_{short-term}$: (A) *biased high* session: 0.04, *biased low* session: -0.02. (B) *biased high* session: 0.19, *biased low* session: 0.28. (C) *biased high* session: -0.01, *biased low* session: 0.02.



Extended data figure 6: Most participants account for stimulus relevance in the *biased low* and *high* sessions.
(A) BIC scores of the full Bayesian model are lower than those of the random-guess model for all except 2 participants in the *biased low* session (yellow dots). Data point: one participant; number of participants = 53. (B) Corresponding plot for the *biased high* session (brown dots); number of participants = 48. In (A, B) Error bars: 95% confidence interval. Average BIC scores and error bars are obtained using model fits to 100 balanced datasets (see Methods; Table 1). *: two-sided Wilcoxon sign-rank test (Table 1). (C, D) Psychometric curves for the two participants that have a higher BIC value for the full Bayesian model compared to the random-guess model. Average curves for *biased* sessions (yellow curve – *biased low*, brown curve – *biased high*) are evaluated using 20 subsampled balanced datasets. Error bars: standard error of the mean. Curves for the *unbiased* session, shown using black lines, are similar to Figs.3 B-E. Respective full Bayesian and random-guess model fits are shown using dotted lines. (E, F) For all participants besides the two marked using black arrows, BIC scores of the full Bayesian model are lower than those of the no-distractor model for both the *biased* sessions *: two-sided Wilcoxon sign-rank test (Table 1). In all plots: $^{\dagger}p<0.05$; $*p<0.01$, $**p<0.001$, $***p<0.0001$.

Extended data table 1: Statistical Comparisons.

Comparison	Figure	Center	Spread	N	Test	Test Statistic	p-value	Effect size
ANOVA for effects of tone position on normalized correlation of tone frequency and category choice	2A	First tone position: 0.33 Second tone position: 0.32 Third tone position: 0.34 (mean)	First tone position: 0.012 Second tone position: 0.011 Third tone position: 0.01 (SEM)	56 participants	One-way repeated measures ANOVA	$F_{\text{tone position}}(2,110) = 0.9$	$p_{\text{tone position}} = 0.41$	$\eta^2_{tone\ position} = 0.016$
Correlation of tone frequency with category choice for signal versus distractor tones	2B, C, D	Trials with no distractors Signal correlation: 0.91 Trials with one distractor Signal correlation: 0.87 Distractor correlation: 0.29 Trials with two distractors Signal correlation: 0.78 Distractor correlation: 0.47 (mean)	Trials with no distractors Signal correlation: 0.007 Trials with one distractor Signal correlation: 0.01 Distractor correlation: 0.03 Trials with two distractors Signal correlation: 0.02 Distractor correlation: 0.02 (SEM)	56 participants	Two-sided paired t-test	Trials with one distractor t-stat: -16.21 Trials with two distractors t-stat: -12.71	Trials with one distractor p: 1.32.e-22 Trials with two distractors p: 5.36e-18	Trials with one distractor Cohen's-d: 3.35 Trials with two distractors Cohen's-d: 1.97
ANOVA for effect of trial type on correlation of signal tone frequency and category choice, and on correlation of distractor tone frequency and category choice	2B	Same as above	Same as above	56 participants	One-way repeated measures ANOVA	$F_{\text{correlation of signal tones}}(2,110) = 34.55$ $F_{\text{correlation of distractor tones}}(1,55) = 24.36$	$p_{\text{correlation of signal tones}} = 2.28e\text{-}12$ $p_{\text{correlation of distractor tones}} = 8e\text{-}6$	$\eta^2_{correlation\ of\ signal\ tones} = 0.39$ $\eta^2_{correlation\ of\ distractor\ tones} = 0.31$
Accuracy versus number of distractors *unbiased session: both categories combined*	2E	n/a	n/a	56 participants	Spearman Correlation	$\rho = -1$	$p = 0$	n/a

