

Feedback Effects on Cost-Benefit Learning in Perceptual Categorization

C. J. Bohil, W. T. Maddox, & J. L. Dodd

University of Texas at Austin

OVERVIEW

We conducted an experiment designed to assess the impact of different feedback types, error costs, and category discriminability (d') levels in a hypothetical medical diagnosis (i.e., categorization) task. Performance was generally closer to optimal when feedback was based on an attainable level of performance ("optimal-classifier" feedback) rather than based on perfect performance ("objective feedback"), when incorrect responses were accompanied by a cost of zero rather than a negative cost, and in the larger of two d' conditions. Modeling analyses uncovered individual differences among the subjects: about half were sensitive to the feedback and cost manipulations. Model parameters suggest that subjects placed too much weight on accuracy to maximize reward in objective feedback conditions, in negative cost conditions, and in the smaller d' condition.

INTRODUCTION

Categorization judgments are commonly based on uncertain information. For example, determining whether a patient "has the flu" or "does not have the flu" based on their symptoms can be thought of a categorization problem. The patient could exhibit a wide range of symptoms and yet the diagnosis would reflect only one of these two categories. In order to solve categorization problems of this type, the optimal classifier (a hypothetical construct that maximizes long-run reward) sets a decision criterion along the stimulus dimension (e.g., body temperature) and gives one response for dimensional values above this criterion, and the other response for values below the criterion. The location of the optimal criterion is determined by payoff matrix values (i.e., benefits and costs associated with correct & incorrect responses). Importantly, the optimal classifier sacrifices

accuracy in order to maximize long-run reward, and category discriminability (d' ; Green & Swets, 1966) determines the degree of accuracy sacrifice required (higher discriminability = smaller sacrifice). The optimal classifier provides a benchmark for examining human performance in categorization tasks.

Maddox & Bohil (2001) found that categorization performance is affected by benefits and costs as well as category discriminability, and also by the type of feedback provided. In conditions where subjects received "objective" trial-by-trial feedback (i.e., based on the objectively correct response), performance was further from optimal than when subjects received "optimal-classifier" feedback (i.e., based on the performance of the optimal classifier). When the stimulus sets from the two categories overlap, thus making perfect performance impossible, objective feedback portrays a level of performance that is unachievable (i.e., 100% correct). Optimal-classifier based feedback appears to diminish attention to accuracy, resulting in closer-to-optimal performance.

In the study described below, subjects performed a hypothetical medical diagnosis (i.e., categorization) task, and completed several different conditions (i.e., a within-subjects design was used). The study was designed to replicate Maddox & Bohil's (2001) study by factorially combining two levels of category discriminability with two types of feedback (objective & optimal-classifier). The current study extends the previous research by including different cost conditions (either zero cost associated with incorrect responses, as used previously, or negative costs associated with incorrect responses). The details of the study are presented below, followed by a summary of our model-based analyses, which includes some discussion of important hypotheses regarding category discriminability and the trade-off between accuracy and reward maximization.

METHODS

Study Objective: Designed to replicate and extend Maddox & Bohil's (2001) earlier results by testing the hypothesis that negative costs in the payoff matrix will lead to greater weight on accuracy than

when only zero costs exist, thus diminishing performance (i.e., increasing deviation from optimal performance) by de-emphasizing the goal of reward-maximization.

Design

- 8 conditions [2 levels of payoff matrix] x [2 levels of d'] x [2 types of feedback]
- Optimal β of 3.0 for all experimental conditions.

