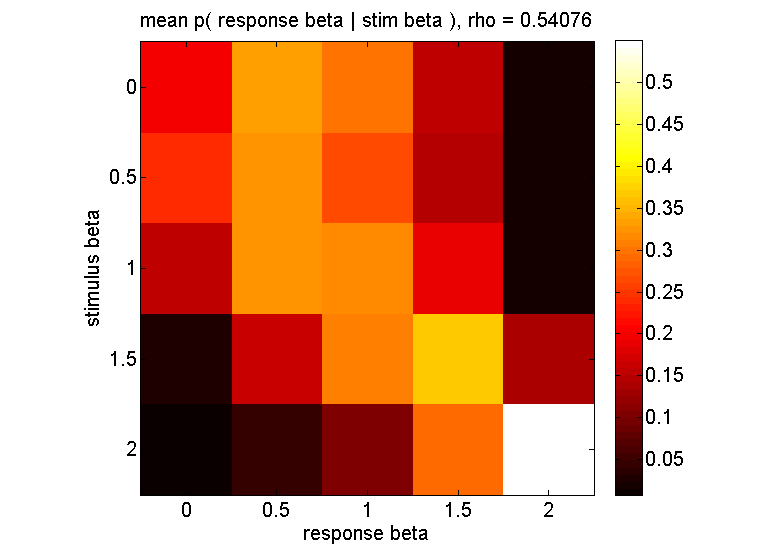
**2/12/14**

**Analysis for SFA experiment 2**

N = 10 subjects have been tested on the behavioral version of experiment 2.

2 subjects are excluded from all below analyses due to extremely poor performance on the beta discrimination task. (cumulative d’ for stimulus beta = 2, relative to stimulus beta = 0, were 0.3 and 0.6)

**Beta discrimination performance**



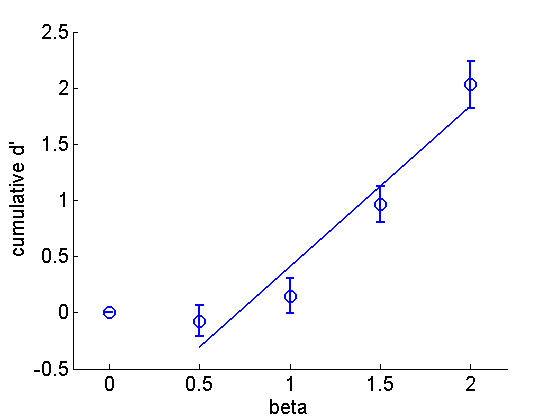
Heat map of p( response beta | stimulus beta ), averaged across subjects. Rows sum to 1. “rho” in the title refers to the mean value of the Spearman’s rho correlation between stimulus beta and response beta, which was calculated for each subject and averaged.

Overall performance is OK. A one sample t-test on the Spearman’s rho for stimulus beta and response beta (first transformed by Fisher’s r-to-z conversion) shows that overall beta discrimination is well above chance (p = 8e-6). However, subjects have a very hard time discriminating betas <= 1.01. Within this range of stimulus beta, mean rho = 0.06, which is not significantly above zero across subject (p = .3).

This pattern is not evident in Schmuckler & Gilden 1993. They had subjects discriminate between beta = 0, 1, and 2 defined over the pitch, duration, and loudness of a sequence of tones. In particular, for discriminating autocorrelations in pitch, subjects correctly identified beta = 0 sequences 66% of the time, and beta = 1 sequences 58.9% of the time (chance = 33.3%).

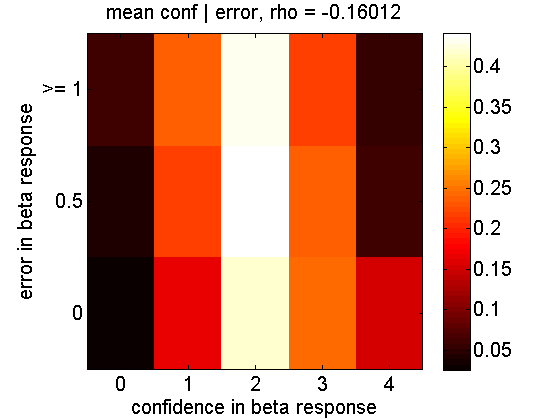
However, although their design was very similar to ours, some relevant differences might explain the different findings.

