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 or music, into classes, such as syllables or notes. Auditory categorization depends not only on the acoustic waveform, but also on variability and uncertainty in how the listener perceives the sound – including sensory and stimulus uncertainty, the listener's estimated relevance of the particular sound to the task, and their ability to learn the past statistics of the acoustic environment. Whereas these factors have been studied in isolation, whether and how these factors *interact* to shape categorization remains unknown. Here, we measured human participants' performance on a multi-tone categorization task and modeled each participant's behavior using a Bayesian framework. Task-relevant tones contributed more to category choice than task-irrelevant tones, confirming that participants combined information about sensory features with task relevance. Conversely, participants' poor estimates of task-relevant tones or high-sensory uncertainty adversely impacted category choice. Learning the statistics of sound category over both short- and long-timescales also affected decisions, biasing the decisions toward the overrepresented category. The magnitude of this effect correlated inversely with participants' relevance estimates. Our results demonstrate that individual participants idiosyncratically weigh sensory uncertainty, task relevance, and statistics over both short and long timescales, providing a novel understanding of and a computational framework for how sensory decisions are made under several simultaneous behavioral demands.

Introduction

Making sensory decisions in everyday settings is a complex task due to the presence of multiple forms of uncertainty(1–8). One of many such decisions is identifying what a friend said when the conversation takes place in a crowded restaurant versus in a quiet room. Despite the apparent ease by which we can accomplish this task, it is a very complicated computational process involving many factors. First, the sensory transduction process and neural-processing stages are noisy, generating a form of uncertainty called *sensory uncertainty*(9–12). A second form of uncertainty is *stimulus relevance*: of all stimuli within a sensory scene, only some are relevant to the current decision(13–18). For example, when trying to identify a particular sound uttered by our friend, we have to disregard sounds uttered by other speakers, which can be considered to be distractors. A third form of uncertainty is *stimulus uncertainty*, which is a listener's uncertainty in estimating the variability in a sound source's generation of the nominal same stimulus(19). Even once our sensory system isolates the signals that were produced by our friend from background noise or irrelevant signals, we may have uncertainty regarding the correspondence between the relevant signals and a specific sound or a word. Furthermore, when presented with any form of uncertainty, observers may rely on *the learned past statistics of stimuli*(20–26) to make their decisions, such as their long-term knowledge of how their friend typically talks or their short-term knowledge of their friend's recent speech sounds(27). Indeed, observers can learn both *short-* (28,29) and *long-term stimulus statistics*(30–32) and this learning is considered crucial for efficient decision making(33,34).

Whereas the effects of the factors – sensory uncertainty, stimulus relevance, stimulus uncertainty and short- and long-term statistics – on sensory decision-making, such as categorization, have previously been studied *separately*(19,35), in everyday situations we often encounter them *simultaneously*. Thus, these factors may *interact* with each other, differentially affecting category decisions. For example, 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 prior expectations of what our friend is saying may play a greater role when our sensory uncertainty or stimulus uncertainty is high or when there are more distracting sounds in the environment 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 factors shapes sensory decisions, we measured the performance of human participants on an auditory categorization task(36–39). 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 how their task performance was affected by learning. Participants categorized trial sequences of three tones, as either high or low in a two-alternative forced choice

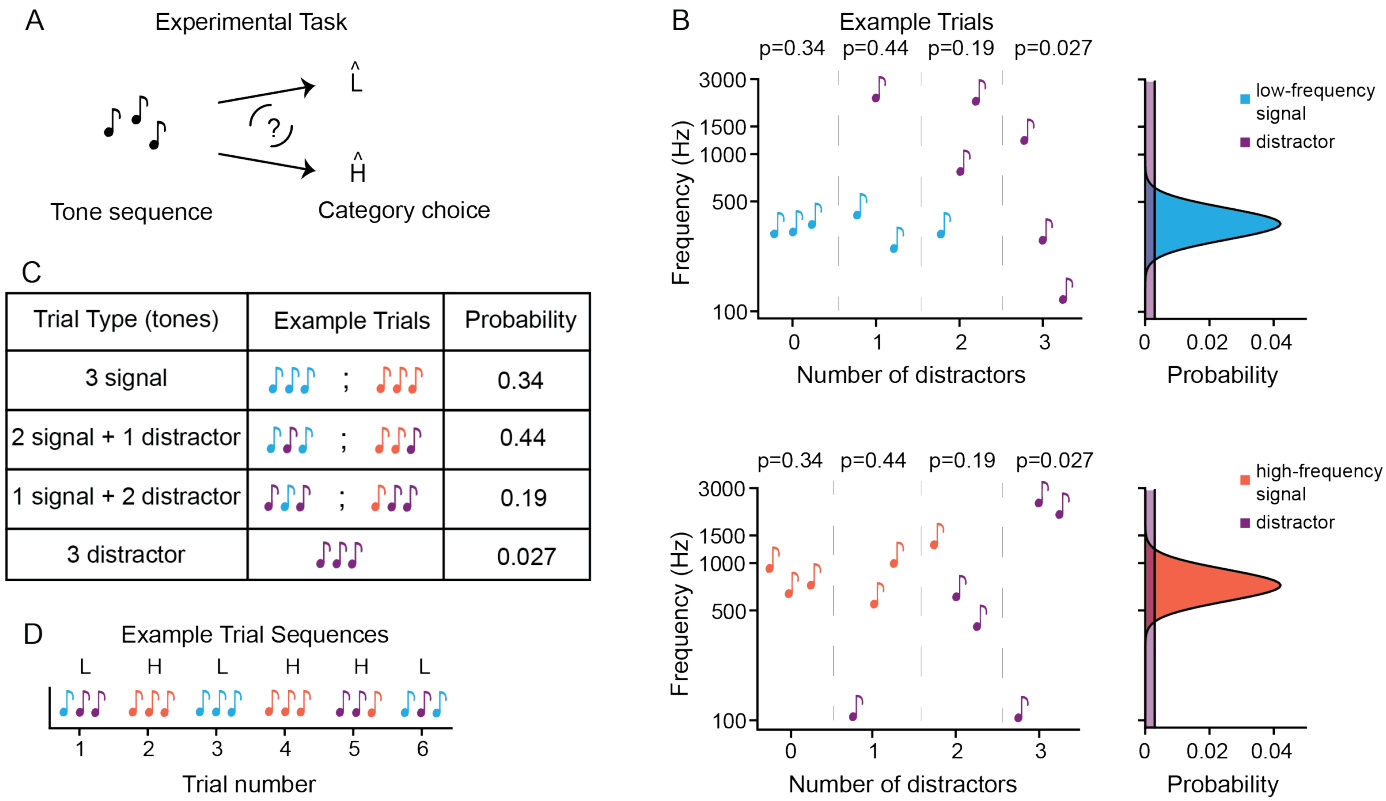
task, as we varied the number of task-irrelevant (distractor) and task-relevant (category-specific) tones across the trials. Additionally, to test learning, in a subset of experimental sessions, we modulated the proportion of trials biased toward the high or the low category.

Because it has been suggested that perceptual systems optimally integrate information from relevant inputs while ignoring the irrelevant ones(40), we characterized participants' behavior using normative Bayesian models of sensory decision-making(41–46). The Bayesian formalism is a well-established framework to understand complex interactions among many factors governing decision-making(13,14,30,31,47). On average, when categorizing auditory stimuli, participants integrated information across all three tones but weighed them by their task relevance. Stimulus uncertainty was similar across participants. However, 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 learning short-term statistics (i.e., stimulus category of the previous trial) and learning long-term statistics (i.e., session-level bias toward high versus low category). However, individual participants' knowledge of short- and long-term statistics was inversely correlated with their estimates of the likelihood of irrelevant stimuli. Together, our results demonstrate that humans differentially combine information about sensory uncertainty, stimulus relevance, stimulus uncertainty, and short-term and long-term learning, providing an integrated framework for understanding how the interplay of different mechanisms shapes auditory processing.

Results

Participants weigh tones according to stimulus relevance

Auditory categorization can depend on multiple factors, such as sensory uncertainty, stimulus relevance, stimulus uncertainty, and stimulus statistics that can vary over both short and long timescales. Because these factors often co-occur, they may interact with each other to affect decision-making and considering their relations they, indeed, computationally should interact. Here, we designed a categorization task to measure how these factors interacted in decision-making. 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 a signal or distractor (Fig 1B-D).



Trial Type (tones)	Example Trials	Probability
3 signal	;	0.34
2 signal + 1 distractor	;	0.44
1 signal + 2 distractor	;	0.19
3 distractor		0.027

Fig. 1: Schematic diagram of the auditory categorization task.
(A) A participant categorizes a given three-tone trial sequence as high or low. (B) Example low (top) and high (bottom) category trials from an *unbiased* session. The underlying probability distribution for signal tones is Gaussian (low-frequency Gaussian: light blue; high-frequency Gaussian: red) and for distractor tones is uniform (purple). 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 (red), and distractor tones from the uniform distribution (purple). (C) The 4 types of trial combinations in the task and their corresponding probabilities. (D) Example trial sequences from one of the participant's data in the *unbiased* session.

We first tested whether participants accounted for relevance of individual tones when categorizing the tone sequences. Compared to signal tones, distractors should have less of an impact on participants' decisions, because they have no relevance for the task. Indeed, signal tones were more strongly correlated with participants' choice probability than distractor tones (Fig 2A; two-sided paired t-test, trials with one distractor: p = 1.32e-22; trials with two distractors: p = 5.36e-18; Figs S1C-H and Table S1). This confirms that participants used tone relevance when making decisions.

Conversely, because distractor tones may still be integrated in the decision with some probability(48,49), they may impair participants' performance by providing information that should have been ignored. Indeed, as the number of distractors increased, the influence of distractor tones on category choice also increased whereas that of signal tones decreased (Fig 2A; Figs S1B-D; one-way rmANOVA, significant main effect for number of distractors on correlations with signal tone frequencies: $F(2,110) = 34.55$, $p = 2.28e\text{-}12$, and on correlations with distractor tone frequencies: $F(1,55) = 24.36$, $p = 8e\text{-}6$). Additionally, accuracy decreased as the number of distractors increased (Fig 2B; N = 56, correlation of accuracy with the number of distractors: Spearman, $\varrho = -1$, $p = 0$). These findings collectively demonstrate that because participants were uncertain about the relevance of the different tones, the distractors adversely impacted their decisions.

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 2C; Table S1 – 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 (Fig S1J). A qualitative comparison of these curves to the psychometric curve for trials with only signal tones (Fig S1B) 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.

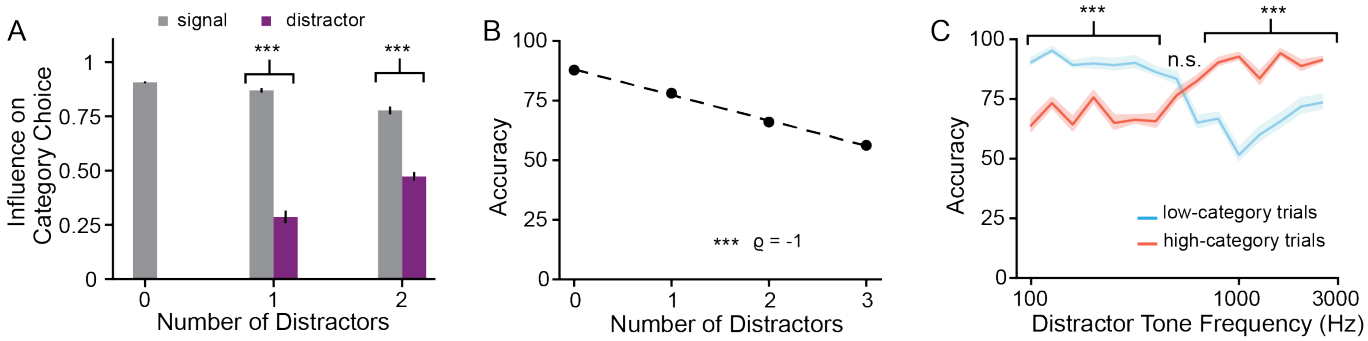


Fig. 2: On average, participants weighted tones according to their relevance when making category decisions. (A) For both signal and distractor tones, their influence on category choice is computed using correlation of the respective tone frequencies with category choice probability (Methods). Error bars: standard error of the mean. Statistics: two-sided paired t-test between influence of signal and distractor tones. (B) Average accuracy in *unbiased* trials decreased with the number of distractor tones. Error bars: standard error of the mean. Statistics: Spearman correlation. (C) 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. Statistics: two-tailed Wilcoxon signed-rank test between accuracy in high- and low-category trials computed at each distractor frequency. N = 56 participants. †p<0.05, *p<0.01, **p<0.001, ***p<0.0001.

Participants exhibit high individual variability in performance

Whereas the previous analysis confirmed the expected effects of tone relevance on the mean performance of the participants, choice correlation varies idiosyncratically across individuals. This diversity may be due to participants' distinct internal priors, estimates of relevance, measure of stimulus uncertainty or their sensory uncertainty. Indeed, there was significant heterogeneity in the influence of tone frequency for both signal and distractor tones, on category choice probability (Figs 3A, B) and on accuracy (Fig 3C), across trials with one and two distractors. 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, stimulus uncertainty, sensory uncertainty, and learning of stimulus statistics on the average

performance across the population but also on inter-participant variability. This allowed us to probe how these different real-life factors might mediate individual participant's decision-making process.

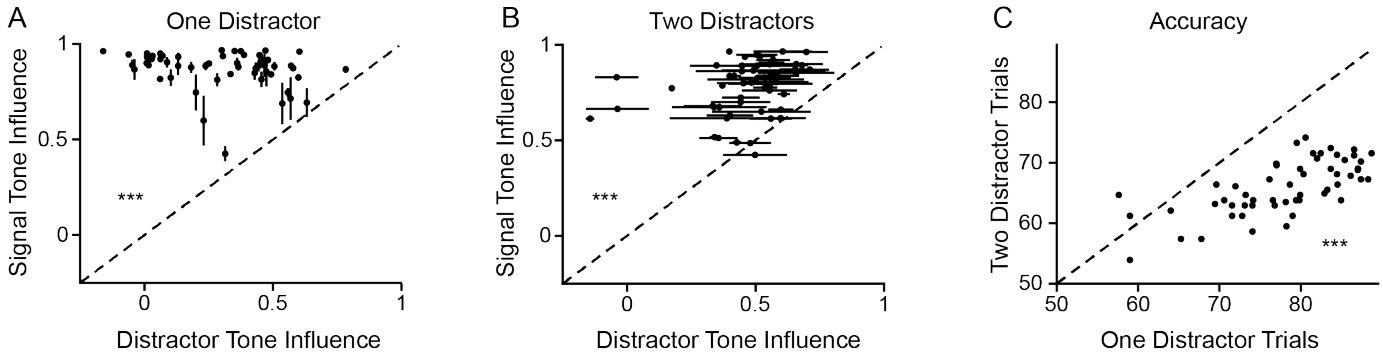


Fig. 3: Individual participants differed in their measure of relevance.
(A, B) For each participant, influence of tone type on category choice is computed using correlation of respective tone frequencies with category choice probability. (A) Influence of signal versus distractor tones for trials with one distractor. Statistics: two-tailed Wilcoxon signed-rank test for the 56 participants with greater signal tone frequency correlation. (B) Influence of signal versus distractor tones for trials with two distractors. Statistics: two-tailed Wilcoxon signed-rank test for the 51 participants with greater signal tone frequency correlation. (C) Categorization accuracy for the tone sequences with one versus two distractor tones. Statistics: two-tailed Wilcoxon signed-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. $^{\dagger}p<0.05$, $^{*}p<0.01$, $^{**}p<0.001$, $^{***}p<0.0001$.