Accuracy of trials with one distractor versus accuracy of trials with two distractors	2F	One distractor: 79.46 Two distractors: 66.24 (median)	One distractor: 11.26 Two distractors: 6.03 (IQR)	54 participants	Two-sided Wilcoxon signed-rank test	Z = -6.39	p = 1.63e-10	$\frac{Z}{\sqrt{n}} = -0.87$

Accuracy of low category trials vs high category trials (variation in frequency of distractor tone)	2G	Frequency: 100 Hz
		Low – 90.91,
		High – 68.75;
		Frequency: 126 Hz
		Low – 100,
		High – 73.33;
		Frequency: 159 Hz
		Low – 88.89,
		High – 68.75;
		Frequency: 200 Hz
		Low – 100,
		High – 71.43;
		Frequency: 252 Hz
		Low – 100,
		High – 71.43;
		Frequency: 317 Hz
		Low – 100,
		High – 66.67;
		Frequency: 399 Hz
		Low – 85.71,
		High – 71.43;
		Frequency: 502 Hz
		Low – 100,
		High – 75;
		Frequency: 631 Hz
		Low – 66.67,
		High – 83.33;
			Frequency: 100 Hz
			Low – 18.18,
			High – 25;
			Frequency: 126 Hz
			Low – 0,
			High – 20;
			Frequency: 159 Hz
			Low – 11.11,
			High – 28.13;
			Frequency: 200 Hz
			Low – 20,
			High – 21.43;
			Frequency: 252 Hz
			Low – 20,
			High – 42.86;
			Frequency: 317 Hz
			Low – 25,
			High – 14.05;
			Frequency: 399 Hz
			Low – 17.86,
			High – 17.86;
			Frequency: 502 Hz
			Low – 25,
			High – 25;
			Frequency: 631 Hz
			Low – 33.33,
			High – 16.67;
				56 participants	Two-sided Wilcoxon signed-rank test	Frequency: 100 Hz
						Z = 6.16;
						Frequency: 126 Hz
						Z = 6.27;
						Frequency: 159 Hz
						Z = 5.7;
						Frequency: 200 Hz
						Z = 4.09;
						Frequency: 252 Hz
						Z = 5.31
						Frequency: 317 Hz
						Z = 5.31
						Frequency: 399 Hz
						Z = 5.15
						Frequency: 502 Hz
						Z = 1.9
						Frequency: 631 Hz
						Z = -4.55
						Frequency: 795 Hz
						Z = -6.23
						Frequency: 1001 Hz
						Z = -6.25
						Frequency: 1259 Hz
						Z = -4.58
							Frequency: 100 Hz
							p = 7.4e-10;
							Frequency: 126 Hz
							p = 3.5e-10;
							Frequency: 159 Hz
							p = 1.22e-8;
							Frequency: 200 Hz
							p = 4.4e-5;
							Frequency: 252 Hz
							p = 1.09e-7
							Frequency: 317 Hz
							p = 1.11e-7
							Frequency: 399 Hz
							p = 2.65e-7
							Frequency: 502 Hz
							p = 0.057
							Frequency: 631 Hz
							p = 5.3e-6
							Frequency: 795 Hz
							p = 4.59e-10
							Frequency: 1001 Hz
							p = 4e-10
							Frequency: 1259 Hz
							p = 4.67e-6	