* S&G use 20 sec long sequences of 200 ms tones, so each sequence has 100 tones. We use 10.2 sec long sequences of 300 ms tones, so each sequence has 34 tones. So subjects in S&G had significantly more information for each sequence with which to discriminate beta. This seems like the most likely source of the discrepancy.
* S&G do not seem to report the nitty-gritty details of how they generated the beta = 1 stimuli. As far as I can tell, they only say that they drew random numbers from a 1/f distribution, without specifying any further. Since using an exact value of beta = 1 is mathematically problematic, they likely used a slightly higher value of beta (cf Patel & Balaban 2000, who used beta = 1.3 for what they called the 1/f condition). So S&G may have used a relatively high beta such as e.g. beta = 1.1 or 1.2. We use beta = 1.01.
* S&G presented 3 levels of beta overall, whereas we present 5. It is possible that requiring finer grained distinctions in beta makes it more difficult to learn and discern coarser differences between e.g. beta = 0 and beta = 1.



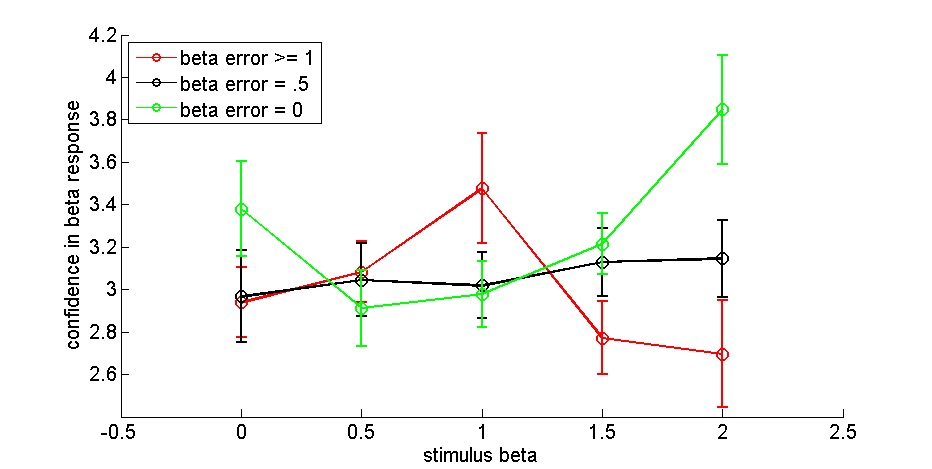
Cumulative d’ as a function of beta, averaged across subjects. D’ is measured here with respect to the beta = 0 stimulus. This graph should be interpreted heuristically, since it assumes the standard deviations of evidence for each level of stimulus beta are equal, whereas the heat map seems to indicate that evidence associated with beta <= 1 is more variable than for beta > 1. Even so, the graph makes it even more evident that subjects cannot discriminate well in the fGn range for these stimuli.