Mapping out decision strategies for auditory categorization

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 a Bayesian approach. We first discuss how tone frequency should affect individual category choice probability and then lay out three models for understanding decisions and their predictions for this relation.

On a standard two alternative forced choice (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. As the tone frequencies move further away from the decision boundary, tones become more informative about the trial's category (Figs 4A, B) 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, we expect participants' performance to be substantially different. Because tones whose frequencies are far from the category-specific Gaussians are more likely to be distractors than signal tones (Figs 4C, D), participants may assign lower relevance to those tones. This effect would result in poorer accuracy for tones lying at the tails of the Gaussian distributions (Fig. 4D).

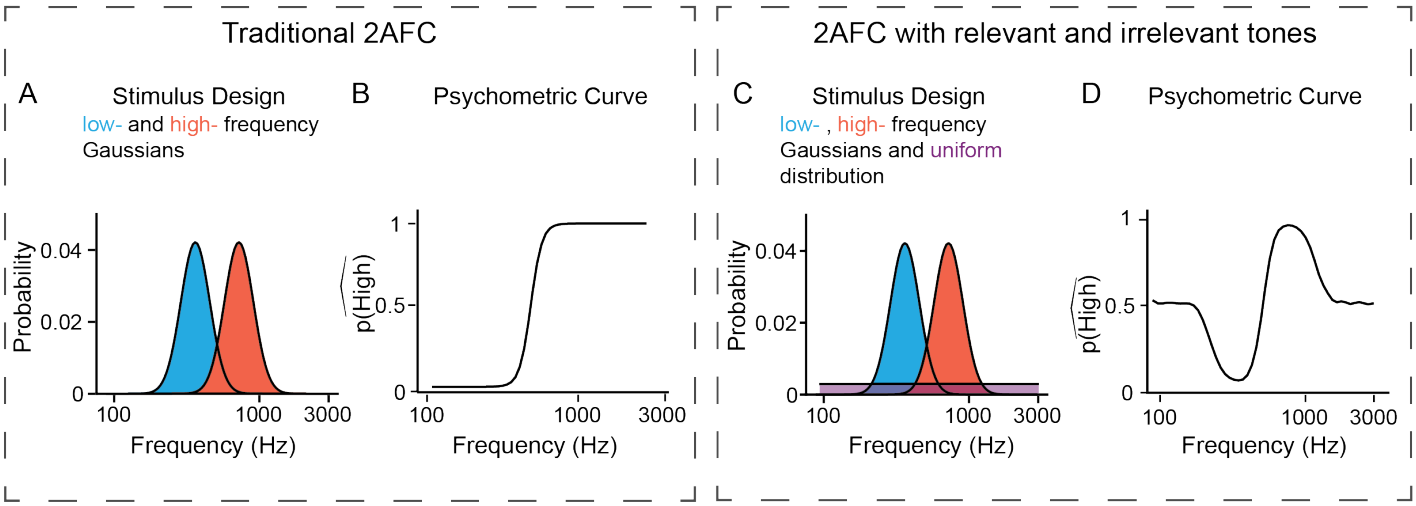
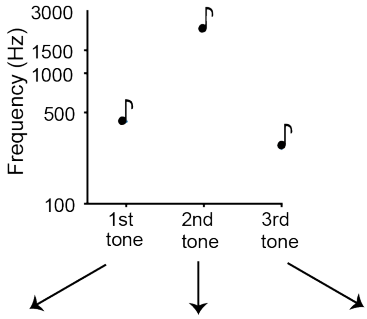


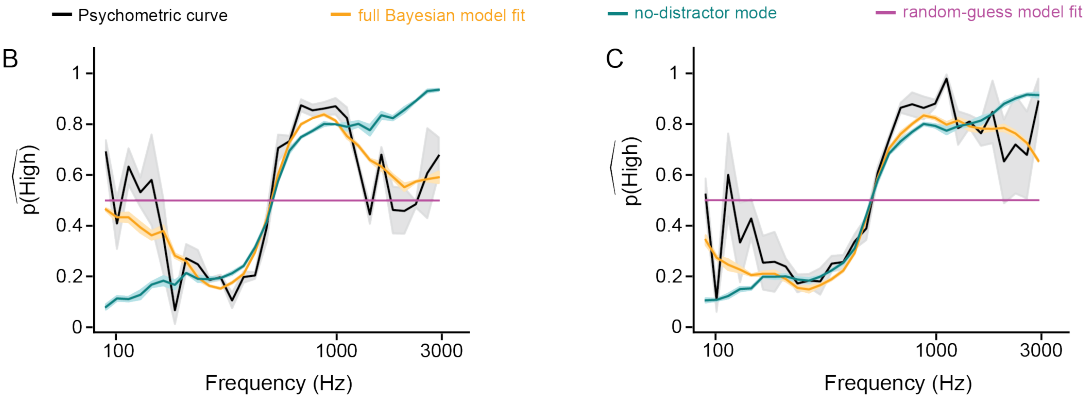
Fig. 4: Schematic diagram demonstrating the effect of distractor tones on a participant's psychometric curve.
(A, B) If a trial contained only signal tones that were drawn from either the low- or high- frequency signal Gaussian distributions, the psychometric curve would be a sigmoid as in the traditional 2AFC task. (C, D) 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 theoretically plateaus at 0.5 because frequency values at the tails of the Gaussian distributions 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 weighted equally (Fig S1A), we averaged the three resulting $p(\widehat{High})$ values.

We developed a general framework to capture potential strategies that a participant could use when categorizing the three-tone sequences (41,50–52). 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 5A). The full Bayesian model reflects the optimal decision-making strategy for a participant (Fig 5A table, top). 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 alternative models, we considered two reduced (suboptimal) variants of this model. In the no-distractor model, the participant assumed that all three tones were 'signal' (Fig 5A table, center). 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 (Fig 5A table, bottom). This model controls for the possibility that participants ignored all of the tones when they made their decision. Henceforth, we interpret our experimental findings using this three-model Bayesian framework.

A



full Bayesian model	Initial belief of trial categories	Perceived as 'signal low'	Perceived as 'distractor'	Perceived as 'signal low'
no-distractor model	Initial belief of trial categories	Perceived as 'signal low'	Perceived as 'signal high'	Perceived as 'signal low'
random-guess model	Initial belief of trial categories	Perceived as 'distractor'	Perceived as 'distractor'	Perceived as 'distractor'



Example participant 1:
High accuracy in trials with 1 or 2 distractors

Example participant 2:
High accuracy in trials with 1 distractor;
Low accuracy in trials with 2 distractors

Example participant 3:
Low accuracy in trials with 1 or 2 distractors

Average behavior across participants

Fig. 5: Schematic diagram of the 3 decision-making models; psychometric curves and model fits shown for both representative participants as well as averaged across all participants.
(A) A schematic of all 3 decision-making models. (top) The full Bayesian model: 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. (center) The no-distractor model: the participant considers all tones to be 'signal' tones. (bottom) The random-guess model: the participant considers all tones to be 'distractor' tones and thus, makes their decision randomly. $\hat{L}$: $\widehat{Low}$-, $\hat{H}$: $\widehat{High}$ - category decision. (B, C, D) Psychometric curves (black lines) and respective model fits (full Bayesian model: orange lines, no-distractor model: teal lines, random-guess model: mediumorchid lines) for 3 example participants (for posteriors of the full Bayesian model, and for GLM analysis, see Fig S2; for more participants see Fig S3). 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.

Population-wide effect of stimulus relevance demonstrated using the Bayesian approach

We first examined the performance of a representative participant with high accuracy on trials with distractors (Fig 5B, one-distractor accuracy 80.59% and two-distractor accuracy 74.14%). Tones that were far from the two category distributions and thus more likely to be distractors were correlated less strongly with behavioral performance than tones closer to the middle of the distributions (Fig 5B, black). Additionally, the optimal full Bayesian model (Fig 5B, orange) fit the participant's nonmonotonic behavior considerably better than the no-distractor or random-guess models (Fig 5B, teal and mediumorchid, respectively; Table S1 – two-sided Wilcoxon signed-rank test). We also designed a simpler, more descriptive model – the Generalized Linear Model and found it to be consistent with the full Bayesian model (Supplementary Information; Figs S2A, E). These analyses collectively demonstrated that this participant used the relevance of the different tones when categorizing trials.

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 5C, D). Furthermore, the difference in accuracy on trials with one and two distractors was lower for the low-accuracy than high-accuracy participants (Fig 5C: one-distractor accuracy 78.24% and two-distractor accuracy 59.48%, Fig 5D: one-distractor accuracy 59% and two-distractor accuracy 61.21%). This observation was substantiated by analysis from the corresponding GLMs (Figs S2F, G). 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 (orange curve versus the teal and mediumorchid curves; Table S1 – two-sided Wilcoxon signed-rank test).

These three examples suggest that potentially all the participants, irrespective of their accuracy, take into account stimulus relevance when choosing the stimulus category, albeit to varying degrees. To test this hypothesis across the entire population, we plotted the average psychometric curve and the average model fits (Fig 5E). We found that there the data were well fit by the full Bayesian model, which qualitatively illustrates a population-wide effect of stimulus relevance on category choice decisions.

Next, we quantified this effect by comparing the statistics for the fits of the Bayesian models. 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 Fig S4)), we used Bayesian Information Criterion (BIC) scores to compare them. A lower BIC score indicates a better model fit. For all participants, the full Bayesian model was better than the random-guess model (Fig 6A; population statistics: two-sided Wilcoxon signed-rank test on BIC, $Z = -6.51$, $p = 7.55e-11$) at describing categorization behavior. This implies that participants do not simply rely on their prior beliefs when classifying the tone sequences.

For most participants, except for two, the full Bayesian model was also better than the no-distractor model (Fig 6B; same test, Z = -6.49, p = 8.41e-11). For the two outlier participants (Fig 6B, C; light pink), the full Bayesian and no-distractor model fits were similar (Table S1 – two-sided Wilcoxon signed-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 5B-E, Figs S3A, D, G). These results confirm that nearly all participants took stimulus relevance into account when making category decisions.

Finally, we correlated model fits to participants' task accuracy. Across the population, the difference in BICs of the full Bayesian and the no-distractor models was inversely correlated with accuracy (Fig 6C, 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.

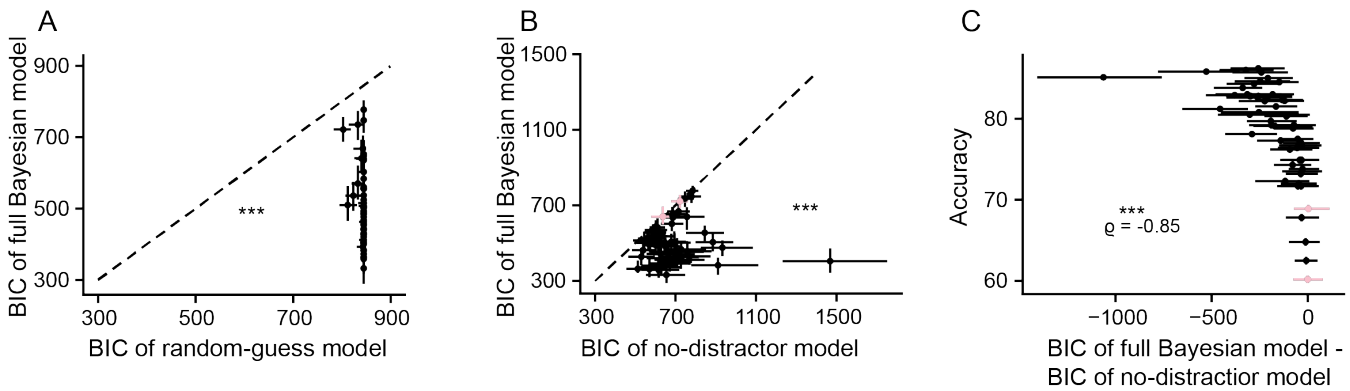


Fig. 6: Nearly all participants were sensitive to stimulus relevance.
On *unbiased* trials, BIC scores of the full Bayesian model were lower than that of the (A) random-guess model for all participants and (B) the no-distractor model for all except two participants. The two outliers are shown in light pink. Statistics: two-tailed Wilcoxon signed-rank test comparing model fits across the population. (C) Participants' accuracy was inversely correlated with difference in BIC scores between the full Bayesian and no-distractor models. Statistics: Spearman correlation. 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) and horizontal error bars in (C): 95% confidence intervals derived from respective model fits to 100 bootstrapped samples. Vertical error bars in (C): standard error of the mean, obtained by assuming that accuracy follows a binomial distribution; vertical error bars not large enough to be distinguishable.

Sensory uncertainty and estimated task relevance underlie inter-participant variability

However, individual participants' ability to discriminate between signal and distractor tones may depend not only on their subjective estimates of the tones' task relevance (or more specifically, their estimate of the probability of distractors) but also on their measure of stimulus uncertainty (which is the standard deviation of the two Gaussian distributions) and their sensory uncertainty. For instance, some participants may have different prior convictions about the number of distractors (e.g., some may spend more time in noisy environments than others), which could bias how they integrate information over the three tones. Additionally, we expect that participants with low stimulus and sensory uncertainty would distinguish better between signal and distractor tone frequencies than participants with high uncertainty. We thus conjectured that estimated priors for the probability of distractors, stimulus uncertainty and, sensory uncertainty would drive individual participants' category choice and task accuracy contributing to inter-individual variability.

To test the effect of these three factors on category choice, we first analyzed the inter-participant variability in these factors. We found that across the 56 participants, stimulus uncertainty (denoted by model parameter σ) was nearly constant (Fig S4D), whereas sensory uncertainty (denoted by parameter $\sigma_{sensory}$) and the estimation for the probability of distractors (denoted by model parameter $p_{distractor}$) were participant-specific

(Figs S4A,E,F). To quantify a relationship between $p_{distractor}$ and $\sigma_{sensory}$, we defined a metric 'sigmoidicity' (Fig 7A; see Methods) which quantitatively captured participant's category decisions across all trials. We found that this metric was significantly influenced by $p_{distractor}$ and $\sigma_{sensory}$ as well as by the interaction between these two parameters (Fig 7B; Table S1 – two-factor regression model). Specifically, sigmoidicity increased with $\sigma_{sensory}$ and decreased with $p_{distractor}$. We then 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 7C; Table S1 – two-factor regression 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 had a similar measure of stimulus uncertainty but differed in their sensory uncertainty and measure of relevance uncertainty. Further, although participants' category decisions were shaped by both their estimate for the probability of distractors and their sensory uncertainty, much of their differences in accuracy may be driven by differences in their ability to distinguish between the tone frequencies.