		Frequency: 795 Hz Low – 63.64, High – 100; Frequency: 1001 Hz Low – 42.86, High – 100; Frequency: 1259 Hz Low – 66.67, High – 100; Frequency: 1585 Hz Low – 72.73, High – 100; Frequency: 1996 Hz Low – 75, High – 100; Frequency: 2551 Hz Low – 75, High – 94.44; (median)	Frequency: 795 Hz Low – 10, High – 16.67; Frequency: 1001 Hz Low – 28.57, High – 14.29; Frequency: 1259 Hz Low – 16.67, High – 25; Frequency: 1585 Hz Low – 27.27, High – 14.29; Frequency: 1996 Hz Low – 37.5, High – 25; Frequency: 2551 Hz Low – 25, High – 16.67; (IQR)			Frequency: 1585 Hz Z = -6.16 Frequency: 1996 Hz Z = -4.15 Frequency: 2551 Hz Z = -4.75	Frequency: 1585 Hz p = 7.06e-10 Frequency: 1996 Hz p = 3.31e-5 Frequency: 2551 Hz p = 2.03e-6	
BIC scores for participant 1	3B	full Bayesian: 243.09 no-distractor: 436.34 random-guess: 415.89 (median)	full Bayesian: 47.57 no-distractor: 93.47 random-guess: 0 (95th Percentile – 5th Percentile)	100 bootstrapped trials	Two-sided Wilcoxon signed-rank test	Z (full Bayesian versus random-guess) = -8.68 Z (full Bayesian versus no-distractor) = -8.68	p (full Bayesian versus random-guess) = 3.96e-18 p (full Bayesian versus no-distractor) = 3.96e-18	$\frac{Z}{\sqrt{n}}$ (full Bayesian versus random-guess) = -0.87 $\frac{Z}{\sqrt{n}}$ (full Bayesian versus no-distractor) = -0.87
BIC scores for participant 2	3C	full Bayesian: 221.86 no-distractor: 263.29 random-guess: 415.89 (median)	full Bayesian: 45.21 no-distractor: 58.79	100 bootstrapped trials	Two-sided Wilcoxon signed-rank test	Z (full Bayesian versus random-guess) = -8.68	p (full Bayesian versus random-guess) = 3.96e-18	$\frac{Z}{\sqrt{n}}$ (full Bayesian versus

			random-guess: 4.78 (95th Percentile – 5th Percentile)			Z (full Bayesian versus no-distractor) = -8.46	p (full Bayesian versus no-distractor) = 2.67e-17	random-guess) = -0.87 $\frac{Z}{\sqrt{n}}$ (full Bayesian versus no-distractor) = -0.85
BIC scores for participant 3	3D	full Bayesian: 358.03 no-distractor: 366.72 random-guess: 409.89 (median)	full Bayesian: 40.93 no-distractor: 39.34 random-guess: 15.35 (95th Percentile – 5th Percentile)	100 bootstrapped trials	Two-sided Wilcoxon signed-rank test	Z (full Bayesian versus random-guess) = -8.68 Z (full Bayesian versus no-distractor) = -5.06	p (full Bayesian versus random-guess) = 3.96e-18 p (full Bayesian versus no-distractor) = 4.13e-7	$\frac{Z}{\sqrt{n}}$ (full Bayesian versus random-guess) = -0.87 $\frac{Z}{\sqrt{n}}$ (full Bayesian versus no-distractor) = -0.51
BIC scores of full Bayesian model versus random-guess model	4A	full Bayesian: 502.41 random-guess: 844.57 (median)	full Bayesian: 127.23 random-guess: 0 (IQR)	56 participants	Two-sided Wilcoxon signed-rank test	Z = -6.51	p = 7.55e-11	$\frac{Z}{\sqrt{n}} = -0.87$
BIC scores of full Bayesian model versus no-distractor model	4B	full Bayesian: 502.41 no-distractor: 659.66 (median)	full Bayesian: 127.23 no-distractor: 117.39 (IQR)	56 participants	Two-sided Wilcoxon signed-rank test	Z = -6.49	p = 8.41e-11	$\frac{Z}{\sqrt{n}} = -0.86$
Outlier participant 1	4B	full Bayesian: 310.75 no-distractor: 312.29 (median)	full Bayesian: 46.89 no-distractor: 45.29 (95th Percentile – 5th Percentile)	100 bootstrapped trials	Two-sided Wilcoxon signed-rank test	Z (full Bayesian versus no-distractor) = -0.47	p (full Bayesian versus no-distractor) = 0.64	$\frac{Z}{\sqrt{n}}$ (full Bayesian versus no-distractor) = -0.047
Outlier participant 2	4B	full Bayesian: 351.1 no-distractor: 354.53 (median)	full Bayesian: 34.84 no-distractor: 41.51	100 bootstrapped trials	Two-sided Wilcoxon signed-rank test	Z (full Bayesian versus no-distractor) = -0.28	p (full Bayesian versus no-distractor) = 0.77	$\frac{Z}{\sqrt{n}}$ (full Bayesian versus no-distractor) = -0.028