**Confidence rating for beta discrimination**

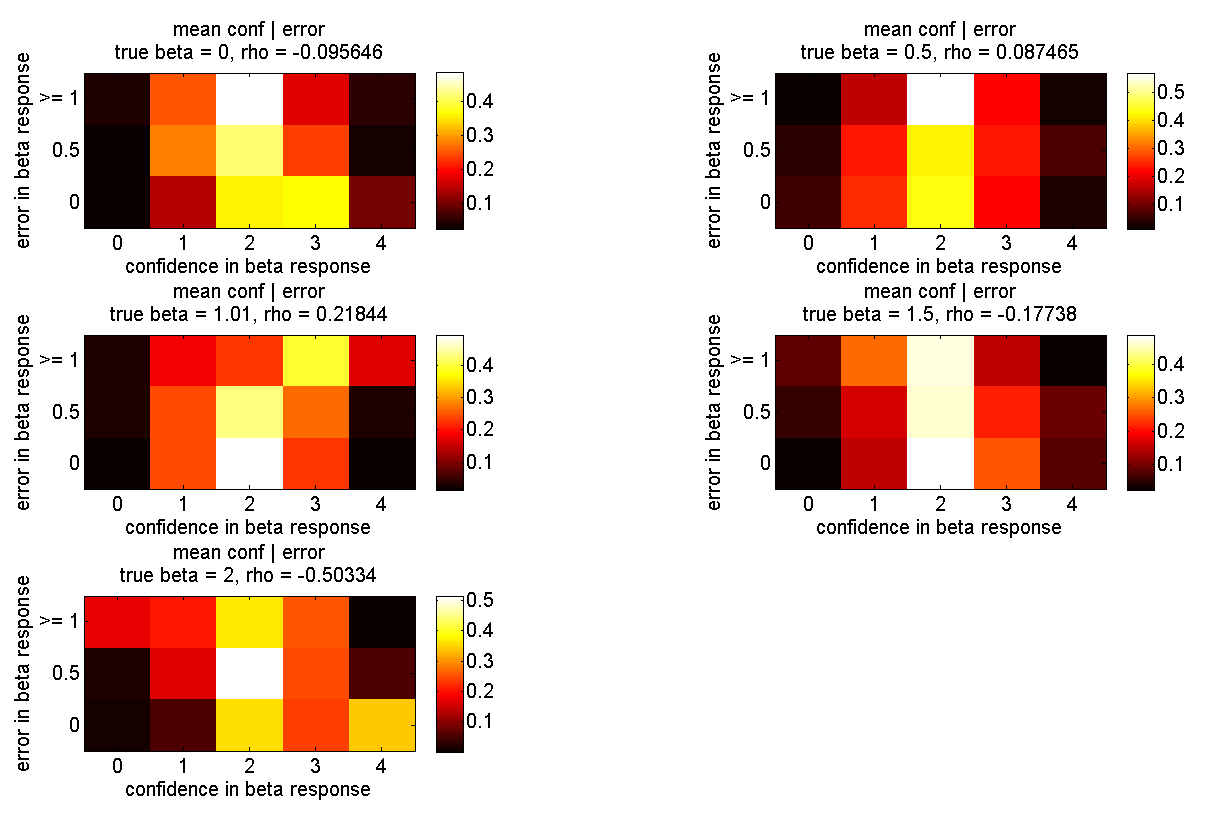
****

Overall, subjects have a hard time placing confidence ratings appropriately. If confidence carries information about task performance, we should expect that higher confidence would be associated with smaller error in the beta discrimination. This is in fact the case, although the effect size is small. The average Spearman’s rho between confidence and error is -.16, which (after r-to-z transformation) is significant in a one-sample t-test at p = .0016.

*Does the informational content of confidence ratings depend on stimulus beta?*

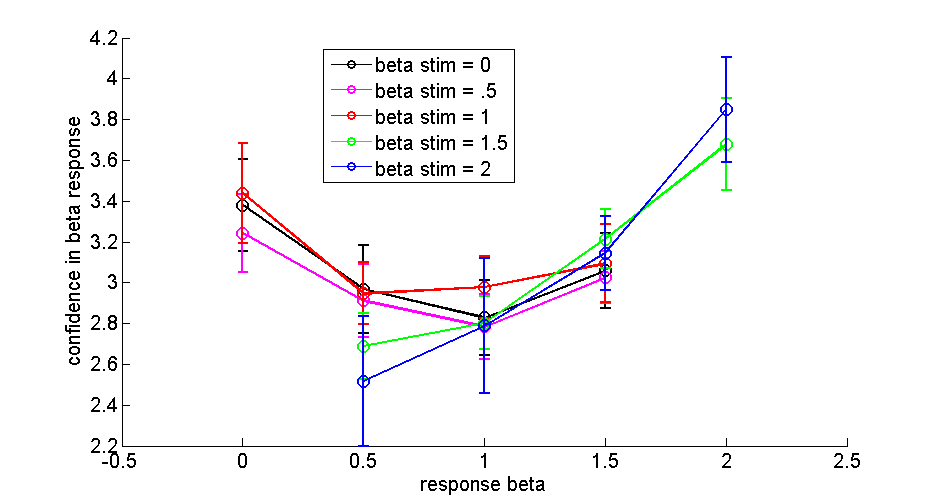
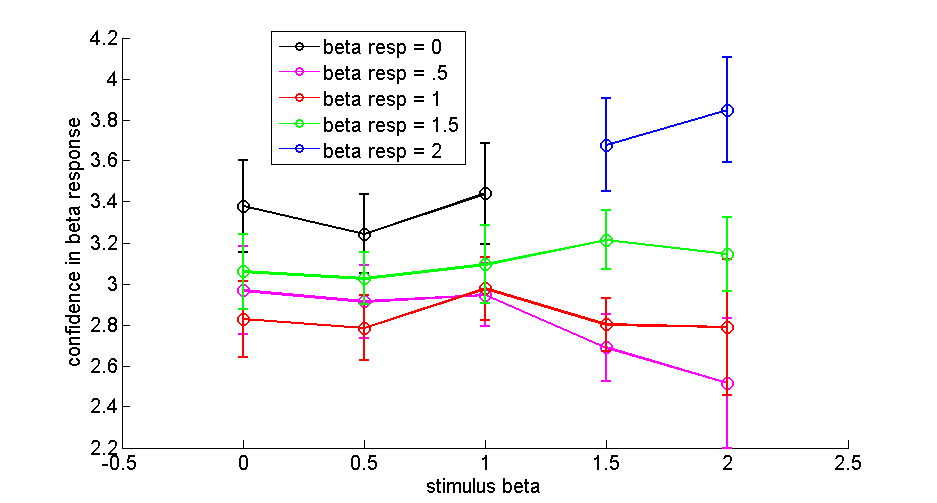
****