Next, we tested whether listeners differentially weighed tones based on their position in the sequence when making their category decision. To test this idea, we developed an alternative Bayesian model with a nonuniform psychophysical kernel in which we substituted $p_{distractor}$ for 3 tone-position dependent versions of $p_{distractor}$. We found that participants weighed the first two tone positions ($p_{distractor_{position1}}$: Mdn = 0.7, IQR = 0.24; $p_{distractor_{position2}}$: Mdn = 0.7, IQR = 0.28) equally but slightly more than the third tone position ($p_{distractor_{position3}}$: Mdn = 0.85, IQR = 0.17; Fig S5G; Table S1 – one-way rmANOVA, post hoc two-sided Wilcoxon signed-rank test) when making their category decision. Additionally, values for $p_{distractor_{position3}}$ but not $p_{distractor_{position1}}$ and $p_{distractor_{position2}}$ were statistically different than $p_{distractor}$ (Fig S5G; Mdn = 0.59, IQR = 0.43; Table S1 – two-sided Wilcoxon signed-rank test). The influence of all three position-dependent weights on sigmoidicity and accuracy was consistent with results (Figs. S5A-F; Table S1 – Spearman) from the full Bayesian model (Figs. 7B, C). These findings are also supported by those from the simpler GLM model (Fig. S5H; Table S1 – one-way rmANOVA, post hoc two-sided Wilcoxon signed-rank test). Overall, these data show that participants' used information from all three tone positions but weighed the third tone somewhat less than the first two.

Interestingly, for both the full Bayesian model and the tone-position dependent Bayesian model we found the $p_{distractor}$ values to be higher than the true experimental value of 0.3. Participants thus had a tendency to ignore tones more than the statistics of the task would require.

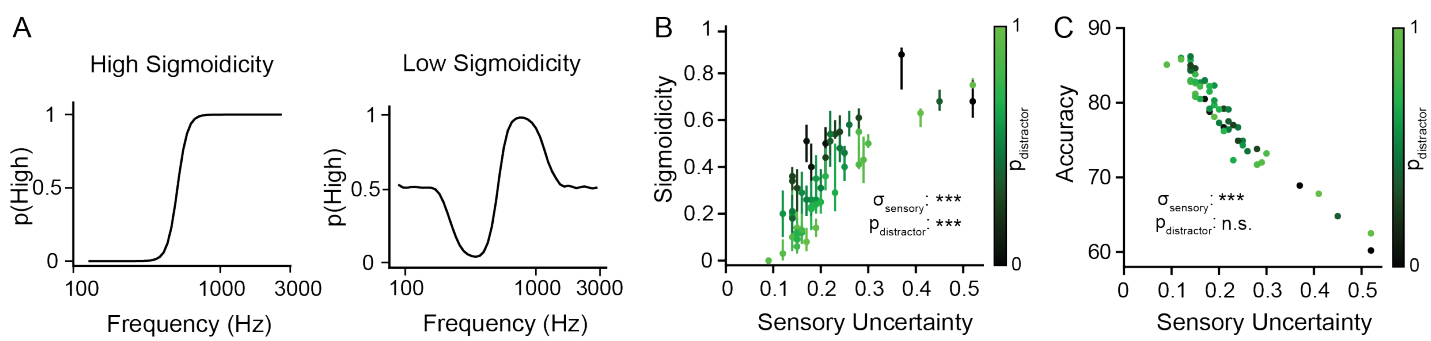


Fig. 7: For an individual participant their category choice was driven by both their sensory uncertainty and their subjective estimates of the probability of distractors, but their accuracy was mostly driven by their sensory uncertainty.
(A) Cartoon models of different psychometric curves illustrate the qualitative range of sigmoidicity. (B) Each data point is a participant; its hue is given by $p_{distractor}$ which is the participant's estimated probability of distractors (see color bar); Black: $p_{distractor} \sim 0$; Green: $p_{distractor} \sim 1$. $p_{distractor}$ is a fitting parameter in the full Bayesian model and has value 0 in the no-distractor model. The model parameter $\sigma_{sensory}$ captures participant's sensory uncertainty. Statistics: effect of $p_{distractor}$ and $\sigma_{sensory}$ on sigmoidicity using a two-factor regression model. (C) Participants' accuracy did not significantly depend on $p_{distractor}$

but was inversely correlated with their sensory uncertainty. $p_{distractor}$ is color coded as in (B). Statistics: effect of $p_{distractor}$ and $\sigma_{sensory}$ on accuracy using a two-factor regression model. In all panels, data point: one participant; †p<0.05; *p<0.01, **p<0.001, ***p<0.0001. Error bars in (B) and horizontal error bars in (C): 95% confidence intervals derived from respective model fits to 100 bootstrapped samples. Vertical error bars in (C): standard error of the mean, obtained by assuming that accuracy follows a binomial distribution; vertical error bars aren't large enough to be distinguishable.

Participants rely on past statistics of category probabilities

Having established how stimulus uncertainty, sensory uncertainty and stimulus relevance shape auditory categorization, we next asked whether and how participants' learning of 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 category of the previous trial ($t-1$), $\mathrm{p}(\widehat{\mathrm{H}_t}|\mathrm{H}_{t-1})$ and $\mathrm{p}\left(\widehat{\mathrm{H}_t}\middle|\mathrm{L}_{t-1}\right)$ (Fig 8A). 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 8C, D), others' decisions were not (Fig 8B). Overall, across the population, the category of the previous trial (i.e., short-term stimulus statistics) significantly influenced participants' choices on the current trial (Fig 8E; $\Delta_{\text{short-term}}$: difference between $\mathrm{p}(\widehat{\mathrm{H}_t}|\mathrm{H}_{t-1})$ and $\mathrm{p}\left(\widehat{\mathrm{H}_t}\middle|\mathrm{L}_{t-1}\right)$ at central frequency of 511Hz; central frequency: frequency which is the mid-point of the experimental frequency range in log-space; two-sided Wilcoxon signed-rank test on $\Delta_{\text{short-term}}$, Z = 6.04, p = 1.58e-9). Moreover, we found that this influence was only restricted to the category of the previous trial and did not substantially depend on whether the participant responded accurately to the trial (Fig S6; Table S1 – two-sided Wilcoxon signed-rank test on $\Delta_{\text{CorrectVsIncorrect}}$). Combined, these data confirm the presence of category-driven short-term learning in decision-making.

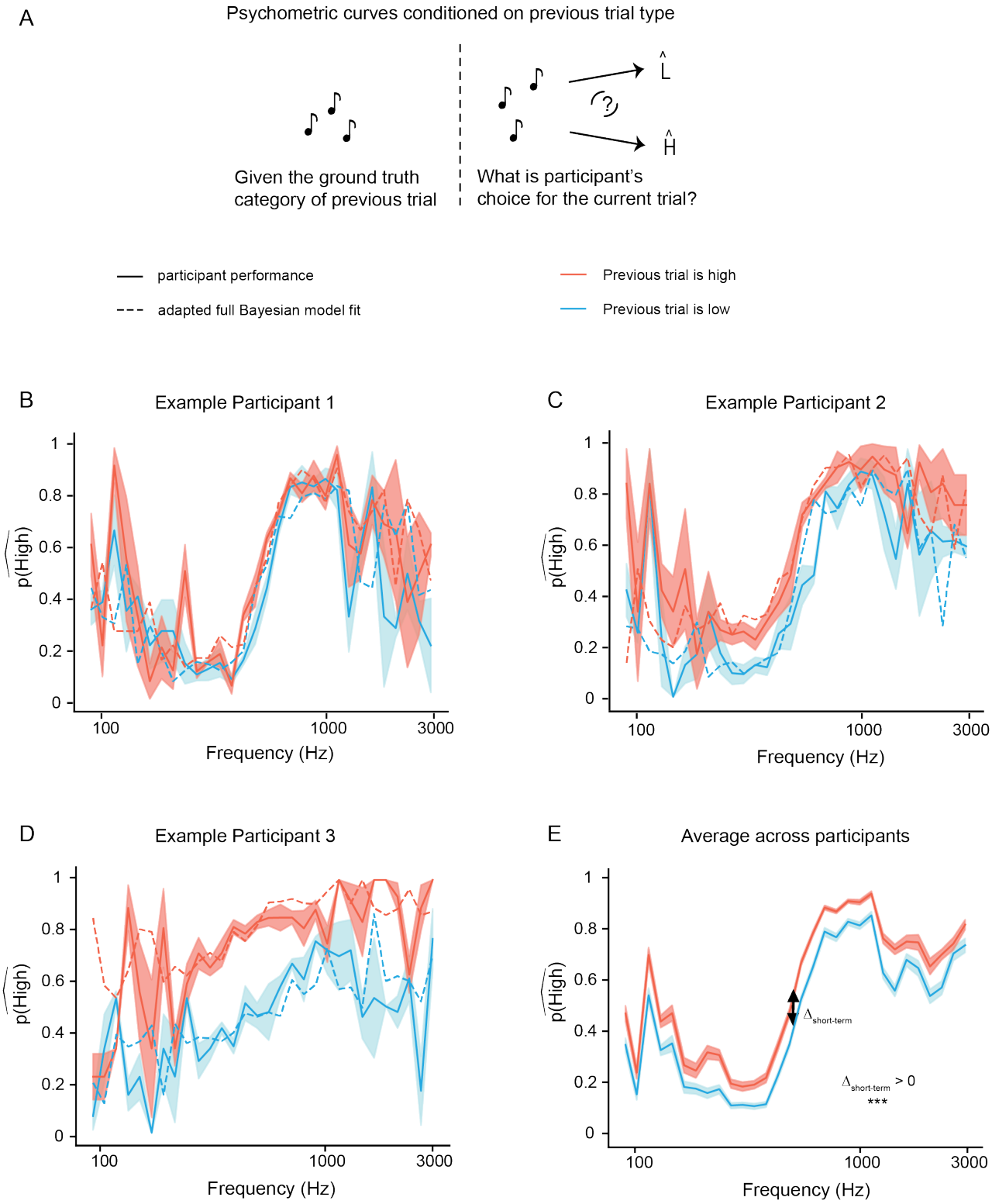


Fig. 8: To a varying degree, participants incorporated short-term statistics of category probabilities in their decisions. (A) Schematic shows the effect of learning short-term statistics. (B-D) Average psychometric curves (light blue: previous trial is low; red: previous trial is high) and the adapted full Bayesian model fits (same colors, dashed lines) conditioned on the previous trial's category 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 ($\Delta_{short-term}$): (B) ~0.1, (C) ~0.22 and (D) ~0.32. (E) Corresponding psychometric

curves averaged across all participants in the *unbiased session*; $\Delta_{short-term}$: ~0.12. Statistics: two-tailed Wilcoxon signed-rank test on $\Delta_{short-term}$. In panel (E), $^{\dagger}p<0.05$; $^{*}p<0.01$, $^{**}p<0.001$, $^{***}p<0.0001$.

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 short- and/or long-term learning, in separate behavioral sessions, we changed the probability of the low-category trials. Specifically, $p_L = 0.7$ in *biased low* sessions and $p_L = 0.3$ in *biased high* sessions (Fig 9A; Participants: N = 53 in *biased low* and N = 48 jointly in *biased low* and *biased high*). Anecdotally, individual participants with high values of $p_{distractor}$ did not use either type of learning when making decisions (Fig 9B; Fig S7A). However, participants with lower values of $p_{distractor}$ seemed to use previous trial statistics to varying degrees (Figs 9C, D). Some participants used both short- and long-term learning (Fig 9C; Fig S7B), whereas others primarily used long-term learning (Fig 9D; Fig S7C). Overall, in the population, there was a clear effect of both the session type and the last trial on category choice (Fig 9E, Figs S6-8; $\Delta_{\text{long-term}}$: 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, data across the three sessions – *unbiased, biased low* and *biased high* – suggests that both, short- and long-term learning influenced participants' decision-making, presumably depending on their relevance estimates.

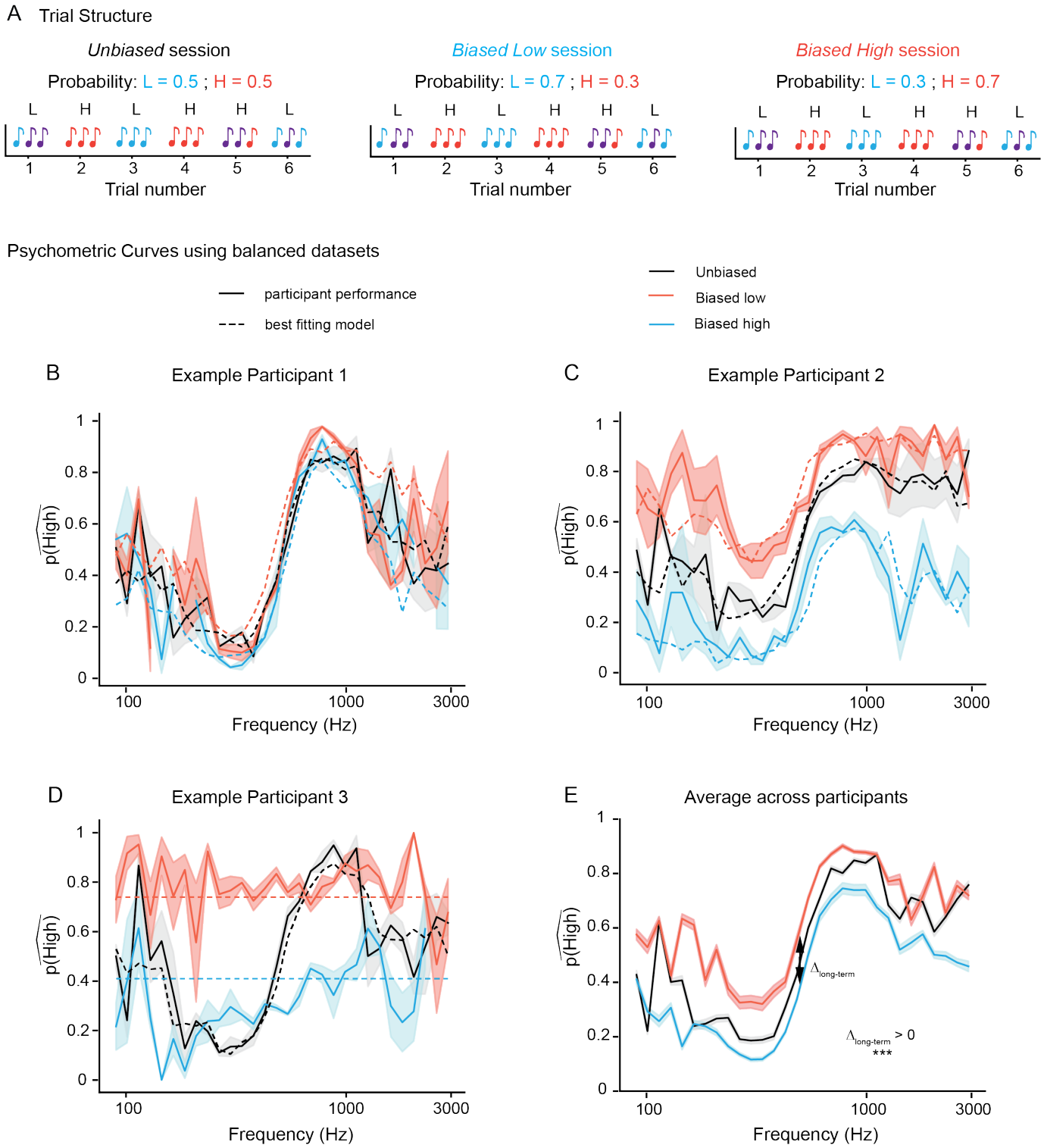


Fig. 9: To a varying degree, participants' decisions were also affected by long-term statistics of category probabilities. (A) Example stimuli from *unbiased*, *biased low*, and *biased high* trials. Individual tones are color-coded according to their underlying distribution as shown in Fig. 1B. (B-D) Average psychometric curves for three participants in the *unbiased* (black), *biased low* (light blue) and *biased high* (red) sessions. Respective model fits also shown using dashed lines; (B, C) fit using the full Bayesian model and (D) 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 5B-D. Difference between $p(\widehat{High})$ for *biased low* and *biased high* trials computed at the central tone frequency 511 Hz ($\Delta_{long-term}$): (B) ~ 0.06, (C) ~ 0.35 and (D) ~

0.51. (E) Psychometric curves averaged across all participants; $\Delta_{long-term}$: ~0.2. Statistics: two-tailed Wilcoxon signed-rank test on $\Delta_{long-term}$. In panel (E), $^{\dagger}p<0.05$; $^{*}p<0.01$, $^{**}p<0.001$, $^{***}p<0.0001$.