			(95th Percentile – 5th Percentile)					
Accuracy versus difference in BIC scores of full Bayesian and no-distractor models	4C	n/a	n/a	56 participants	Spearman correlation	ρ = -0.85 95% confidence interval for both = [-0.91,-76]	p = 8.53e-17	n/a
2-factor model: Sigmoidicity ~ sensory uncertainty + $p_{distractor}$ + sensory uncertainty * $p_{distractor}$	4E	Model coefficients Estimate	Model coefficients Standard error	56 participants	Two-sided paired t-test	Model coefficients t-stat	Model coefficients p-value	n/a
		intercept: 0.34 sensory uncertainty: 0.97 $p_{distractor}$: -0.61 sensory uncertainty* $p_{distractor}$: 1.23	intercept: 0.056 sensory uncertainty: 0.2 $p_{distractor}$: 0.082 sensory uncertainty* $p_{distractor}$: 0.29			intercept: 6.09 sensory uncertainty: 4.8 $p_{distractor}$: -7.4 sensory uncertainty* $p_{distractor}$: 4.2	intercept: 0 sensory uncertainty: 0 $p_{distractor}$: 0 sensory uncertainty* $p_{distractor}$: 0	
2-factor model: Accuracy ~ sensory uncertainty + $p_{distractor}$ + sensory uncertainty * $p_{distractor}$	4F	Model coefficients Estimate	Model coefficients Standard error	56 participants	Two-sided paired t-test	Model coefficients t-stat	Model coefficients p-value	n/a
		intercept: 92.41 sensory uncertainty: -65 $p_{distractor}$: -1.08 sensory uncertainty* $p_{distractor}$: 4.06	intercept:1.28 sensory uncertainty: 4.6 $p_{distractor}$: 1.88 sensory uncertainty* $p_{distractor}$: 6.7			intercept: 72.36 sensory uncertainty: -14.14 $p_{distractor}$: -0.58 sensory uncertainty* $p_{distractor}$: 0.61	intercept: 0 sensory uncertainty: 0 $p_{distractor}$: 0.57 sensory uncertainty* $p_{distractor}$: 0.55	
$\Delta_{\text{short-term}}$ *unbiased session*	5D	0.12 (median)	0.15 (IQR)	56 participants	Two-sided Wilcoxon signed-rank test	Z = 6.04	p = 1.58e-9	$\frac{Z}{\sqrt{n}} = 0.81$
$p(\widehat{\text{high}})_{\text{central tone}}$	5H	*biased high*: 0.42 *biased low*: 0.62 (median)	*biased high*: 0.11	*biased low*: 53 participants	Two-sided Wilcoxon rank sum test for	Z = 8.63	p = 6.34e-18	$\frac{Z}{\sqrt{n}} = 1.19$