Procedure

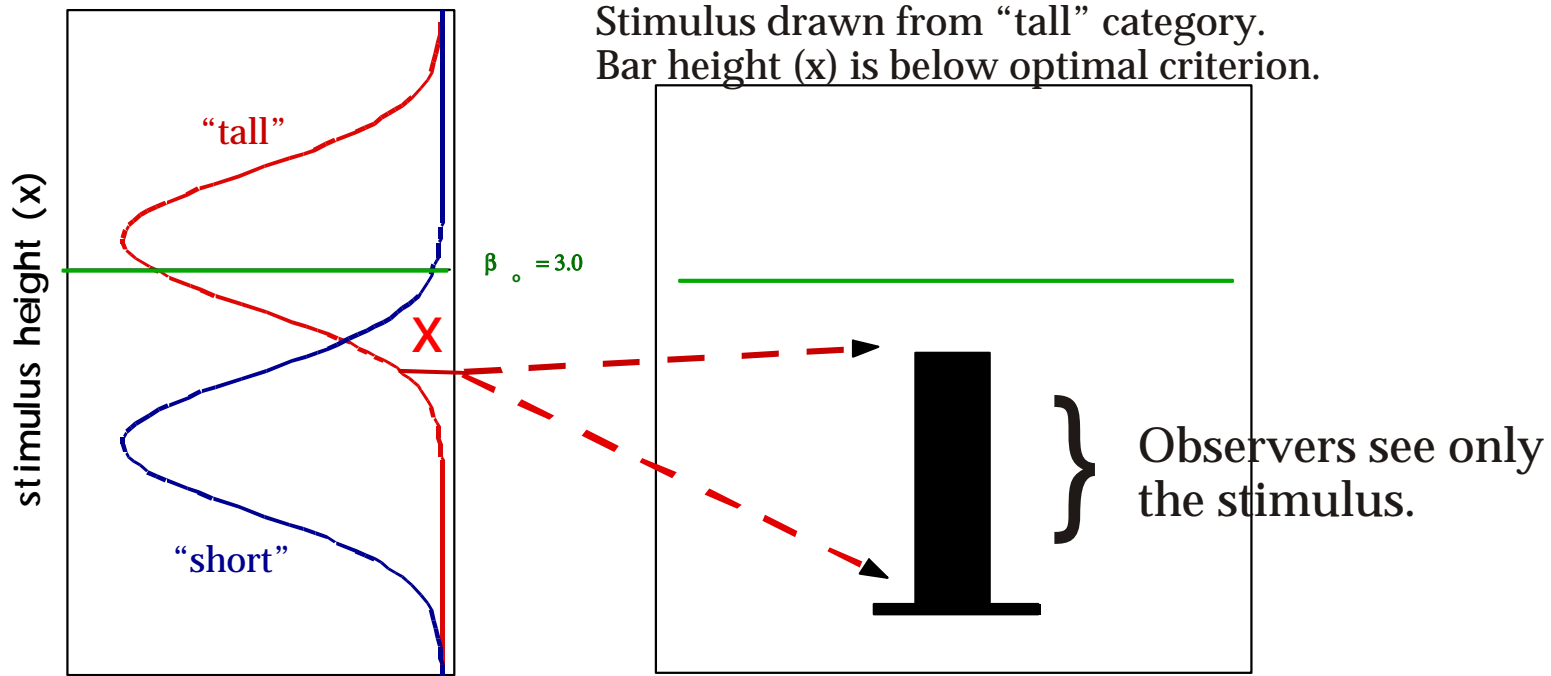
- 60+ warm-up trials (until observer reaches minimum performance criterion)
- 3 blocks of 60 training trials with feedback
- 1 block of 60 transfer trials with no feedback (results not presented here).
- Observers earned points for each response based on payoff matrix values; instructed to maximize point total.

Negative Cost

		Category	
		A	B
Response	a	2	-1
	b	-1	0

Zero Cost

		Category	
		A	B
Response	a	3	0
	b	0	1



Stimulus drawn from "tall" category.
Bar height (x) is below optimal criterion.

If Observer Responds "short" (incorrect)