To further quantify how learning modulates category decision-making, we adapted the full Bayesian model to incorporate both short- (exponential weighted: $W_1 e^{-t/\tau}$) and long-term (constant: $W_{constant}$) components (Values of W_1 and $W_{constant} \sim 0$ implies that there was not any learning, whereas values ~1 implies learning). This adapted model captured the combined effects of sensory uncertainty, stimulus relevance, stimulus uncertainty, as well as learning of long- and short-term stimulus statistics.

Using this framework, we asked how participants measure of stimulus relevance and their sensory uncertainty modulated their learning of stimulus statistics. Thus, we first tested whether participants continued to account for stimulus relevance when categorizing trials during *biased low* and *high* sessions. We found that the performance of most participants, except for two, was best fit by the full Bayesian model (Fig S9). These two outlier participants primarily used their learned prior and their performance was thus best fit by the random-guess model (Figs S9A-D). These data reinforce our previous finding that stimulus relevance is central to categorization.

Next, we assessed the degree to which participants learnt the short-term statistics of trials by 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 10A; $W_1 > 0$ for N = 55). However, there was substantial variability across participants such that for some $W_1 \sim 0$ whereas for some others it was ~ 1. Similar heterogeneity was also observed in the time constant for learning such that the participants' time constant was inversely correlated with their accuracy (Fig 10B; Spearman, ϱ = -0.27, p = 0.042). However, across the population, the average time constant (τ) in the *unbiased* session was ~1 trial (Mdn: 0.63 trial, IQR: 0.44; Figs 10A, B). The nature of W_1 (Fig 10C) and τ (Fig S10; Table S1 – Spearman correlation with accuracy) was similar in the *biased* sessions. Jointly, these results indicate that although most participants use short-term learning when making category decisions, this learning was largely restricted to the immediately preceding trial.

Similar to short-term learning, we also expected some participants to exhibit effects of long-term learning. Indeed, most participants learned the long-term trial statistics (Fig 10C), but this learning was not correlated to their short-term learning (Spearman; Table S1). Thus, participants appear to use both the bias in the session and their knowledge of previous trials albeit to different degrees when making decisions.

Next, we probed how these two learning mechanisms relate to participants' estimated probability of distractors. We found that the weight of the preceding trial ($W_1 e^{-1/\tau}$) was inversely correlated with $p_{distractor}$ for all three sessions (Figs 10D, E; 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 10F; Spearman, *biased low and high*: ϱ = -0.25, p = 0.013). This indicates that participants' measure of relevance uncertainty was inversely correlated with their stimulus-category knowledge derived over both short- and long-time scales. This suggests that participants who are poor at differentiating between signal versus distractor tones likely rely on other sources of information such as past knowledge when making category choices.

Finally, we asked how the various parameters that drive performance – sensory uncertainty, stimulus relevance, stimulus uncertainty, and learning relate to task accuracy. We found that sensory uncertainty was a strong predictor of accuracy, accounting for much of the variance (Figs 7C, S11; Spearman, *biased low and high*: ϱ = -0.75, p = 2.38e-19). Thus, 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 the basic assumption of Bayesian models that sensory uncertainty is a fundamental factor in decision making.

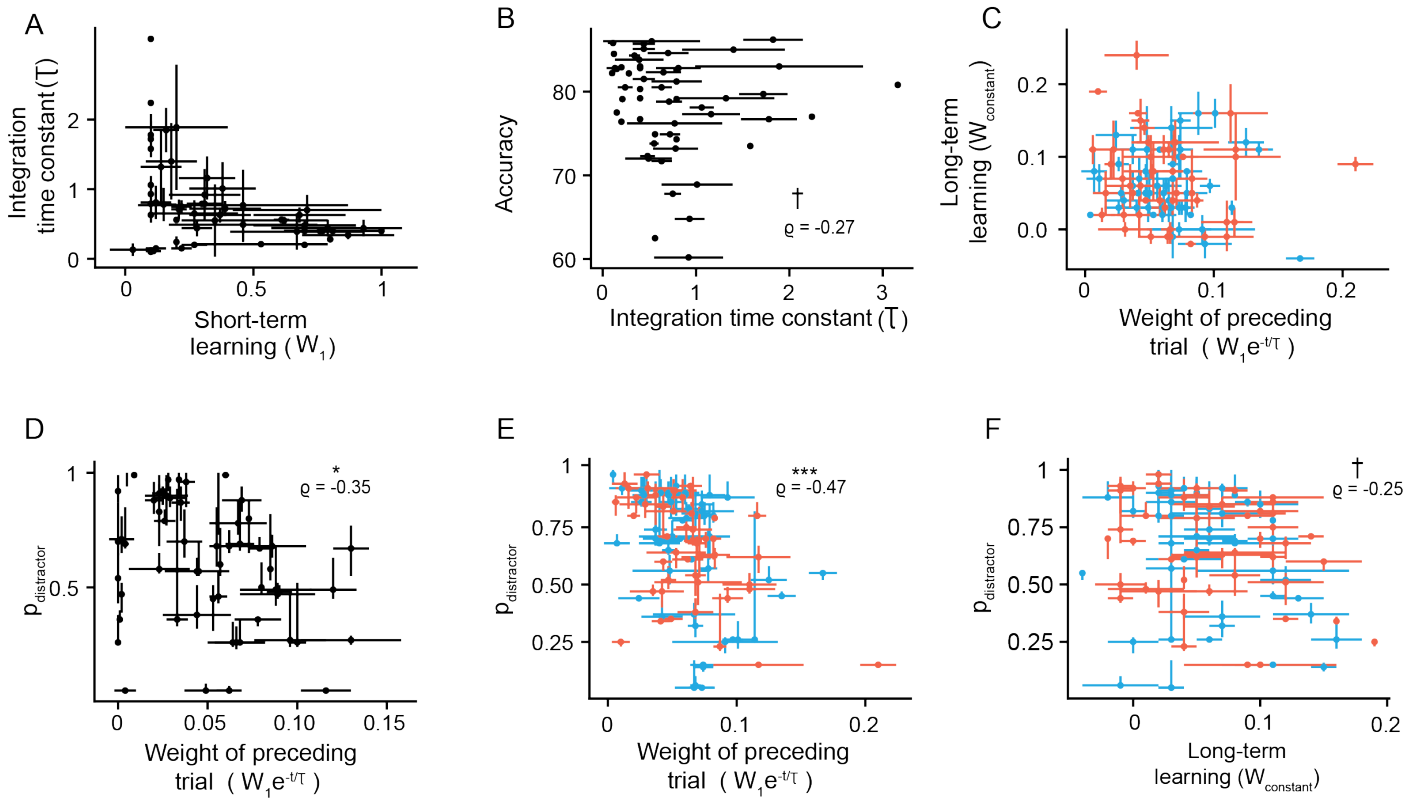


Fig. 10: Participants subjective estimate of distractor probability varies inversely with their learning of stimulus statistics when making category decision.

(A) On *unbiased* trials, nearly all participants (black, N=55) exhibited short-term learning ($W_1 > 0$). (B) Their accuracy was inversely correlated with the time constant (τ) for such learning. Further, the median number of previous trials over which participants' retained information was 0.63; effectively participants retained information over the immediately preceding trial. (C) On *biased* trials (*biased low*: light blue, N=53; *biased high*: red, N=48), the effect of short- and long-term learning varied substantially across participants. Across all three sessions, $p_{distractor}$ was inversely correlated with (D, E) short-term learning given by $W_1 e^{-1/\tau}$ and in the *biased* sessions, it was also inversely related to (F) long-term learning captured using $W_{constant}$. All panels, data point: one participant; $^{\dagger}p<0.05$; $^{*}p<0.01$, $^{**}p<0.001$, $^{***}p<0.0001$; Statistics: Spearman correlation; error bars: standard error of the mean.

Discussion

It has long been known that humans and other animals can isolate stimuli of interest from background in noisy, crowded environments. When presented with two competing auditory streams, participants' neural responses to the task-relevant stimuli were enhanced(53,54), but, their processing of the irrelevant distractor stimuli was mediated by attention(55,56). Similar findings in the visual object-processing literature suggest that spatial and feature attention affect the neural representation of task-irrelevant sensory information(57,58). 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(30,59–61) – 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(62). Indeed, such adaptation underscores context-dependent speech perception and categorization(27). In our work, we investigated this complexity in perceptual decision-making by testing how participants' sensory uncertainty, measure of stimulus relevance, stimulus uncertainty, and learning interacted to shape their auditory categorization behavior.

We found that participants differentially combined these factors when making category decisions. Specifically, when categorizing tone sequences consisting of both category-irrelevant (distractors) and relevant tones, participants weighed individual tones by their task relevance (Figs 2,3; Figs S1-3,9). Additionally, nearly all participants had similar stimulus uncertainty but diverged in their sensory uncertainty and their estimates of stimulus relevance (Fig S4). In fact, participants' category choices were jointly determined by both their sensory uncertainty and their subjective estimate of stimulus relevance, whereas their accuracy was mostly driven by their sensory uncertainty (Fig 7). 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 8-10; Fig S6-9).

Flexibility of experimental paradigm

Compared to previous work(30,63), our experimental paradigm is unique in that we could modulate the multiple factors of sensory uncertainty, stimulus relevance, stimulus uncertainty and learning all within the same stimulus sequence. For example, because our stimulus was a three-tone sequence and because we could manipulate the number of task-irrelevant distractors (Fig 1), we could 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 2,5-7; Fig S1). Additionally, although humans use both short- and long-term knowledge to efficiently make decisions(26,33), most studies probing the influence of learned expectations on categorization have largely examined either short- or long-term effects without testing for their interaction(64–66). 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 10). Thus, our experimental setup allowed us to simultaneously probe multiple factors that often underlie category decision-making.

Bayesian approach to study variability and interplay of relevance and uncertainty

To quantify the interplay between all these factors, we leveraged a Bayesian model (Figs 1,5-10). Bayesian models are a powerful tool that has been successfully used to study human behavior in sensorimotor learning(12,67), sensory perception(25,68), categorization(30,42) etc. Our Bayesian analysis allowed us to characterize 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 6,10; Figs S3,9). 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 (Fig S9C, D). In addition to this heterogeneity across strategies, we found that even among the participants using the optimal strategy there was substantial variability (Figs 7,10; Figs S2-5,7) resulting from participants' sensory uncertainty, their subjective measure of relevance, and their degree of short- and long-term learning. For instance, by modeling sensory uncertainty using the parameter $\sigma_{sensory}$, we found that although, on average, the value of participants' sensory uncertainty is consistent with previous literature(30), some participants have low sensory uncertainty and high task accuracy whereas some others have substantially high sensory uncertainty and poor task accuracy (Fig 7). This variability also underscored participants' measure of stimulus relevance and their learning of stimulus statistics. 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. Future work could use similar Bayesian approaches to probe whether in such categorization tasks, participants use the same strategy throughout the experiment or switch between different strategies(69) based on the complexity of the trial.

The Bayesian analysis also allowed us to explore how the different factors – sensory uncertainty, stimulus relevance, stimulus uncertainty and learning – interplay during categorization. We found that participants' sensory uncertainty and their estimated likelihood of the probability of distractors taken together determine the nature of their psychometric curves. Additionally, though the effects of the two time scales of learning were independent (Fig 10) of each other(60), both were inversely correlated with the relevance parameter $p_{distractor}$. This suggests that when making decisions, participants combined their uncertainty about stimulus relevance with other sources of information, a finding that supports similar observations from previous work(70). A shortcoming of our experimental design is that we do not vary stimulus uncertainty. Future studies could examine how this parameter interacts with the other forms of uncertainty when making category decisions.

Future directions

The interplay between multiple factors might also have implications for the broader understanding of how humans navigate real-life crowded noisy environments. For example, spatial cues(71,72) and attention(73–75) are integral to parsing multiple sound sources, 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 greatly affected by tone positions(Fig S5). However, real-life stimuli such as speech(76) 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' estimates of stimulus relevance varied inversely with their learning of stimulus statistics over different time scales. Future work should explore the role of relevance in real-life scenarios and whether prior knowledge of one's environment may 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(77–81), uncertainty in relevance of stimuli(57), uncertainty in category of trials(82) and sensory uncertainty may all be captured by distinct circuit elements and brain regions(83,84). 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 in writing 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(85) 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 (each tone is a sinusoid; 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.

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

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

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

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 10 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. All participants self-reported that they had normal hearing or hearing aids.

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.

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(30,41,86–88).

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. 2B: Before we analyzed 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 correlation of each of the three tone frequencies with category choice probability $\mathrm{p}(\widehat{\mathrm{High}})$ for all trials in the *unbiased* session (correlations are normalized to 1) and found that these correlations were independent of tone positions (Fig S1A; Table S1, one-way repeated measures ANOVA (rmANOVA)). This analysis suggests that participants weighed all three tone positions equally within a trial and integrated information across the tones when making their category choice, and therefore our analyses could treat tone types in all three tone positions the same.

