Is the use of confidence ratings in signal detection tasks fundamentally flawed?

Shane T. Mueller & Christoph T. Weidemann

Indiana University, Bloomington

{stmuelle,cweidema}@indiana.edu

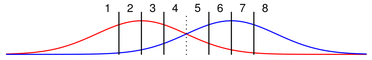
http://mypage.iu.edu/~stmuelle/MuellerWeidemann2005.pdf

Abstract

Recent criticisms of signal detection theory (SDT) have suggested that it is fundamentally flawed in its representations and assumptions (e.g., Balakrishnan, 1999). Evidence for this position includes demonstrations that the detection criterion in confidence ROC functions *does not change* in response to base rate manipulations, whereas the shape of the ROC functions *does* change. We propose an account that explains these findings in terms of participants' well-documented inability to use deterministic decision criteria. We conclude that confidence ratings appear to be unreliable and inaccurate measures of decision certainty, and that the evidence for or against SDT based on these unreliable measures must be used cautiously.

Introduction

Signal detection theory (SDT) assumes that stimuli in a discrimination task give rise to perceptual distributions and that likelihood criteria are set to determine which response to produce:

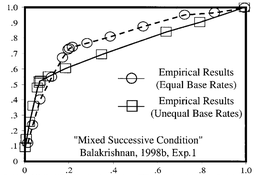


- SDT includes a flexible likelihood ratio decision threshold (β) that controls how willing an observer is to make a particular response. By varying β under a single sensitivity condition, Receiver Operating Characteristic (ROC) functions can be formed.
- One traditional way of forming ROC functions is to collect confidence ratings. Presumably, differing confidence levels correspond to different regions of likelihood (in above figure denoted 1-8).
- Prominent critiques of SDT have questioned (1) how the decision criteria is set, and (2) whether responses are made according to deterministic decision rules. (e.g., Healy & Kubovy, 1981; Treisman & Williams, 1984).
- Recent research using confidence ratings has uncovered two empirical phenomena inconsistent with SDT: (1) The confidence ROC functions produced by base rate manipulations change shape (see below figure); and (2) distributional measures of decision threshold do not change in response to base rate manipulations, despite changes in β.
- Such results suggest that signal detection theory may be fundamentally flawed (Balakrishnan, 1999).
- These new criticisms may stem from the same factors that led to earlier critiques, especially the inability of observers to make deterministic responses to identical perceptual states.
- We conducted a signal detection experiment using an external noise paradigm to determine.

1. How confidence responses are affected by changes in experimental stimulus base rate.
2. The extent to which participants can reliably distinguish multiple confidence responses.

ROC Functions Vary With Baserates

When confidence is collected at varying baserates, the hits and false alarm rates trace out a different ROC functions for each baserate, which is incompatible with SDT.



Part of Figure 1 from Balakrishnan & MacDonald (2002). The data are from Balakrishnan (1998). The false alarm rate is plotted on the abscissa, the hit rate on the ordinate.

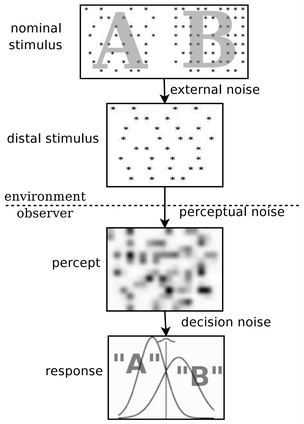
The Decision Noise Hypothesis

Consistent with previous accounts that have attempted to explain probability matching and the learning of criteria placement (e.g., Mueller, 1998; Schoeffler, 1965; Healy & Kubovy, 1981; Erev, 1998), we formalize a generic version of SDT that allows for traditional perceptual noise along with non-deterministic mappings between perceptual state and response: Decision Noise.

- Initially, a nominal stimulus is selected
- The distal stimulus is typically a noisy version of the ideal nominal stimulus.
- Perceptual processes often introduce more noise, producing a noisy percept.
- The mapping from perceptual state onto responses may be non-deterministic.

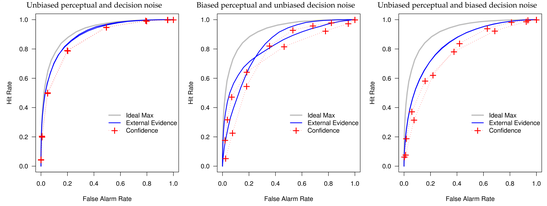
The sources of noise can be discriminated using different forms of ROC functions.