Confidence as a function of stimulus beta for different levels of beta discrimination error. Confidence is most informative by far for beta = 2. Paradoxically, for beta = 1, confidence is negatively predictive of accuracy; subjects are more confident when they make large errors for beta = 1 than when they get it right.



Heat maps showing the distribution of mean confidence as a function of error magnitude and stimulus beta. As in the previous figure, the most striking patterns are the strong performance of confidence for beta = 2 and the below-chance performance for beta = 1. The p-values for the r-to-z transforms of rho(confidence, error) in the one-sample t-test at each level of beta are .1, .07, .03, .003, and 8e-4.

*Why is confidence below chance for beta = 1?*



(Curves are incomplete where no data is available for all 8 subjects. E.g. since at least one subject never reported that beta = 2 when the true beta = 1, no corresponding data point is shown on these plots.)

The strange behavior of confidence for beta = 1 appears to just be a general effect of how subjects rate confidence when they enter a response of beta <= 1. In these cases, subjects just generally have higher confidence when they report beta = 0 than when they report beta = .5 or 1, and this pattern does not change according to the true stimulus beta. So the below-chance performance of confidence for beta = 1 is likely attributable to (1) pre-existing patterns of response bias for confidence for each beta response, (2) the fact that these patterns do not differentiate between true stimulus beta.

**Final tone probability rating**

Performance for the final tone probability rating task can’t be assessed without defining the objective features of the stimulus that are to be evaluated. For these analyses we consider 4 possibilities for what the final tone probability rating might be tracking:

1. | log f(t+1) – log f(t) |

i.e. the absolute difference between the pitch of the final tone in the series (t+1) and the penultimate tone (t)

1. | log f(t+1) – E[ log f(t+1) ] |

i.e. the absolute difference between the pitch of the final tone and its *expected* value, which depends on the beta of the tone sequence

1. | log f(t+1) – log f(t) | / σε

i.e. the same as (1), but normalized by the standard deviation of the distribution of pitch values for t+1

1. | log f(t+1) – E[ log f(t+1) ] | / σε

i.e. the same as (2), but normalized by the standard deviation of the distribution of pitch values for t+1

We can call these measures “tone distance metrics.”

Essentially, we want to see if subjects can extract two pieces of information from the tone sequences: first, the expected value of the upcoming tone; and second, the standard deviation of the distribution for the upcoming tone (which depends on the overall standard deviation of the series along with beta). An ideal subject would leverage all possible statistical information from the sequence and use strategy (4).

*Correlations between final tone probability rating and tone distance metrics*

If subjects can meaningfully assess the final tone probability, then higher ratings of probability should be associated with lower tone distance metrics. In other words, we should expect a negative correlation.