Objective Feedback

Actual Gain:	0
Potential Gain:	3
Your Total Points:	4
Potential Point Total:	10

Optimal Feedback

Actual Gain:	0
Potential Gain:	0
Your Total Points:	4
Potential Point Total:	7

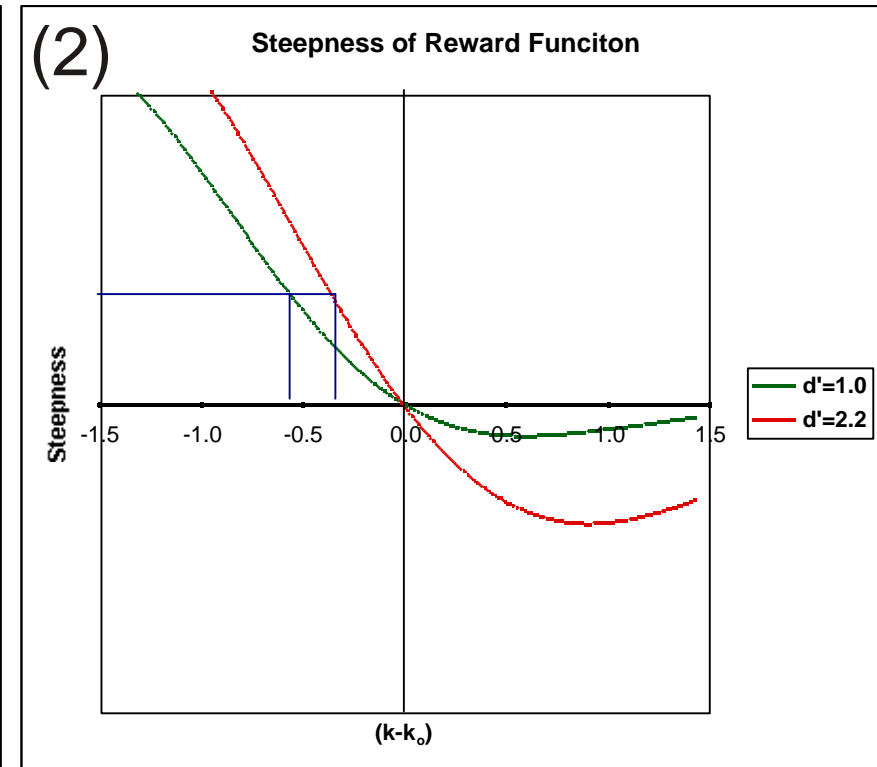
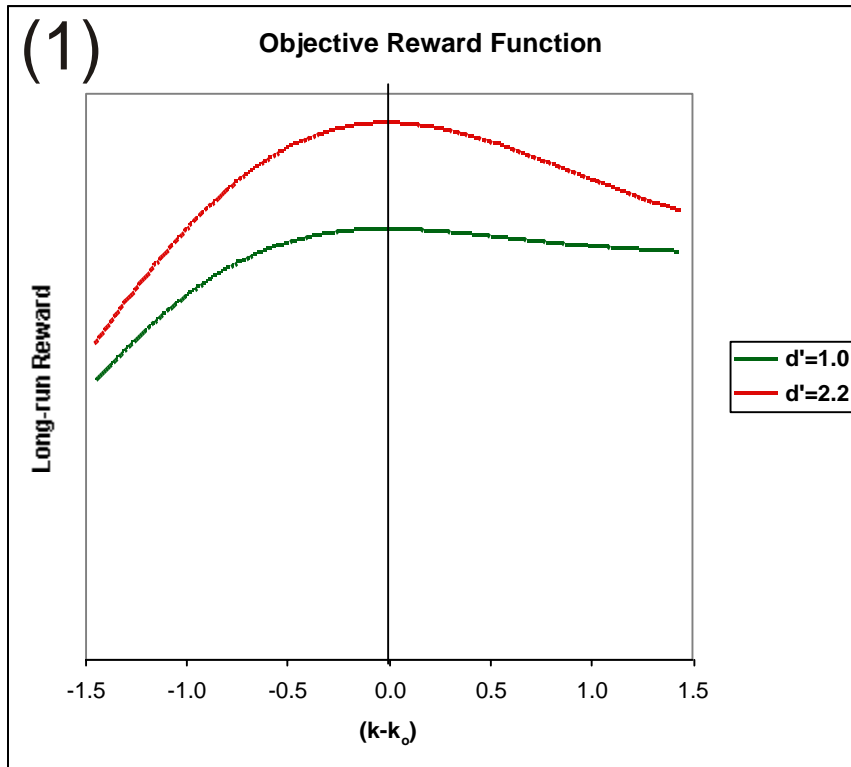
MODELING ANALYSES

Results: On average, performance was closer to optimal (based on signal detection β values, Green & Swets, 1966) when $d'=2.2$ than when $d'=1.0$, with optimal-classifier feedback than with objective feedback, and with zero-costs than with negative-costs in the payoff matrix.

We examined performance more closely using a set of models derived from Decision Bound Theory (Maddox & Ashby, 1993) that were fit to the data from each experimental condition simultaneously, but separately by subject and block. Two important hypotheses regarding the effects of category discriminability (the flat-maxima hypothesis; von Winterfeldt & Edwards, 1982) and the trade-off between benefits and costs (the COBRA hypothesis; Maddox & Bohil, 1998) are discussed below. The assumptions of these hypotheses are incorporated into the models fit to the data.

Category Discriminability and the Steepness of the Objective Reward Function

- (1) Objective Reward Functions (ORFs) for the levels of category discriminability used in the study. ORFs show expected reward as a function of deviation from the optimal decision criterion ($k-k_0$).
- (2) Steepness of the ORF for each level of d' . A single steepness value leads to different $k-k_0$ values depending on d' (note smaller deviation from optimal when $d'=2.2$).