			biased low: 0.085 (IQR)	*biased high*: 48 participants	$\Delta_{\text{long-term}}$			
Accuracy versus time constant of short-term learning *unbiased session*	6B	n/a	n/a	56 participants	Spearman Correlation	ρ = -0.27 95% confidence interval = [-0.5, -0.01]	p = 0.042	n/a
$p_{distractor}$ versus weight of preceding trial *unbiased session*	6C	n/a	n/a	56 participants	Spearman Correlation	ρ = -0.36 95% confidence interval = [-0.57, -0.1]	p = 0.007	n/a
Long-term learning versus weight of preceding trial *biased sessions*	6D	n/a	n/a	*biased low*: 53 participants *biased high*: 48 participants	Spearman Correlation	ρ = -0.15 95% confidence interval = [-0.34, 0.05]	p = 0.15	n/a
Accuracy versus time constant of short-term learning *biased sessions*	6E	n/a	n/a	*biased low*: 53 participants *biased high*: 48 participants	Spearman Correlation	ρ = -0.32 95% confidence interval = [-0.48, -0.12]	p = 0.0016	n/a
$p_{distractor}$ versus weight of preceding trial *biased sessions*	6F	n/a	n/a	*biased low*: 53 participants *biased high*: 48 participants	Spearman Correlation	ρ = -0.47 95% confidence interval = [-0.61, -0.3]	p = 1e-6	n/a
$p_{distractor}$ versus long-term learning *biased sessions*	6G	n/a	n/a	*biased low*: 53 participants *biased high*: 48 participants	Spearman Correlation	ρ = -0.25 95% confidence interval = [-0.43, -0.06]	p = 0.013	n/a
Accuracy versus sensory uncertainty *biased sessions*	6H	n/a	n/a	*biased low*: 53 participants	Spearman Correlation	ρ = -0.75 95% confidence interval = [-0.82, -0.65]	p = 2.38e-19	n/a

				biased high: 48 participants				
Accuracy versus number of distractors *unbiased session: low and high categories*	Extended data 1H	n/a	n/a	56 participants	Spearman Correlation	$\rho_{\text{low category}} = -1$ $\rho_{\text{high category}} = -1$	$p_{\text{low category}} = 0$ $p_{\text{high category}} = 0$	n/a
BIC scores for participant 1	Extended data 3A	full Bayesian: 181.97 no-distractor: 460.32 random-guess: 415.89 (median)	full Bayesian: 45.39 no-distractor: 169.68 random-guess: 0 (95th Percentile – 5th Percentile)	100 bootstrapped trials	Two-sided Wilcoxon signed-rank test	Z (full Bayesian versus random-guess) = -8.68 Z (full Bayesian versus no-distractor) = -8.68	p (full Bayesian versus random-guess) = 3.96e-18 p (full Bayesian versus no-distractor) = 3.96e-18	$\frac{Z}{\sqrt{n}}$ (full Bayesian versus random-guess) = -0.87 $\frac{Z}{\sqrt{n}}$ (full Bayesian versus no-distractor) = -0.87
BIC scores for participant 2	Extended data 3D	full Bayesian: 232.48 no-distractor: 318.37 random-guess: 415.89 (median)	full Bayesian: 39.44 no-distractor: 69.63 random-guess: 0 (95th Percentile – 5th Percentile)	100 bootstrapped trials	Two-sided Wilcoxon signed-rank test	Z (full Bayesian versus random-guess) = -8.68 Z (full Bayesian versus no-distractor) = -8.68	p (full Bayesian versus random-guess) = 3.96e-18 p (full Bayesian versus no-distractor) = 3.96e-18	$\frac{Z}{\sqrt{n}}$ (full Bayesian versus random-guess) = -0.87 $\frac{Z}{\sqrt{n}}$ (full Bayesian versus no-distractor) = -0.87
BIC scores for participant 3	Extended data 3G	full Bayesian: 378.93 no-distractor: 386.01 random-guess: 415.89 (median)	full Bayesian: 28.87 no-distractor: 30.6 random-guess: 1.58 (95th Percentile – 5th Percentile)	100 bootstrapped trials	Two-sided Wilcoxon signed-rank test	Z (full Bayesian versus random-guess) = -8.68 Z (full Bayesian versus no-distractor) = -4.16	p (full Bayesian versus random-guess) = 3.96e-18 p (full Bayesian versus no-distractor) = 0.000032	$\frac{Z}{\sqrt{n}}$ (full Bayesian versus random-guess) = -0.87 $\frac{Z}{\sqrt{n}}$ (full Bayesian versus no-