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 ($\widehat{\mathrm{High}}$ ($\hat{\mathrm{H}}$) or $\widehat{\mathrm{Low}}$ ($\hat{\mathrm{L}}$)). 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(89,90), 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 almost equally correlated with behavior (Fig S1A, S5), the relevance of a tone (denoted by parameter $p_{distractor}$) 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,

$$\sum_{r_{p_1}=R,I}\sum_{r_{p_2}=R,I}\sum_{r_{p_3}=R,I} p\left(\log\left(t_{H_1}\right)\middle|H,r_{p_1}\right)p\left(\log\left(t_{H_2}\right)\middle|H,r_{p_2}\right)p\left(\log\left(t_{H_3}\right)\middle|H,r_{p_3}\right)p\left(r_{p_1},r_{p_2},r_{p_3}\middle|H\right)$$
$$=\prod_{i=1,2,3} p\left(\log\left(t_{H_i}\right)\middle|H,r_{p_i}=R\right)p(R|H)+p\left(\log\left(t_{H_i}\right)\middle|H,r_{p_i}=I\right)p(I|H) \quad (11)$$
$$=\prod_{i=1,2,3} p_{GH}\left(\log\left(t_{H_i}\right)\middle|\mu_{high},\sigma\right)*(1-p_{distractor})+p_U(\log(t_{H_i}))*p_{distractor}. \quad (12)$$

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 main 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. In these three models, stimulus relevance is agnostic of tone position (Fig 5). We also compared results from the full Bayesian model to those from an alternative model with position-dependent weights for each tone in the sequence (Fig S5).

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. 10), 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.

Last, the model with position-dependent weights has 8 fitting parameters – $\mu_{low}, \mu_{high}, \sigma, \sigma_{sensory}, p_{distractor_{position_1}}, p_{distractor_{position_2}}, p_{distractor_{position_3}}$ and p_{low}. $p_{distractor_{position_1}}, p_{distractor_{position_2}}$ and $p_{distractor_{position_3}}$ are participants' estimates of the probabilities that the first, second and third tone respectively in the sequence are distractors.

To fit each of the above models to the participants' psychometric curves (see Figs 6,8,9; Figs S3,4), we applied a decision threshold to the Bayesian model posterior such that,

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

Similar to Eq. 9, we considered that each tone t_i generated sensory evidence t_{p_i} according to the probability density $p\big(\log(t_{p_i})\,\big|\,\log(t_i)\big)$, 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 + second-order Tikhonov) regularized binary logistic regression model(91) for each participant such that the model predictors are the indicator functions for the three tones in a trial. An indicator function $\mathbb{I}_{t_i}$ 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,

$$y = \begin{cases} 1 & \text{expit}(w_0 + w_1\mathbb{I}_1 \dots w_{30}\mathbb{I}_{30}) > 0.5 \\ 0 & \text{otherwise.} \end{cases} \quad (16)$$

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

To compute the individual influence of the three tone positions in the sequence (Fig S5H), we expand the model to have 90 predictors – 30 for each tone position.

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 Figs S4G, H) 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. These values are consistent with previous findings(30).

Next, like the *unbiased* session, we also fit data from the two *biased* sessions using the second fitting approach. We tested our fitting procedure by simulating 4 virtual participants (Table S2). These were modeled based on 4 actual participants from our dataset to ensure a consistency check for performance accuracy. Additionally, for the virtual participants, we computed the confidence intervals on their fitted parameters by simulating behavior 10 times and following the bootstrapping procedure as detailed below.

Specifically, we generated confidence intervals by creating either 100 bootstrapped (with replacement) datasets comprising of 600 trials each for the *unbiased* session (Fig 6, Fig S4) or 100 balanced datasets for the *biased* tasks (Fig 10, Fig S9). 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 9A has negligible internalized expectation of bias, whereas participants in Figs S9C, 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 $\{0, 0.1, 0.2, \ldots, 1\}$.

Full Bayesian model with position-dependent weights. We reduce the 8 parameters in this model to 4. Participants are assumed to be veridical about the Gaussian parameters and for $\sigma_{sensory}$ we use the corresponding median value. Thus, we only fit $p_{distractor_{position_1}}$, $p_{distractor_{position_2}} p_{distractor_{position_3}}$ and p_{low}. The fitting procedure is similar to the Full Bayesian model.

Metrics. Across the three experimental sessions, for nearly all the participants, we used the final model parameter values of the Full Bayesian model to compute their psychometric-curve fits and their Bayesian model posteriors (Figs 5,7-9; Figs S2,3,9). Because we had our trials had 3 tones, we plotted the average psychometric curve across all 3 tone positions.

Average curve = $(p(High|tone_{position_1}) + p(High|tone_{position_2}) + p(High|tone_{position_3}))/3$.

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 6; Fig S9). We

also tested the results using the Akaike Information Criterion and the corrected Akaike Information Criterion(92), 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 shape of the psychometric curve and thus, in turn the category choice of the participant. It is ~ 0 for participants who can accurately identify distractor tones as irrelevant to their category decisions (example subject in Fig 5B) and is ~ 1 for participants whose psychometric curve resembles a step function. Specifically, for each participant sigmoidicity is computed using their respective posterior curve (examples in Figs S2B-D, S3B, E, H) 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 second-order Tikhonov regularization. $\alpha = 0$ implies only Lasso regularization while $\alpha = 1$ implies only Tikohonov regularization. For example, α for the curve fit in Fig S2E is 0.3 whereas it is 0 for those in Figs S2F and S2G. The Tikhonov matrix (τ) is designed such that $\tau_{ii} = -1, \tau_{i,i-1} = 2$ and $\tau_{i,i+1} = -1$.

The individual influence of the three tone positions in the sequence (Fig S5H) is computed by taking the sum of the absolute values of the corresponding GLM weights for each of the 30 tone frequencies. The weights for the first tone position are $w_1 \ldots w_{30}$, for the second tone position are $w_{31} \ldots w_{60}$ and for the third tone position are $w_{61} \ldots w_{90}$. We tested our fitting procedure by simulating 6 virtual participants (Table S3). The virtual participants were based on real experimental participants.

Data and Code Availability

The de-identified data and code associated with this paper are publicly available and can be accessed at https://github.com/JanakiSheth/Relevance-uncertainty-and-learning-influence-auditory-categorization. We have also used Zenodo to assign a DOI to the repository: 10.5281/zenodo.7439086.

1. Heekeren HR, Marrett S, Ungerleider LG. The neural systems that mediate human perceptual decision making. Nat Rev Neurosci. 2008 Jun;9(6):467–79.

2. Chen SY, Ross BH, Murphy GL. Decision making under uncertain categorization. Front Psychol [Internet]. 2014 [cited 2022 Jan 21];5. Available from: https://www.frontiersin.org/article/10.3389/fpsyg.2014.00991

3. Niwa M, Ditterich J. Perceptual Decisions between Multiple Directions of Visual Motion. J Neurosci. 2008 Apr 23;28(17):4435–45.

4. Bushdid C, Magnasco MO, Vosshall LB, Keller A. Humans Can Discriminate More than 1 Trillion Olfactory Stimuli. Science [Internet]. 2014 Mar 21 [cited 2022 Jan 21]; Available from: https://www.science.org/doi/abs/10.1126/science.1249168

5. Garcia SE, Jones PR, Rubin GS, Nardini M. Auditory Localisation Biases Increase with Sensory Uncertainty. Sci Rep. 2017 Jan 11;7(1):40567.

6. Heron J, Whitaker D, McGraw PV. Sensory uncertainty governs the extent of audio-visual interaction. Vision Res. 2004 Nov 1;44(25):2875–84.

7. Pouget A, Beck JM, Ma WJ, Latham PE. Probabilistic brains: knowns and unknowns. Nat Neurosci. 2013 Sep;16(9):1170–8.

8. Qamar AT, Cotton RJ, George RG, Beck JM, Prezhdo E, Laudano A, et al. Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. Proc Natl Acad Sci. 2013 Dec 10;110(50):20332–7.

9. Beierholm U, Rohe T, Ferrari A, Stegle O, Noppeney U. Using the past to estimate sensory uncertainty. Reichenbach T, King AJ, Reichenbach T, Acerbi L, editors. eLife. 2020 Dec 15;9:e54172.

10. Barthelmé S, Mamassian P. Evaluation of Objective Uncertainty in the Visual System. PLOS Comput Biol. 2009 Sep 11;5(9):e1000504.

11. Zhou Y, Acerbi L, Ma WJ. The role of sensory uncertainty in simple contour integration. PLOS Comput Biol. 2020 Nov 30;16(11):e1006308.

12. Körding KP, Wolpert DM. Bayesian integration in sensorimotor learning. Nature. 2004 Jan;427(6971):244–7.

13. Wei K, Körding K. Relevance of Error: What Drives Motor Adaptation? J Neurophysiol. 2009 Feb 1;101(2):655–64.

14. Daliri A, Dittman J. Successful auditory motor adaptation requires task-relevant auditory errors. J Neurophysiol. 2019 Aug 1;122(2):552–62.

15. Anders UM, McLean CS, Ouyang B, Ditterich J. Perceptual Decisions in the Presence of Relevant and Irrelevant Sensory Evidence. Front Neurosci [Internet]. 2017 [cited 2022 Jan 19];11. Available from: https://www.frontiersin.org/article/10.3389/fnins.2017.00618

16. Mirza MB, Adams RA, Friston K, Parr T. Introducing a Bayesian model of selective attention based on active inference. Sci Rep. 2019 Sep 26;9(1):13915.

17. Lutfi RA, Kistler DJ, Callahan MR, Wightman FL. Psychometric functions for informational masking. J Acoust Soc Am. 2003 Dec;114(6 Pt 1):3273–82.

18. Oh EL, Wightman F, Lutfi RA. Children's detection of pure-tone signals with random multitone maskers. J Acoust Soc Am. 2001 Jun;109(6):2888–95.

19. Lutfi RA. Informational processing of complex sound. III: Interference. J Acoust Soc Am. 1992 Jun;91(6):3391–401.

20. Hansen K, Hillenbrand S, Ungerleider L. Effects of Prior Knowledge on Decisions Made Under Perceptual vs. Categorical Uncertainty. Front Neurosci [Internet]. 2012 [cited 2022 Jan 21];6. Available from: https://www.frontiersin.org/article/10.3389/fnins.2012.00163

21. Kok P, Brouwer GJ, Gerven MAJ van, Lange FP de. Prior Expectations Bias Sensory Representations in Visual Cortex. J Neurosci. 2013 Oct 9;33(41):16275–84.

22. Rahnev D, Lau H, Lange FP de. Prior Expectation Modulates the Interaction between Sensory and Prefrontal Regions in the Human Brain. J Neurosci. 2011 Jul 20;31(29):10741–8.

23. Kok P, Mostert P, de Lange FP. Prior expectations induce prestimulus sensory templates. Proc Natl Acad Sci. 2017 Sep 26;114(39):10473–8.

24. Rohenkohl G, Cravo AM, Wyart V, Nobre AC. Temporal Expectation Improves the Quality of Sensory Information. J Neurosci. 2012 Jun 13;32(24):8424–8.

25. Stocker AA, Simoncelli EP. Noise characteristics and prior expectations in human visual speed perception. Nat Neurosci. 2006 Apr;9(4):578–85.

26. Mendonça AG, Drugowitsch J, Vicente MI, DeWitt EEJ, Pouget A, Mainen ZF. The impact of learning on perceptual decisions and its implication for speed-accuracy tradeoffs. Nat Commun. 2020 Jun 2;11(1):2757.

27. Holt LL. Speech categorization in context: Joint effects of nonspeech and speech precursors. J Acoust Soc Am. 2006 Jun;119(6):4016–26.

28. Kluender KR, Coady JA, Kiefte M. Sensitivity to change in perception of speech. Speech Commun. 2003 Aug;41(1):59–69.

29. Holt LL, Lotto AJ, Kluender KR. Neighboring spectral content influences vowel identification. J Acoust Soc Am. 2000 Aug;108(2):710–22.

30. Gifford AM, Cohen YE, Stocker AA. Characterizing the Impact of Category Uncertainty on Human Auditory Categorization Behavior. PLOS Comput Biol. 2014 Jul 17;10(7):e1003715.

31. Berniker M, Voss M, Kording K. Learning Priors for Bayesian Computations in the Nervous System. PLOS ONE. 2010 Sep 10;5(9):e12686.

32. Simoncelli EP, Olshausen BA. Natural Image Statistics and Neural Representation. Annu Rev Neurosci. 2001;24(1):1193–216.

33. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. Learning the value of information in an uncertain world. Nat Neurosci. 2007 Sep;10(9):1214–21.

34. Green DM, Swets JA. Signal detection theory and psychophysics. Oxford, England: John Wiley; 1966. xi, 455 p. (Signal detection theory and psychophysics).

35. Watson CS, Kidd GR. Studies of tone sequence perception: effects of uncertainty, familiarity, and selective attention. Front Biosci J Virtual Libr. 2007 May 1;12:3355–66.

36. Russ BE, Lee YS, Cohen YE. Neural and behavioral correlates of auditory categorization. Hear Res. 2007 Jul;229(1–2):204–12.

37. Banno T, Lestang JH, Cohen YE. Computational and neurophysiological principles underlying auditory perceptual decisions. Curr Opin Physiol. 2020 Dec;18:20–4.

38. Ley A, Vroomen J, Hausfeld L, Valente G, Weerd PD, Formisano E. Learning of New Sound Categories Shapes Neural Response Patterns in Human Auditory Cortex. J Neurosci. 2012 Sep 19;32(38):13273–80.

39. Tsunada J, Cohen YE. Neural mechanisms of auditory categorization: from across brain areas to within local microcircuits. Front Neurosci [Internet]. 2014 [cited 2022 Jul 11];8. Available from: https://www.frontiersin.org/articles/10.3389/fnins.2014.00161

40. Cao Y, Summerfield C, Park H, Giordano BL, Kayser C. Causal Inference in the Multisensory Brain. Neuron. 2019 Jun 5;102(5):1076-1087.e8.

41. Acerbi L, Dokka K, Angelaki DE, Ma WJ. Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. PLOS Comput Biol. 2018 Jul 27;14(7):e1006110.

42. Rigoli F, Pezzulo G, Dolan R, Friston K. A Goal-Directed Bayesian Framework for Categorization. Front Psychol [Internet]. 2017 [cited 2022 Jan 23];8. Available from: https://www.frontiersin.org/article/10.3389/fpsyg.2017.00408

43. Adler WT, Ma WJ. Comparing Bayesian and non-Bayesian accounts of human confidence reports. PLOS Comput Biol. 2018 Nov 13;14(11):e1006572.

44. Knill DC, Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci. 2004 Dec 1;27(12):712–9.

45. Kording KP. Bayesian statistics: relevant for the brain? Curr Opin Neurobiol. 2014 Apr 1;25:130–3.

46. Piasini E, Balasubramanian V, Gold JI. Effect of Geometric Complexity on Intuitive Model Selection. In: Nicosia G, Ojha V, La Malfa E, La Malfa G, Jansen G, Pardalos PM, et al., editors. Machine Learning, Optimization, and Data Science. Cham: Springer International Publishing; 2022. p. 1–24. (Lecture Notes in Computer Science).

47. Weiss Y, Simoncelli EP, Adelson EH. Motion illusions as optimal percepts. Nat Neurosci. 2002 Jun;5(6):598–604.

48. Rangelov D, Mattingley JB. Evidence accumulation during perceptual decision-making is sensitive to the dynamics of attentional selection. NeuroImage. 2020 Oct 15;220:117093.