Overall, such negative correlations are observed (this time using Pearson’s correlation since the tone metrics span a wider, pseudo-continuous range of values). Here are mean values for Pearson’s r, along with the p-value from a one-sampled t-test on the r-to-z transformed correlations.

metric 1: r = -.29, p =.0045

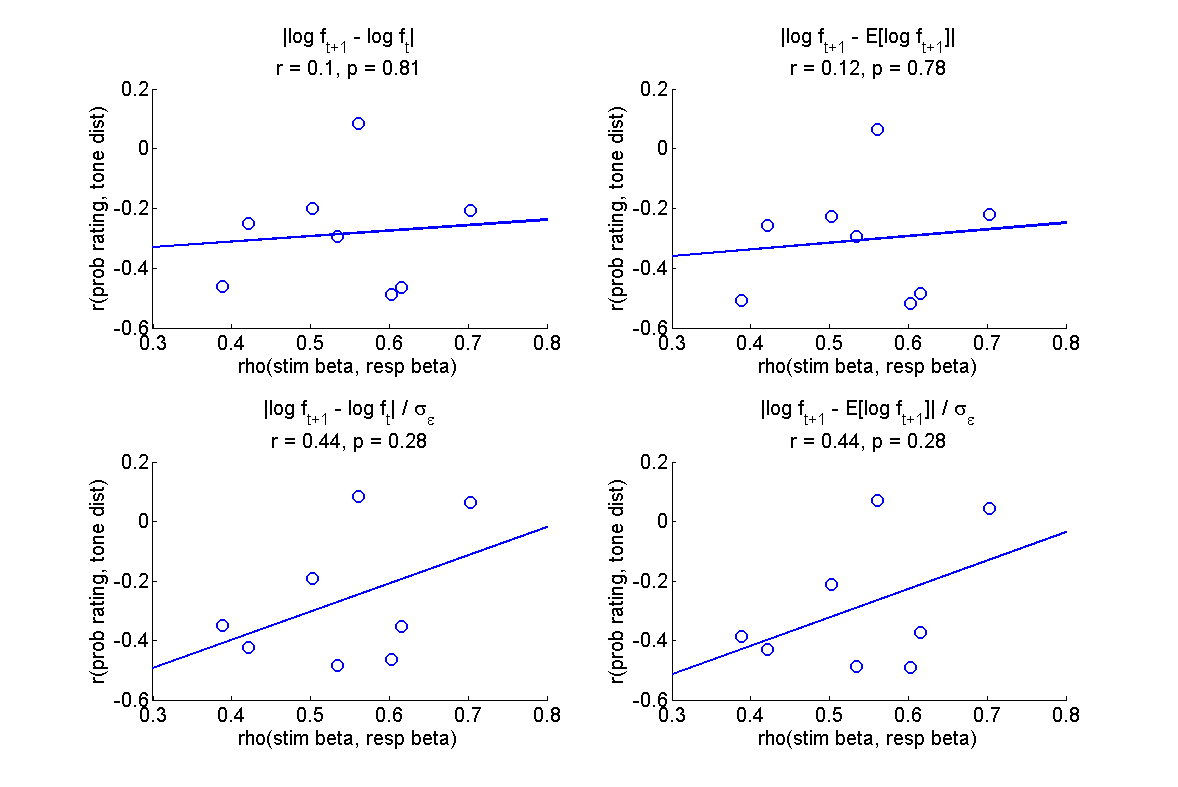
metric 2: r = -.31, p =.0038

metric 3: r = -.26, p = .0132

metric 4: r = -.28, p =.0098

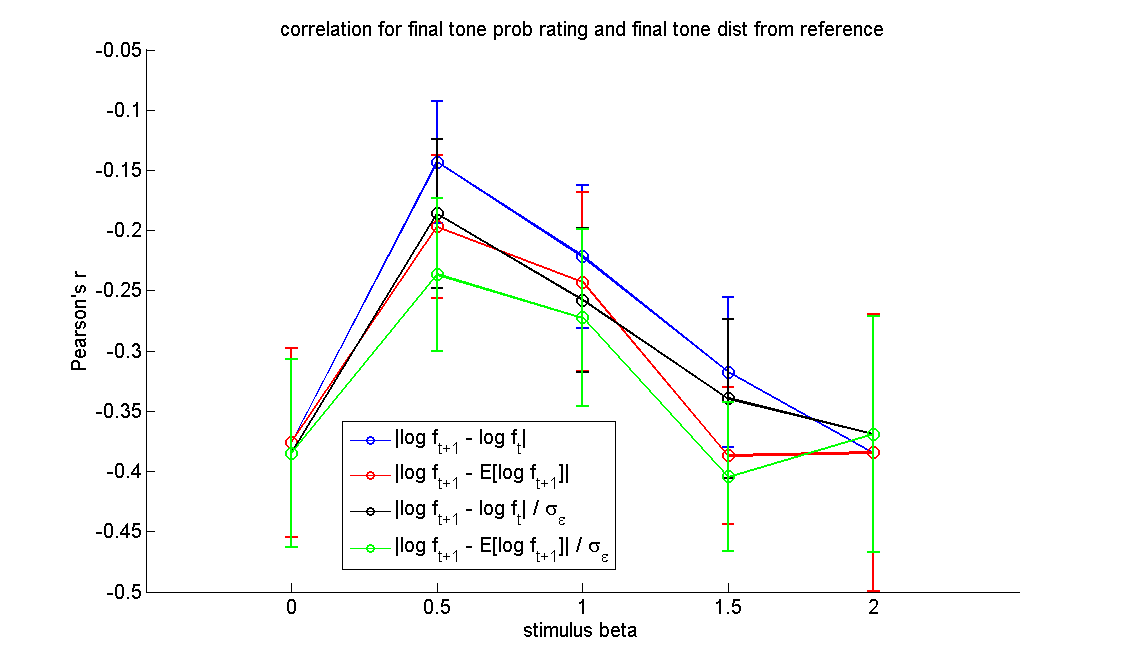
So the final tone probability ratings are informative at a statistically significant level, although the effect sizes are low and similar for the different metrics.