- The unobservable "true" perceptual ROC function is determined by external and perceptual noise, but unaffected by decision noise.
- Observed ROC functions will be weighted averages of the "true" ROC function, and so will fall on or below it.
- Using the presented distal stimulus, one can form an ROC function measuring the response consistency across the range of stimulus levels (blue lines below); using confidence responses, a traditional ROC function can be formed (red lines below).
- Changes in the perceptual distributions should be detectable by both stimulus-based ROC functions and confidence ROC functions (center panel below).
- Changes in response mapping will be detectable in the confidence ROC functions, in ways that will not affect the stimulus-based ROC functions (right panel below).



The Decision Noise Model

We speculate that biased noise may enter into the perceptual-decision process either during perception or decision. Noise during these two phases has different characteristic signatures on observed ROC functions. Left panel shows unbiased perceptual and decision noise; center panel shows biased perceptual and unbiased decision noise; right panel shows unbiased perceptual and biased decision noise.



Experiment

Stimuli: Displays of stars drawn from two overlapping distributions (A: $\mu = 46$ and B: $\mu = 54$; $\sigma = 5$).

Baserates: Stimuli were drawn from the distributions at 3 different ratios (blocked in counterbalanced order): 2:1, 1:1, and 1:2.

Payoff: Rewards of 1, 2, 3, or 4 points and losses of 1, 3, 5, and 7 were given, dependent on accuracy and confidence.

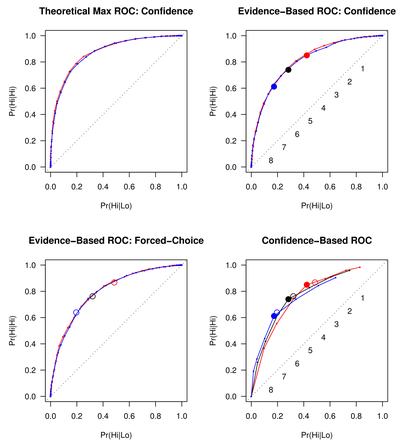
Dependent variable: Confidence ratings with explicit discrimination and forced choice decisions were recorded in different blocks.

Participants: 32 undergraduate students participated in each response and base-rate condition.

Average ROC

ROC functions can be formed based on: the proportions of each nominal stimulus greater than each distal stimulus intensity value (upper left panel); the proportions of "A" and "B" stimuli called "B" greater than each stimulus intensity for the confidence task (upper right) and the forced-choice task (lower left); or the proportions of "A" and "B" stimuli called "B" greater than each confidence rating (lower right panel).

Upper left panel shows the level of external noise present in current experiment. Remaining figures factor out this noise by forming functions based on the ideal distal stimulus category rather than the nominal stimulus.



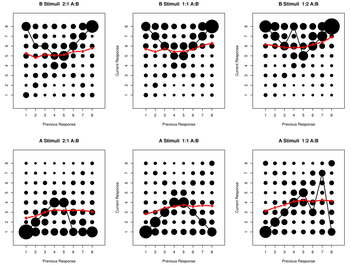
d'/β for forced choice task: 0.87/0.95, 0.82/1.06, 0.9/1.19; confidence task: 0.88/0.91, 0.88/1.01, 0.94/1.13.

Results showed no difference between distal stimulus-based ROC functions as a function of base-rate condition, but did show differences in the confidence ROC function, suggesting that the asymmetries in confidence ratings stem from noise in decisional, rather than the perceptual, processes.

Conditional Response Probabilities

Sequential dependencies in response probabilities for standard SDT tasks have been previously investigated as a source of response inconsistency (e.g., Parducci & Sandusky, 1965; Treisman & Williams, 1984).

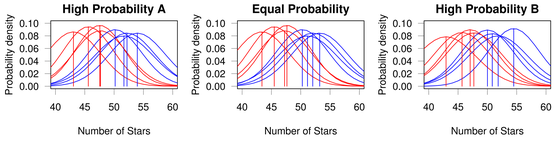
Consistent with these earlier theories (and the decision-noise hypothesis), confidence ratings depend to a large extent on the previous response, especially when the prior stimulus was of the same type as the current stimulus (see figure). Because high base-rate stimuli are more likely to be followed by other high base-rate stimuli while low base-rate stimuli are less likely to be followed by other low base-rate stimuli, sequential dependencies related to repeated stimuli may produce greater response inconsistencies in high base-rate stimuli.