49. Nett N, Bröder A, Frings C. When irrelevance matters: Stimulus-response binding in decision making under uncertainty. J Exp Psychol Learn Mem Cogn. 2015 Nov;41(6):1831–48.

50. Winkel KN de, Katliar M, Bülthoff HH. Forced Fusion in Multisensory Heading Estimation. PLOS ONE. 2015 May 4;10(5):e0127104.

51. Wozny DR, Beierholm UR, Shams L. Probability Matching as a Computational Strategy Used in Perception. PLOS Comput Biol. 2010 Aug 5;6(8):e1000871.

52. Rahnev D, Denison RN. Suboptimality in perceptual decision making. Behav Brain Sci. 2018 ed;41:e223.

53. Ding N, Simon JZ. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci. 2012 Jul 17;109(29):11854–9.

54. Mesgarani N, Chang EF. Selective cortical representation of attended speaker in multi-talker speech perception. Nature. 2012 May;485(7397):233–6.

55. Schwartz ZP, David SV. Focal Suppression of Distractor Sounds by Selective Attention in Auditory Cortex. Cereb Cortex. 2018 Jan 1;28(1):323–39.

56. Jensen A, Merz S, Spence C, Frings C. Interference of irrelevant information in multisensory selection depends on attentional set. Atten Percept Psychophys. 2020 Apr 1;82(3):1176–95.

57. Chen J, Scotti PS, Dowd EW, Golomb JD. Neural Representations of Task-relevant and Task-irrelevant Features of Attended Objects [Internet]. bioRxiv; 2021 [cited 2022 Jul 14]. p. 2021.05.21.445168. Available from: https://www.biorxiv.org/content/10.1101/2021.05.21.445168v1

58. Xu Y. The Neural Fate of Task-Irrelevant Features in Object-Based Processing. J Neurosci. 2010 Oct 20;30(42):14020–8.

59. Hansen K, Hillenbrand S, Ungerleider L. Persistency of Priors-Induced Bias in Decision Behavior and the fMRI Signal. Front Neurosci [Internet]. 2011 [cited 2022 Nov 17];5. Available from: https://www.frontiersin.org/articles/10.3389/fnins.2011.00029

60. Roark CL, Holt LL. Long-term priors constrain category learning in the context of short-term statistical regularities. Psychon Bull Rev [Internet]. 2022 May 6 [cited 2022 May 17]; Available from: https://doi.org/10.3758/s13423-022-02114-z

61. Hansen KA, Hillenbrand SF, Ungerleider LG. Human brain activity predicts individual differences in prior knowledge use during decisions. J Cogn Neurosci. 2012 Jun;24(6):1462–75.

62. Gekas N, McDermott KC, Mamassian P. Disambiguating serial effects of multiple timescales. J Vis. 2019 Jun 28;19(6):24.

63. Tardiff N, Suriya-Arunroj L, Cohen YE, Gold JI. Rule-based and stimulus-based cues bias auditory decisions via different computational and physiological mechanisms [Internet]. bioRxiv; 2021 [cited 2022 Mar 2]. p. 2021.12.09.471952. Available from: https://www.biorxiv.org/content/10.1101/2021.12.09.471952v1

64. Ming VL, Holt LL. Efficient coding in human auditory perception. J Acoust Soc Am. 2009 Sep 1;126(3):1312–20.

65. Smith EC, Lewicki MS. Efficient auditory coding. Nature. 2006 Feb;439(7079):978–82.

66. Stilp CE, Kluender KR. Stimulus Statistics Change Sounds from Near-Indiscriminable to Hyperdiscriminable. PLOS ONE. 2016 Aug 10;11(8):e0161001.

67. Heald JB, Lengyel M, Wolpert DM. Contextual inference underlies the learning of sensorimotor repertoires. Nature. 2021 Dec;600(7889):489–93.

68. Knill DC, Richards W. Perception as Bayesian Inference. Cambridge University Press; 1996. 530 p.

69. Ashwood ZC, Roy NA, Stone IR, Urai AE, Churchland AK, Pouget A, et al. Mice alternate between discrete strategies during perceptual decision-making. Nat Neurosci. 2022 Feb 7;1–12.

70. Honig M, Ma WJ, Fougnie D. Humans incorporate trial-to-trial working memory uncertainty into rewarded decisions. Proc Natl Acad Sci. 2020 Apr 14;117(15):8391–7.

71. Ihlefeld A, Shinn-Cunningham B. Spatial release from energetic and informational masking in a selective speech identification task. J Acoust Soc Am. 2008 Jun;123(6):4369–79.

72. Middlebrooks JC, Waters MF. Spatial Mechanisms for Segregation of Competing Sounds, and a Breakdown in Spatial Hearing. Front Neurosci [Internet]. 2020 [cited 2022 Feb 9];14. Available from: https://www.frontiersin.org/article/10.3389/fnins.2020.571095

73. Bronkhorst AW. The cocktail-party problem revisited: early processing and selection of multi-talker speech. Atten Percept Psychophys. 2015 Jul 1;77(5):1465–87.

74. Shinn-Cunningham BG. Object-based auditory and visual attention. Trends Cogn Sci. 2008 May 1;12(5):182–6.

75. Wöstmann M, Herrmann B, Maess B, Obleser J. Spatiotemporal dynamics of auditory attention synchronize with speech. Proc Natl Acad Sci. 2016 Apr 5;113(14):3873–8.

76. Holt LL, Lotto AJ. Speech perception as categorization. Atten Percept Psychophys. 2010 Jul;72(5):1218–27.

77. Freedman DJ, Riesenhuber M, Poggio T, Miller EK. A Comparison of Primate Prefrontal and Inferior Temporal Cortices during Visual Categorization. J Neurosci. 2003 Jun 15;23(12):5235–46.

78. Zhong L, Zhang Y, Duan CA, Pan J, Xu N long. Dynamic and causal contribution of parietal circuits to perceptual decisions during category learning [Internet]. bioRxiv; 2018 [cited 2022 Jul 14]. p. 304071. Available from: https://www.biorxiv.org/content/10.1101/304071v3

79. Runyan CA, Piasini E, Panzeri S, Harvey CD. Distinct timescales of population coding across cortex. Nature. 2017 Aug;548(7665):92–6.

80. Akrami A, Kopec CD, Diamond ME, Brody CD. Posterior parietal cortex represents sensory history and mediates its effects on behaviour. Nature. 2018 Feb;554(7692):368–72.

81. Bianco R, Harrison PM, Hu M, Bolger C, Picken S, Pearce MT, et al. Long-term implicit memory for sequential auditory patterns in humans. Shinn-Cunningham BG, Obleser J, Schroger E, Bieszczad K, editors. eLife. 2020 May 18;9:e56073.

82. Grinband J, Hirsch J, Ferrera VP. A neural representation of categorization uncertainty in the human brain. Neuron. 2006 Mar 2;49(5):757–63.

83. van Bergen RS, Ji Ma W, Pratte MS, Jehee JFM. Sensory uncertainty decoded from visual cortex predicts behavior. Nat Neurosci. 2015 Dec;18(12):1728–30.

84. Fetsch CR, Pouget A, DeAngelis GC, Angelaki DE. Neural correlates of reliability-based cue weighting during multisensory integration. Nat Neurosci. 2012 Jan;15(1):146–54.

85. Woods KJP, Siegel MH, Traer J, McDermott JH. Headphone screening to facilitate web-based auditory experiments. Atten Percept Psychophys. 2017 Oct 1;79(7):2064–72.

86. Bankieris KR, Bejjanki VR, Aslin RN. Sensory cue-combination in the context of newly learned categories. Sci Rep. 2017 Sep 7;7(1):10890.

87. Bejjanki VR, Clayards M, Knill DC, Aslin RN. Cue Integration in Categorical Tasks: Insights from Audio-Visual Speech Perception. PLOS ONE. 2011 May 26;6(5):e19812.

88. McPherson MJ, McDermott JH. Diversity in pitch perception revealed by task dependence. Nat Hum Behav. 2018 Jan;2(1):52–66.

89. Moller H, Pedersen CS. Hearing at low and infrasonic frequencies. Noise Health. 2004 Apr 1;6(23):37.

90. Ashihara K. Hearing thresholds for pure tones above 16kHz. J Acoust Soc Am. 2007 Sep;122(3):EL52–7.

91. Jas M, Achakulvisut T, Idrizović A, Acuna D, Antalek M, Marques V, et al. Pyglmnet: Python implementation of elastic-net regularized generalized linear models. J Open Source Softw. 2020 Mar 1;5(47):1959.

92. Cavanaugh JE. Unifying the derivations for the Akaike and corrected Akaike information criteria. Stat Probab Lett. 1997 Apr 30;33(2):201–8.

S1 Fig. Participants weigh tones according to their relevance; distractors adversely affect accuracy.

S2 Fig. For 3 representative participants, their posteriors derived from the full Bayesian model fits and weights from a fitted generalized linear model.

S3 Fig. Bayesian and GLM analysis of 3 additional participants.

S4 Fig. Comparison of the two model fitting procedures.

S5 Fig. Participants' measure of stimulus relevance is tone-position dependent. However, similar to the full Bayesian model, tone relevance for all three tone positions influences category choice but not accuracy.

S6 Fig. Participant accuracy on the previous trial has small effect on participants' behavior.

S7 Fig. Participants are variable in their usage of short-term and long-term learning in the biased low and high sessions.

S8 Fig. The influence of long-term learning grows with time within the block.

S9 Fig. Most participants account for stimulus relevance in the *biased low* and *high* sessions.

S10 Fig: Time constant of short-term learning is inversely correlated with accuracy in the *biased low* and *high* sessions.

S11 Fig. Out of all the factors – sensory uncertainty, relevance uncertainty, short-term learning and long-term learning, participants' categorization accuracy is largely determined by their sensory uncertainty.

S1 Table. Statistical Comparisons.

S2 Table. Parameter reconstruction for 4 virtual participants.

S3 Table: Validating the GLM for 6 virtual participants.

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 (R01NS113241) 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

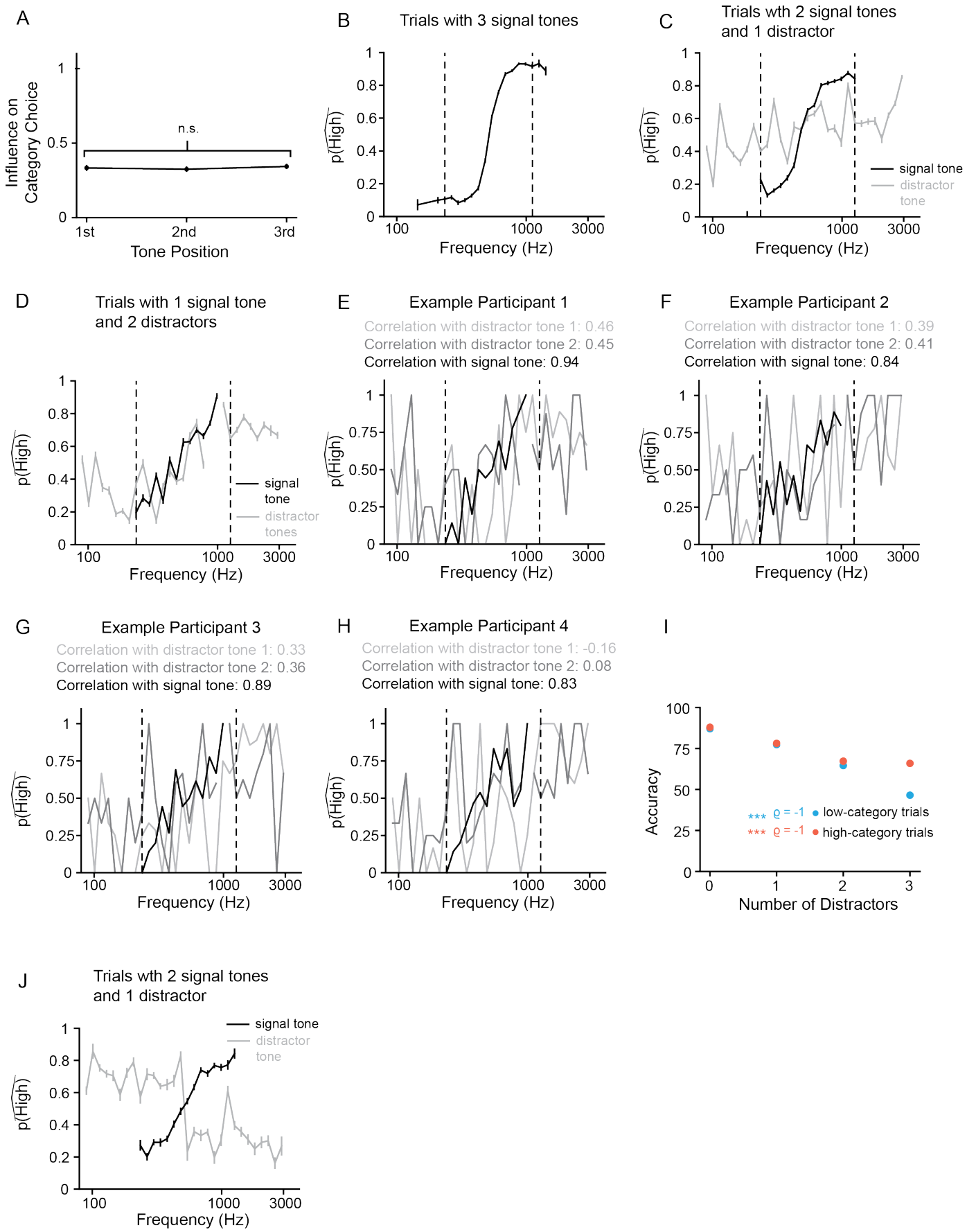
Contributions
J.S., J.C, K.K., E.P., Y.C. and M.N.G. designed the study. J.S. collected and analysed the data. J.J., K.K., Y.C. and M.N.G. wrote the paper.

Corresponding authors
Correspondence to Janaki Sheth or Maria Neimark Geffen.