*Is the ability to predict the upcoming tone related to one’s ability to discriminate beta?*

It would seem the answer would have to be “yes,” at least for metrics 2 – 4, since these metrics depend on knowledge of beta. To get a sense of this, here are the correlations between (1) the Pearson’s r correlation coefficients for final tone probability rating and each tone distance metric, and (2) the Spearman’s rho correlation coefficients for stimulus beta and response beta. If better beta discrimination is associated with better probability rating correlation, the overall correlation here should be negative. (higher beta discrim 🡪 better probability rating 🡪 more negative correlation between prob rating and tone dist)

Metrics 1 and 2 exhibit no correlation. 3 and 4 go in the opposite of the expected direction—the correlation is positive rather than negative (though it is not significant). However, the positive correlation for metrics 3 and 4 appears entirely attributable to the two subjects whose probability ratings have zero correlation with metrics 3 and 4, so the direction of this correlation may be a spurious effect.

*Does the information content of final tone probability ratings depend on stimulus beta?*



For all tone distance metrics, the Pearson correlation with final tone probability rating gets weaker for stimulus beta = 0.5 and 1. This makes sense, given that these are beta values for which (1) the expected value of the final tone is strongly modulated by the preceding sequence, and (2) subjects are poor at performing beta discrimination. However, the drop in correlation appears strongest for metric 1 and weakest for metric 4. This might be a hint that subjects are extracted information pertaining to the expected value and/or the standard deviation of the upcoming tone .

**Model selection for tone distance metrics**

Information-theoretic model selection might be a more powerful method than comparing correlation coefficients in order to determine what information subjects use to rate final tone probability. Here is the basic idea.

We want a model that will convert continuous values (tone distance) into an ordinal value (probability rating). So one way to approach this is to use an ordered probit model. The essential idea behind the ordered probit model is that a set of continuous independent variables **x**  is converted to a continuous dependent variable y\* by the equation

y\* = **βx** + ε, ε ~N(0, σ)

Then the continuous y\* is converted to an ordinal variable y by comparing it to a set of thresholds **μ**.

So for this model comparison analysis, we can use the different tone distance metrics as the possible values for x. Specifically, for each subject, we can find the empirically observed probability rating (y) and the value of the tone distance metric (x) for each trial. By using a maximum likelihood estimation procedure, we can then find the values of **β** and **μ** that assign the highest likelihood to the observed set of behavioral outcomes in y across all trials. This likelihood for the entire data set then becomes the basis of model comparison. Better models will assign a higher likelihood to the observed data set.

We can succinctly summarize model comparison results in an intuitive measure by calculating Akaike weights. These weights sum to 1, and can be interpreted roughly as the likelihood of each model, given the data (relative to the entire class of models being considered).

Note that since the ordered probit models for each distance metric have the same number of parameters, the correction for number of parameters component of the AIC analysis does not figure into things, and in particular gives identical results to a similar analysis based on BIC.

In the attached folder “individual subject prob rating” I present graphs of model comparison results for each individual subject.