Top row indicates responses from distal "B" stimuli; bottom row for distal "A" stimuli. From left to right, panels show response distributions for high probability "A", equal probability "A" and "B", and high probability "B" conditions. The size of the filled circle is monotonically related to the proportion of responses in that cell. Solid (black) lines connect modal response for each column; hashed (red) lines indicate mean response for each column.

Mean Empirical Response Densities

For each confidence rating, variability across stimulus level was large, with very small discriminability between confidence levels.



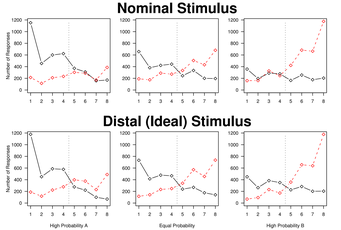
Estimated density of stimulus distributions for each confidence response, computed by averaging fitted mean and standard deviation estimates for each participant.

Decision Threshold is Invariant With Baserate

A second empirical observation inconsistent with SDT is that objective measures of decision threshold do not appear to shift in response to changes in base rate.

This conclusion is based on the sub-optimality seen in the intermediate confidences values for the low-probability response: these responses are given equally often for both stimulus classes.

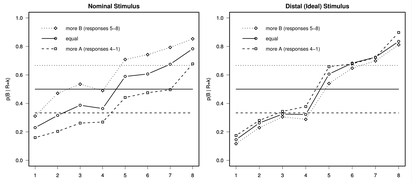
When the distal stimulus is used, the suboptimality disappears.



Response Functions

The demonstration in previous graph shows that observed suboptimalities may occur from truly optimal performance distorted by noise. Clearer evidence can be seen by examining the probability of making the "B" response for each confidence level (below).

For the nominal stimuli (left panel), intermediate confidence levels for low probability stimulus are at or below chance, indicating suboptimality. The hypothesis depicting no criterion shift is supported by the fact that the jump between confidence levels 4 and 5 spans the horizontal line corresponding to each base rate condition. But for the distal stimuli (right panel), suboptimalities are no longer present and objective measures for bias are equivocal regarding whether or not thresholds shift.



Discussion

- Consistent with the decision noise model, results of this experiment indicate that considerable inconsistency exists in the mapping from internal perceptual state onto confidence responses.
- This result is consistent with 40 years of research showing inability of observers to set static cutoffs and tendency to make responses probabilistically in SDT tasks.
- Much of the response inconsistency results from the dependencies of responses on the prior response.
- Such inconsistency could produce suboptimalities noted by Balakrishnan (1999).
- Additionally, similar noise may mask optimal shifts in criterion placement, making responses appear to not shift.
- It may be inappropriate to reject SDT on the basis of data from experiments using confidence ratings, which can introduce suboptimal response mappings that distort ROC functions and measures of bias.
- Based on our data, objective measures could not detect a shift in decision threshold despite changes in β, but they are not incompatible with moderate threshold shifts.
- Future research is needed to determine the extent to which SDT is limited in scope, and the ways it may be extended or elaborated to explain empirical contradictions of its assumptions.

References

- Balakrishnan, J. D. (1998). Measures and interpretations of vigilance performance: Evidence against the detection criterion. *Human Factors, 40*, 601–623.
- Balakrishnan, J. D. (1999). Decision processes in discrimination: fundamental misrepresentations of signal detection theory. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1189–206.
- Balakrishnan, J. D. & MacDonald, J. A. (2002). Decision criteria do not shift: Reply to Treisman. *Psychonomic Bulletin & Review, 9*, 858–865.
- Erev, I. (1998). Signal detection by human observers: A cutoff reinforcement learning model of categorization decisions under uncertainty. *Psychological Review, 105*, 280–298.
- Healy, A. F. and Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 344–354.
- Kac, M. (1969). Some mathematical models in science. *Science, 166* (3906), 695–699.
- Mueller, S. (1998). *Reinforcement learning in signal detection.* Master's Thesis, The University of Michigan, Ann Arbor, MI.
- Parducci, A. and Sandusky, A. (1965). Distribution and sequence effects in judgment. *Journal of Experimental Psychology, 69*, 450–459.
- Schoeffler, M. S. (1965) Theory for psychophysical learning. *The Journal of The Acoustical Society of America, 37*, 1124–1133.
- Treisman, M. and Williams, T. C. (1984). Theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91*, 68–111.

Poster presented at the Annual Meeting of the Society for Judgment and Decision Making, Toronto, ON, CA, 2005.