								distractor) = -0.42
Accuracy versus μ_{low}	Extended data 4A	n/a	n/a	56 participants	Spearman correlation	ρ = 0.14 95% confidence interval = [-0.13, 0.39]	p = 0.31	n/a
Accuracy versus μ_{high}	Extended data 4B	n/a	n/a	56 participants	Spearman correlation	ρ = -0.077 95% confidence interval = [-0.33, 0.19]	p = 0.58	n/a
Accuracy versus σ	Extended data 4C	n/a	n/a	56 participants	Spearman correlation	ρ = 0.17 95% confidence interval = [-0.1, 0.42]	p = 0.2	n/a
Accuracy versus sensory uncertainty	Extended data 4D	n/a	n/a	56 participants	Spearman correlation	ρ = -0.95 95% confidence interval = [-0.97, -0.92]	p = 4.21e-29	n/a
Accuracy versus $p_{distractor}$	Extended data 4E	n/a	n/a	56 participants	Spearman correlation	ρ = 0.08 95% confidence interval = [-0.19, 0.34]	p = 0.56	n/a
BIC scores of full Bayesian model with fixed Gaussian parameters versus when fitting all parameters	Extended data 4F	Fixing Gaussian parameters: 502.41 Fitting all parameters: 514.99 (median)	Fixing Gaussian parameters: 127.23 Fitting all parameters: 115.13 (IQR)	56 participants	Two-sided Wilcoxon sign-rank test	Z = -6.42	p = 1.37e-10	$\frac{Z}{\sqrt{n}} = -0.86$
BIC scores of no-distractor model with fixed Gaussian parameters versus when fitting all parameters	Extended data 4G	Fixing Gaussian parameters: 659.66 Fitting all parameters: 660.13 (median)	Fixing Gaussian parameters: 117.39 Fitting all parameters: 111 (IQR)	56 participants	Two-sided Wilcoxon sign-rank test	Z = -4.45	p = 8.76e-6	$\frac{Z}{\sqrt{n}} = -0.59$
BIC scores of full Bayesian model versus random-guess model *biased low session*	Extended data 6A	full Bayesian: 408.47 random-guess: 663.87 (median)	full Bayesian: 140.54 random-guess: 13.18 (IQR)	53 participants	Two-sided Wilcoxon signed-rank test	Z = -6.19	p = 5.92e-10	$\frac{Z}{\sqrt{n}} = -0.85$

BIC scores of full Bayesian model versus random-guess model *biased high session*	Extended data 6B	full Bayesian: 424.81 random-guess: 629 (median)	full Bayesian: 122.23 random-guess: 35.51 (IQR)	48 participants	Two-sided Wilcoxon signed-rank test	Z = -5.65	p = 1.59e-8	$\frac{Z}{\sqrt{n}} = -0.82$
BIC scores of full Bayesian model versus no-distractor model *biased low session*	Extended data 6E	full Bayesian: 386.99 no-distractor: 534.37 (median)	full Bayesian: 145.69 no-distractor: 97 (IQR)	51 participants	Two-sided Wilcoxon signed-rank test	Z = -6.11	p = 9.87e-10	$\frac{Z}{\sqrt{n}} = -0.86$
BIC scores of full Bayesian model versus no-distractor model *biased high session*	Extended data 6F	full Bayesian: 414.83 no-distractor: 516.55 (median)	full Bayesian: 145.06 no-distractor: 115.61 (IQR)	46 participants	Two-sided Wilcoxon signed-rank test	Z = -5.87	p = 4.3e-9	$\frac{Z}{\sqrt{n}} = -0.87$