Supporting Information

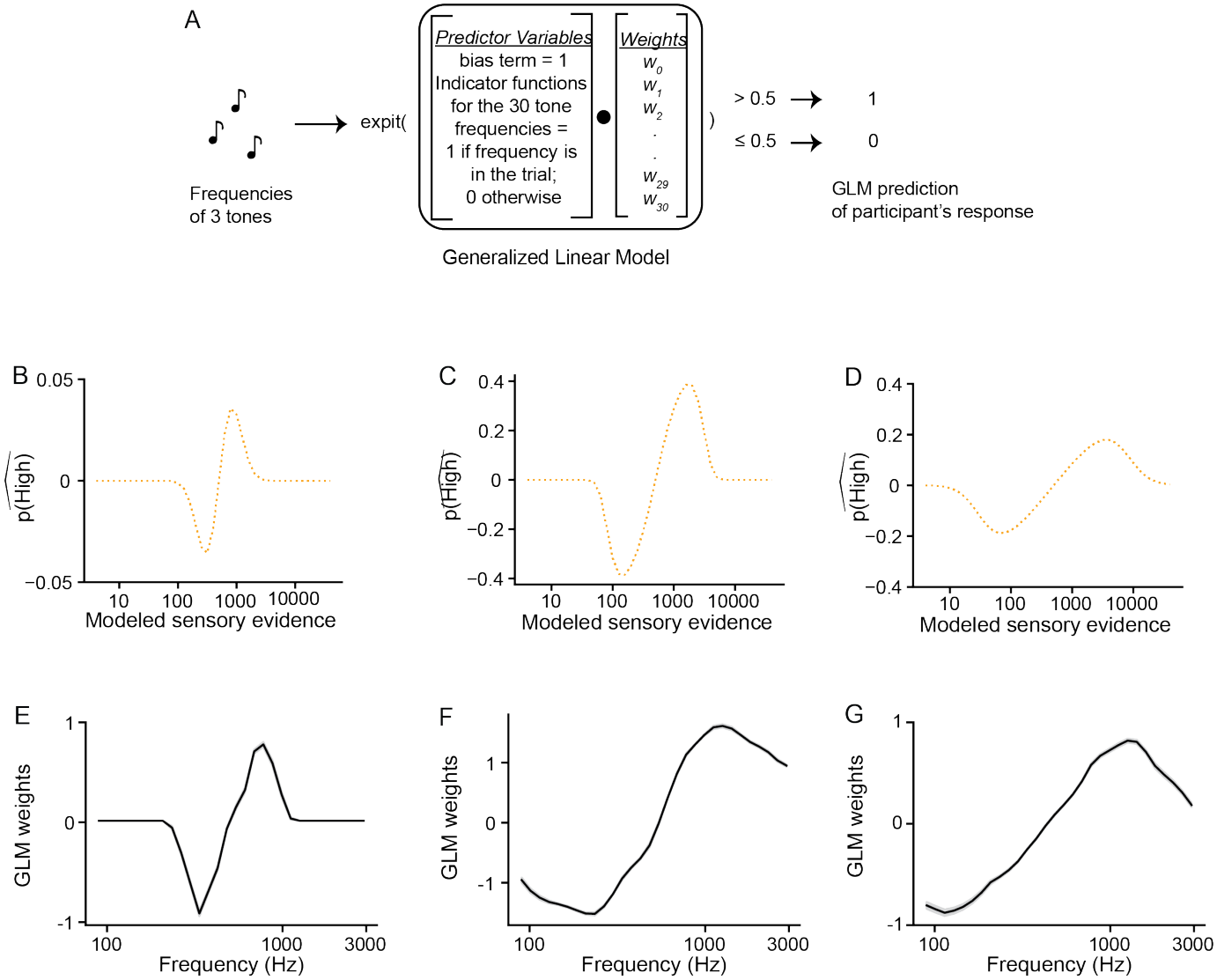
The generalized linear model (GLM) is a descriptive model (Methods). It assigns weights to each of the 30 tone frequencies (Fig S2A) based on a participant's category decisions. The weights are indicative of the contribution of the individual tone frequencies to the participant's decision-making process.

For example, for the representative high-accuracy participant, weights for tones far from the means of the two categories were ~ 0 (Fig S2E, S3A), whereas those for tones that were near the category means had non-zero values. However, for participants with lower accuracy (Fig S2F,G, 3D,G), 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.


A
Influence on Category Choice
n.s.
1st
2nd
3rd
Tone Position
B
Trials with 3 signal tones
p(High)
Frequency (Hz)
C
Trials wth 2 signal tones and 1 distractor
signal tone
distractor tone
D
Trials with 1 signal tone and 2 distractors
signal tone
distractor tones
E
Example Participant 1
Correlation with distractor tone 1: 0.46
Correlation with distractor tone 2: 0.45
Correlation with signal tone: 0.94
F
Example Participant 2
Correlation with distractor tone 1: 0.39
Correlation with distractor tone 2: 0.41
Correlation with signal tone: 0.84
G
Example Participant 3
Correlation with distractor tone 1: 0.33
Correlation with distractor tone 2: 0.36
Correlation with signal tone: 0.89
H
Example Participant 4
Correlation with distractor tone 1: -0.16
Correlation with distractor tone 2: 0.08
Correlation with signal tone: 0.83
I
Accuracy
*** ϱ = -1 low-category trials
*** ϱ = -1 high-category trials
Number of Distractors
J
Trials wth 2 signal tones and 1 distractor
signal tone
distractor tone

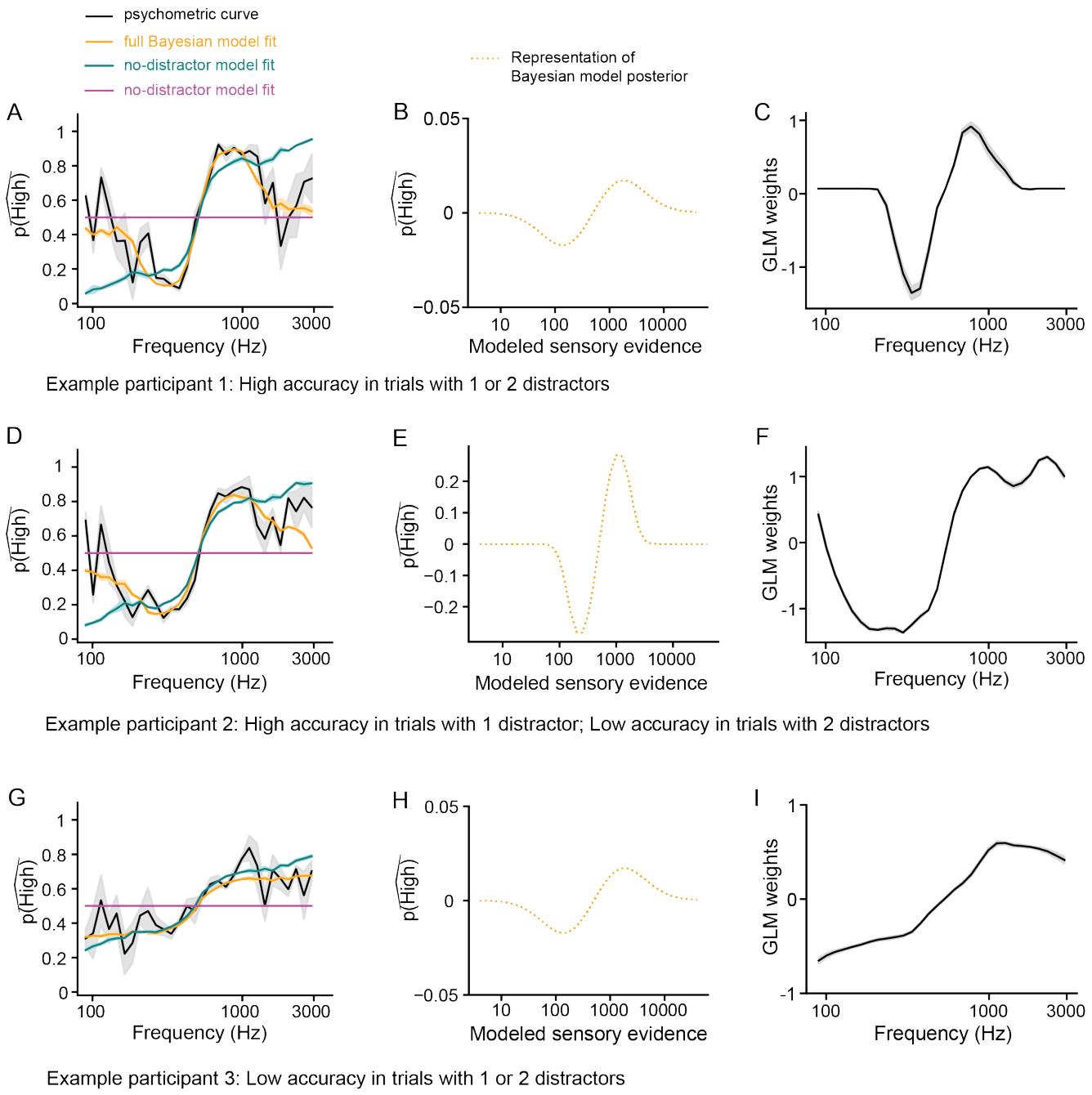
S1 Fig: Participants integrate information over all 3 tones while weighing them according to their relevance; distractors adversely affect accuracy.

(A) In *unbiased* trials, the three temporal positions had equal influence on participants' category choice i.e., the tone frequencies in each of the three positions correlated equally with participants' category choice probability. (B) 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 $p(\widehat{Hıgh})$ values. Average psychometric curves plotted separately for signal and distractor tones (C) for trials with one distractor tone and (D) for trials with two distractor tones. Also, to compute the correlations in Figs 2 and 3, 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. (E-H) Correlations for trials with two distractor tones for 4 example participants. (I) Accuracy for trials from both categories is inversely correlated with number of distractors (Statistics: Spearman correlation; Table S1). (J) 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 (B-D, I, J), mean and SEM over N=56 participants; ***p<0.0001.

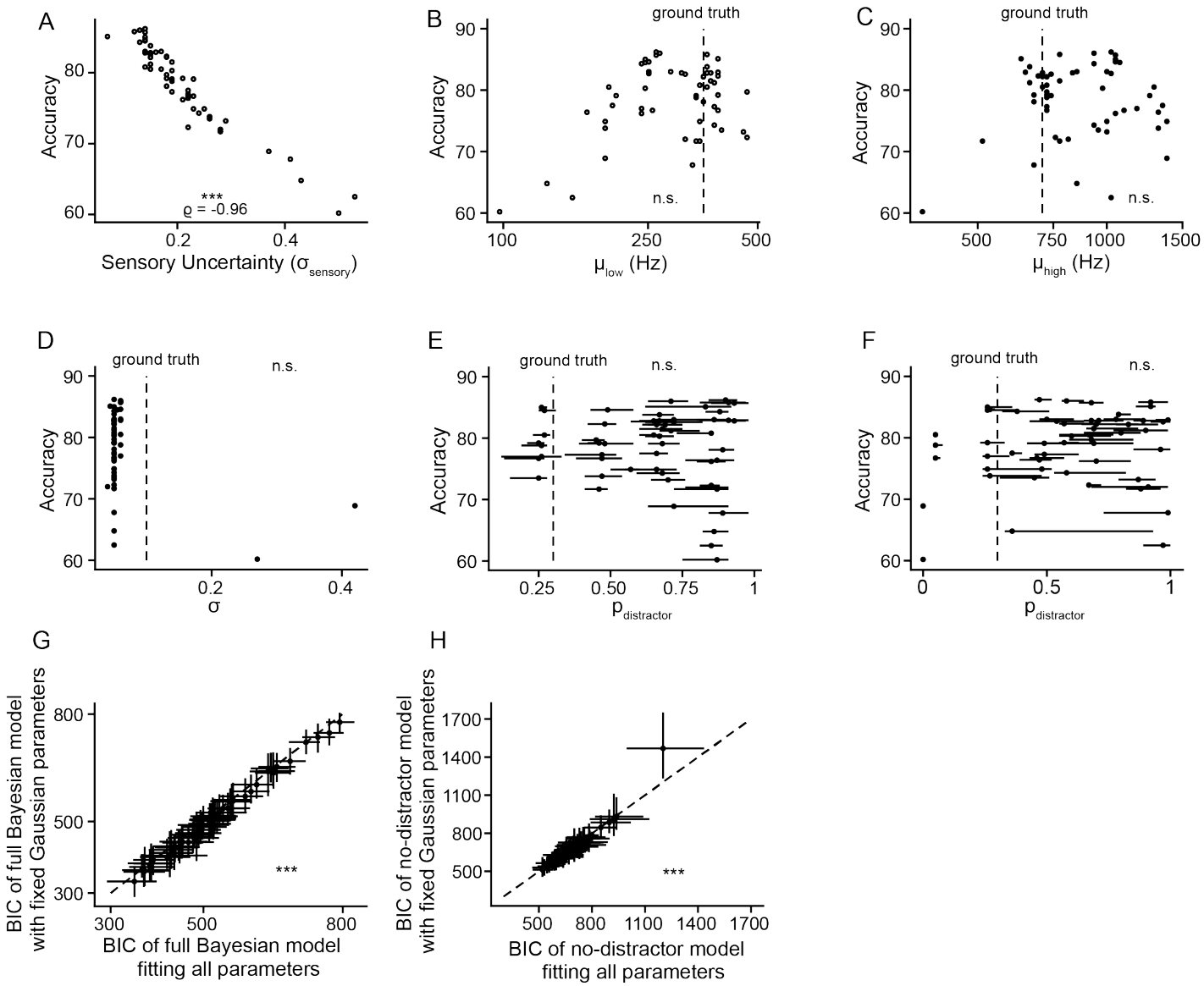


S2 Fig: For 3 representative participants, their posteriors derived from the full Bayesian model fits and weights from a fitted generalized linear model.

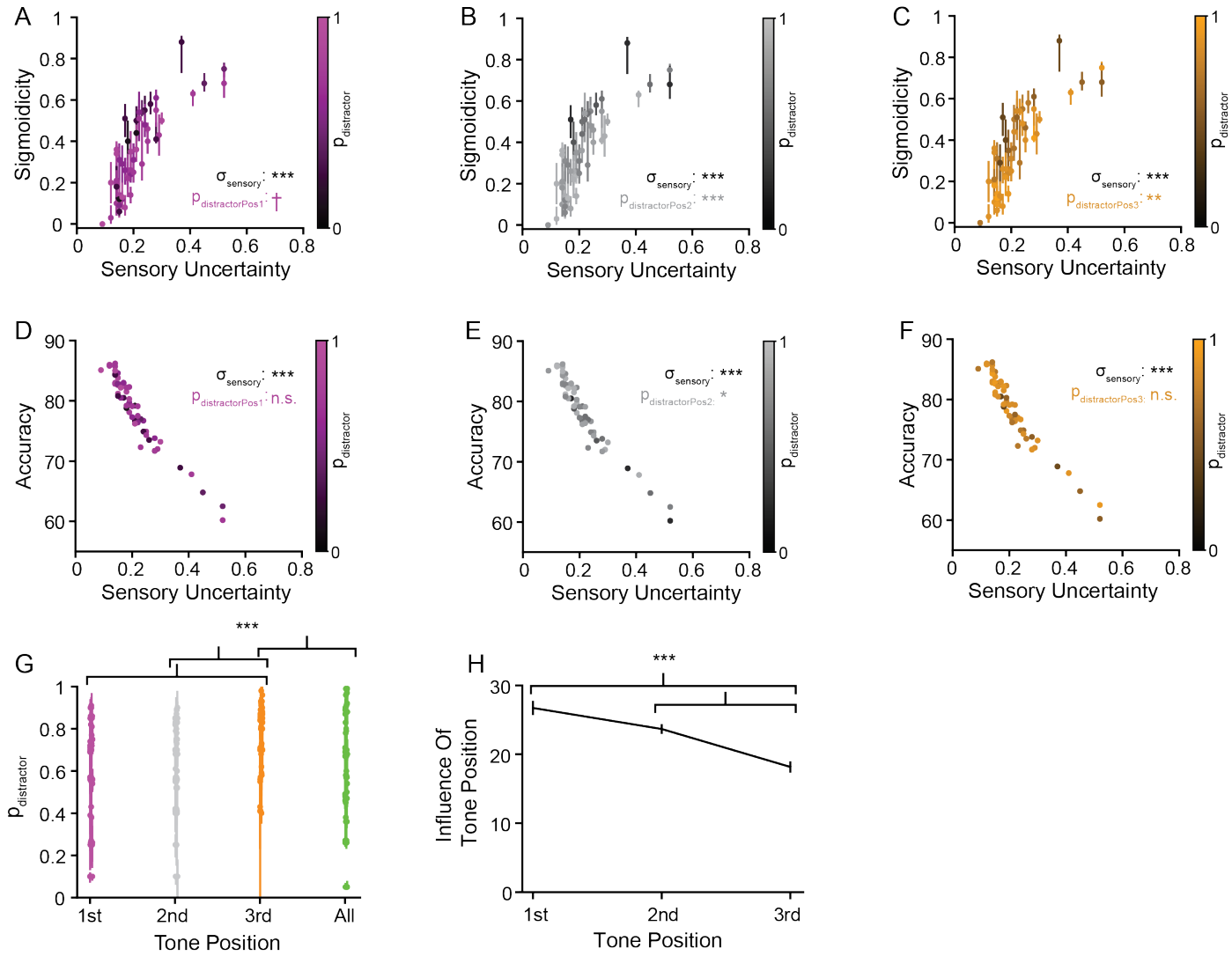
(A) Schematic of the generalized linear model (GLM) that has the frequencies of the three tones in a trial as input and uses the corresponding indicator functions as its predictors to predict the participant's category choice response (see Methods). (B-D) Posteriors generated from the full Bayesian model fits to the representative participants' data in Fig 5. 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-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).