One subject (LR) performed the tone probability rating task so poorly that model selection results are equivocal. For this subject, the overall correlation between final tone probability rating and tone distance was actually positive (though with small magnitude)—i.e. higher final tone probabiliy ratings were weakly associated with objectively more improbable final tones.

Focusing on the remaining 7 subjects, Tone distance metric 2 (i.e. using the upcoming tone’s expected value, but not its standard deviation, to rate its probability) is selected in 5 of the remaining 7 subjects. For the other 2 subjects, model selection chooses metric 4 (i.e. using both the upcoming tone’s expected value and its standard dev to rate its probability).

If we omit LR’s data, then the average Akaike weights for tone distance metrics 1 – 4 are .06, .60, .09, and .25. The sum of the average Akaike weights for models 2 and 4—i.e. the ones that posit that subjects use the expected value when rating tone probability—is .85. Thus, the model selection seems to provide pretty strong support for the idea that subjects do use the tone’s expected value when rating its probability (or at least, they use some piece of information that is closer to the final tone’s expected value than to the value of the penultimate tone).

That’s the good news. However, the model comparison result does not seem to come out in a very compelling way in the visualization of the data. The model comparison seems sensitive to differences in the data that subjectively seem small; either that, or I haven’t found the best way to visualize the results yet. Part of the messiness of the visualization also comes from the fact that the raw behavioral data is not as pretty and clean-cut as it could be.

Please see the .png figures in the “individual subject prob rating” folder to see these visualizations. The figures are too large to view nicely in a Word document. Here is an explanation of these figures.

Columns

Each is organized into 3 rows and 4 columns. Each column corresponds to the analysis for a given tone distance metric, which is identified in the column title.

Top row

The top row visualizes the relationships between (1) the tone distance metric, (2) behavioral probability ratings, and (3) modeled probability ratings.

Scatterplots are not appropriate for visualizing these data, since there are a limited number of probability ratings and values for tone distance, and many trials have identical values for these two variables.

Therefore, instead I plot average tone distance as a function of probability rating. Tone distance is normalized—i.e. the original set of tone distance values is scaled from zero to one. Then the average normalized tone distance is found for each level of final tone probability rating. (This produces smaller mean tone distances for metrics 3 and 4, which contain more extreme values in the raw data set due to scaling tone distances by sigma\_e.)

On these plots, if subjects perform well, we should expect a negative relationship—higher probability rating is associated with lower tone distance.

“Modeled probability ratings” refers to the following. For each trial, given its objective tone distance value and the maximum likelihood fit of the ordered probit model, we can produce the probability rating value that the probit model “thinks” should be produced for this trial. So using the model fit, we can generate a “modeled” or “model-predicted” probability rating for each trial, and thereby compare the relationship between behavioral and modeled probability ratings with tone distance.

There are two salient features to compare for the behavioral and modeled probability ratings: (1) the average tone distance at each level of probability rating (height of the bars), and (2) the shade of gray of each bar (darker shade of gray 🡪 this probability rating was more frequent in the data set).

The MLE model fitting procedure is sensitive to how frequent each probability rating is in the data. For instance, if a subject enters a probability rating of 2 50% of the time and a probability rating of 3 only 10% of the time, the MLE fit may sacrifice accuracy in modeling the trials where rating=3 in favor of capturing the more salient (frequent) behavior for of trials where rating=2. Thus, if a given level of behavioral probability rating is very low (light colored), the model fit may be associated with a very different average tone distance without this necessarily indicating a poor model fit. But behavioral probability ratings that are more frequent (darker) should be associated with modeled probability ratings that are also darker (more frequent in the modeled set of trials) and have comparable average tone distance.

Middle row

This is a heat map of the joint probability distribution of behavioral and modeled final tone probability ratings across trials. The sum across all rows and columns therefore equals 1.

For instance, suppose that for a given matrix, the row corresponding to modeled probability rating = 3, and the column corresponding to behavioral probability rating = 3, has a yellow color corresponding to a value of ~0.2 (see color bar in figure). This means that for 20% of all trials, both the behavioral and modeled probability ratings were equal to 3.

Better model fits to the data should correspond to hotter colors along the negative diagonal where behavioral probability rating = modeled probability rating.

Bottom row

A simple bar graph showing (1) the Spearman’s rho correlation between behavioral probability rating and tone distance; (2) the Spearman’s rho correlation between behavioral and modeled probability rating; and (3) Akaike weight.