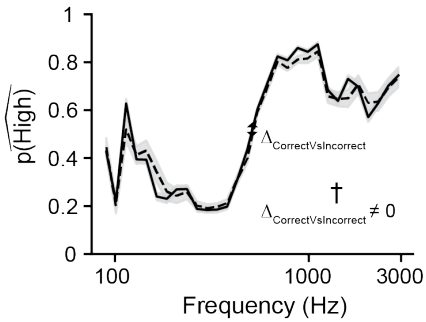
S3 Fig: Bayesian and GLM analysis of 3 additional participants.
(A, D, G) Similar to Fig. 5 psychometric curves for 3 additional participants; black lines: raw data, orange lines: full Bayesian model fit, teal lines: no-distractor model fit, mediumorchid lines: random-guess model fit. (B, E, H) Posteriors 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 5 and Fig S2. (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.



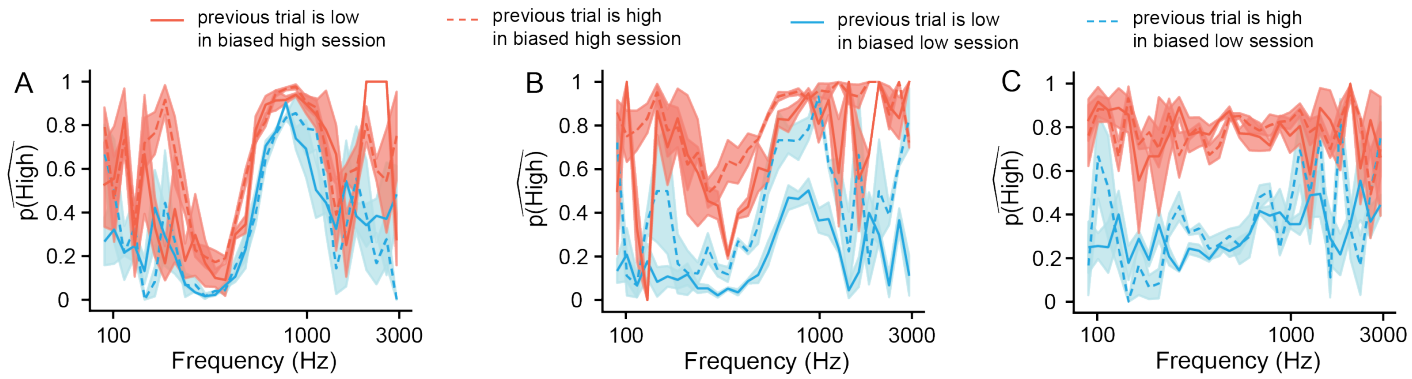
S4 Fig: Comparison of the two model fitting procedures.
(A) Values for $\sigma_{sensory}$ (i.e., sensory uncertainty) are the same for both fitting procedures and are inversely correlated with accuracy. (B-E) Parameter values obtained from fitting all 6 parameters of the full Bayesian model to participants' data. Parameters (B) μ_{low}, (C) μ_{high}, (D) σ of the two Gaussians, and (E) $p_{distractor}$ are uncorrelated with accuracy. (F) corresponds to the parameter values of $p_{distractor}$ obtained using the second fitting procedure wherein participants are considered veridical about the Gaussian parameters. Vertical lines in (B-E) denote 'ground-truth' values, which are the true parameter values used in the generative process. Panels (A-F) Statistics: Spearman Correlation (Table S1). (G, H) 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) Statistics: two-tailed Wilcoxon signed-rank test comparing BIC scores from the two fitting procedures (Table S1). In all panels, data point: one participant; $^{\dagger}p<0.05$; $^{*}p<0.01$, $^{**}p<0.001$, $^{***}p<0.0001$.



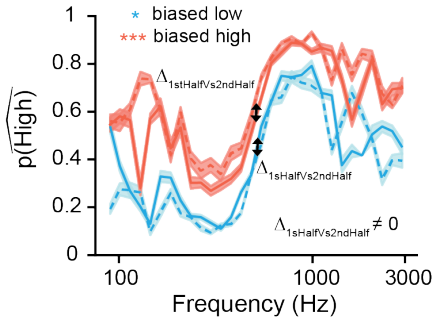
S5 Fig: Participants' measure of stimulus relevance is tone-position dependent. However, similar to the full Bayesian model, tone relevance for all three tone positions influences category choice but not accuracy.
(A) Participants' estimates of the probability of distractors are statistically similar for the first ($p_{distractor_{position_1}}$) and second ($p_{distractor_{position_2}}$) tone positions but are higher for the third ($p_{distractor_{position_3}}$) tone position (one-way rmANOVA; Statistics: two-sided Wilcoxon signed-rank test). Only estimates for the third position are different from those ($p_{distractor}$) derived using the full Bayesian model. (B-D) Similar to Fig. 6B, both $p_{distractor}$ and $\sigma_{sensory}$ affect 'sigmoidicity'; the different colors correspond to the three tone positions as denoted in (A). Statistics: Spearman correlation of respective parameter with 'sigmoidicity'. (E-G) Similar to Fig. 6C, $\sigma_{sensory}$ but not the position-dependent $p_{distractor}$ values influence accuracy; the colors are consistent with (A). Statistics: Spearman correlation of respective parameter with accuracy. (H) Influence of the three tone positions computed using the Generalized Linear Model (see Methods). Statistics: two-sided Wilcoxon signed-rank test. In panels (A-G), data point: one participant; †p<0.05; *p<0.01, **p<0.001, ***p<0.0001.



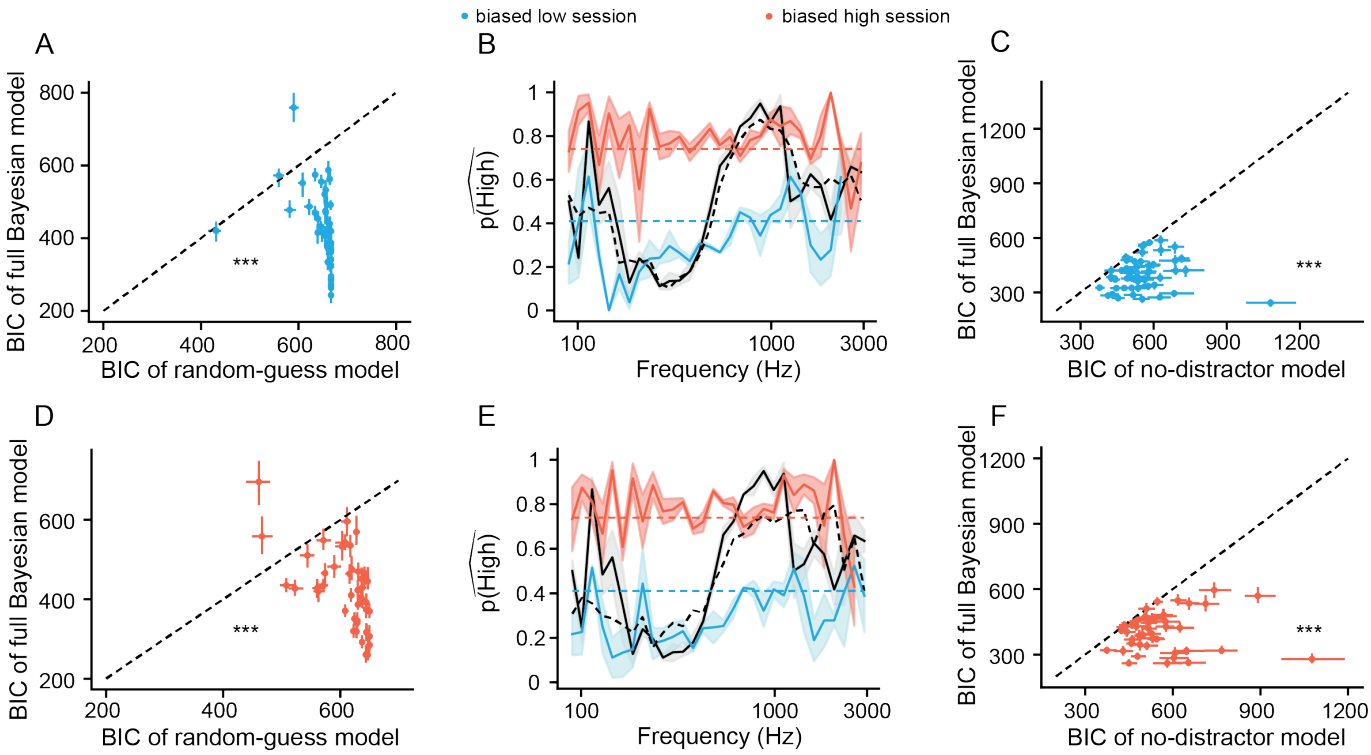
S6 Fig: Participant accuracy on the previous trial has small effect on participants' behavior.
Average psychometric curves conditioned on whether previous choice was correct versus incorrect. $\Delta_{\mathrm{CorrectVsIncorrect}}$: ~0.019; $\Delta_{\mathrm{CorrectVsIncorrect}}$: difference between the two psychometric curves at the central frequency. Error bars: standard error of the mean. Statistics: two-tailed Wilcoxon signed-rank test on $\Delta_{\mathrm{CorrectVsIncorrect}}$ (Table S1), $^{\dagger}p<0.05$.



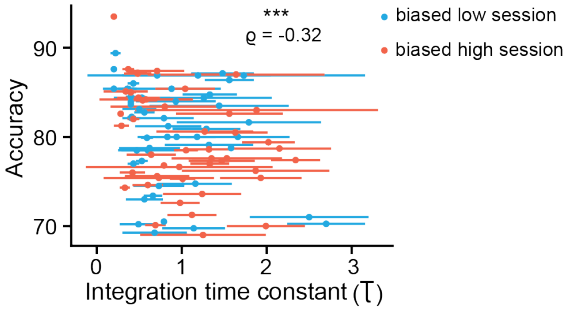
S7 Fig: Participants are variable in their usage of short-term and long-term learning in the *biased low* and *high* sessions. (A-C) Similar to Figs. 8A-C, average psychometric curves conditioned on the previous trial's category type for three participants in the *biased low* (light blue) and *biased high* (red) sessions. These curves are from the same participants as Figs. 9B-D. (A) Example participant who anecdotally does not demonstrate any reliance on the previous trial category nor on session-level bias. (B) Example participant who relies on both the previous trial category and on session-level bias. (C) Example participant who relies on session-level 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.



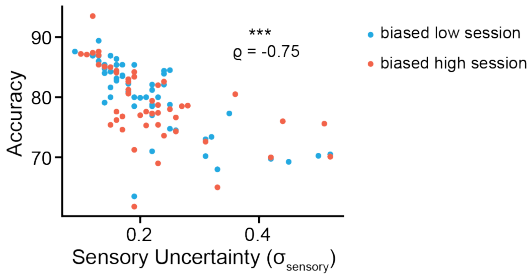
S8 Fig: The influence of long-term learning grows with time within the block.
Psychometric curves for the *biased low* (light blue) and *biased high* (red) sessions, averaged across all participants and conditioned on time within the block; $\Delta_{1stHalfVs2ndHalf}$: ~0.033 (for *biased low*) and ~ -0.063 (for *biased high*). $\Delta_{1stHalfVs2ndHalf}$: difference between the two psychometric curves for each respective session at the central frequency. Statistics: two-tailed Wilcoxon signed-rank test on $\Delta_{1stHalfVs2ndHalf}$ (Table S1), $^{\dagger}p<0.05$; $^{*}p<0.01$, $^{**}p<0.001$, $^{***}p<0.0001$. Error bars: standard error of the mean.



S9 Fig: 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 (light blue dots). Data point: one participant; number of participants = 53. (B) Corresponding plot for the *biased high* session (red 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 S1). Statistics: two-sided Wilcoxon signed-rank test comparing the two model fits (Table S1). (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 (light blue curve – *biased low*, red 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. 5B-E. Respective full Bayesian and random-guess model fits are shown using dotted lines. (E, F) For all other participants besides these two, BIC scores of the full Bayesian model are lower than those of the no-distractor model for both the *biased* sessions. Statistics: two-sided Wilcoxon signed-rank test comparing the two model fits (Table S1). In all plots: ***$p<0.0001$.



S10 Fig: Time constant of short-term learning is inversely correlated with accuracy in the *biased low* and *high* sessions. For participants whose performance was fit using the adapted full Bayesian model (biased low: light blue, N=51; biased high: red, N=46), similar to the *unbiased* trials (Fig 10B), their short-term learning time constant was inversely correlated with their accuracy. Statistics: Spearman correlation; *p< 0.01



S11 Fig: Out of all the factors – sensory uncertainty, relevance uncertainty, short-term learning and long-term learning, participants' categorization accuracy is largely determined by their sensory uncertainty.
For all participants, similar to Fig. 7C, category accuracy was impaired by sensory uncertainty. Data points: one participant; Error bars: standard error of the mean; Statistics: Spearman correlation; ***p<0.0001

S1 Table: Statistical Comparisons.

S2 Table: Parameter reconstruction for 4 virtual participants.

S3 Table: Validating the GLM for 6 virtual participants.