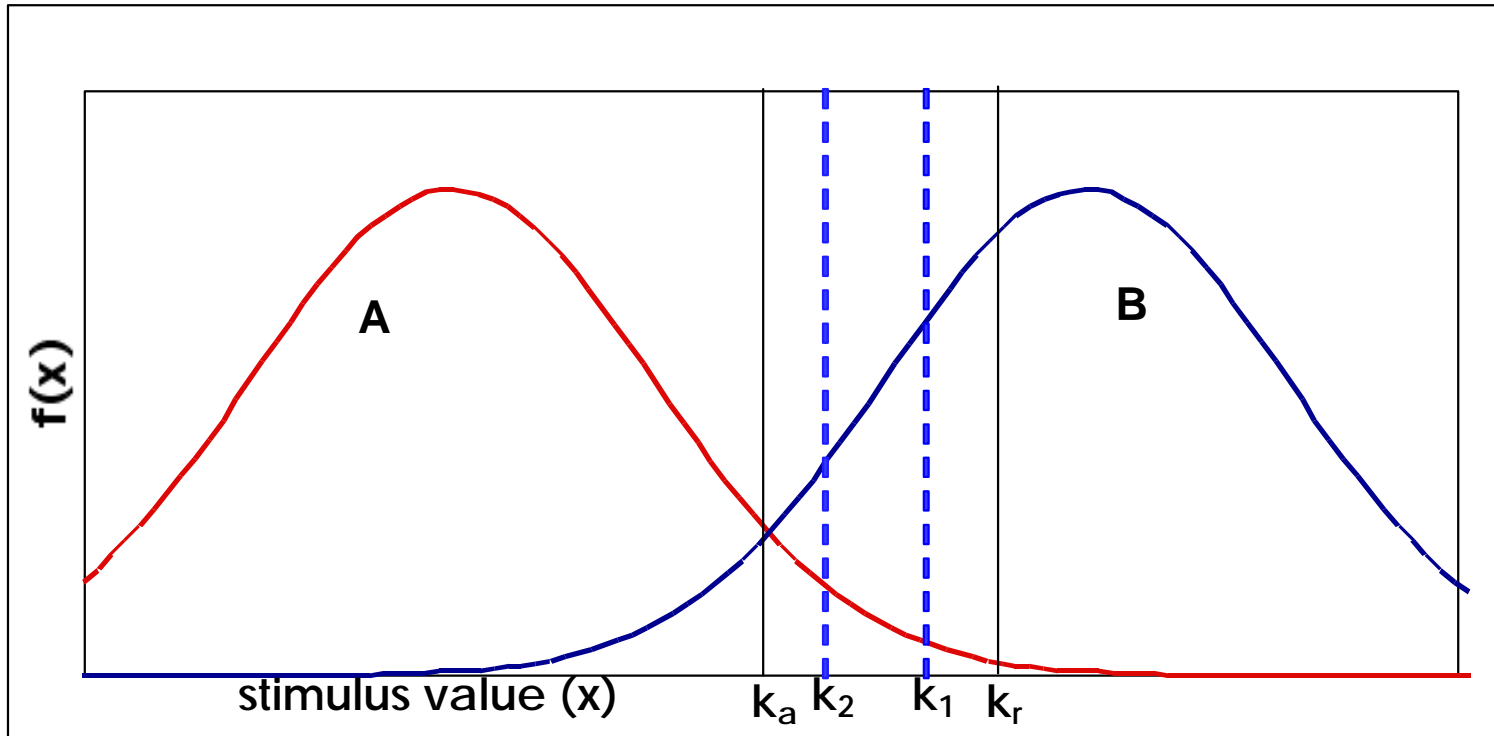
$$k = \ln(\beta) / d'; \quad k_0 = \ln(\beta_0) / d'$$

β is based on subject's response rates; β_0 is based on optimal response rates.

Prediction: Function steepness might influence noticeability of change in reward associated with criterion adjustment. If so, steeper function ($d' = 2.2$ ORF) should lead to better performance than flatter function ($d' = 1.0$ ORF).

Competition Between Reward and Accuracy (COBRA)

Schematic illustration of competition between reward and accuracy (COBRA)



$k_{r(\text{eward})}$: criterion used by the subject to try and maximize reward

$k_{a(\text{ccuracy})}$: criterion that maximizes expected accuracy

k_1 & k_2 : criteria used by two subjects that place different weight on accuracy

Prediction: less weight placed on accuracy should lead to increased long-run reward

Decision Bound Models

The models were Hybrid Flat-maxima/COBRA models (Maddox & Dodd, 2001), which instantiate assumptions of the Flat-maxima and COBRA hypotheses together in a single framework in order to gauge the hypotheses' ability to account for each subject's data. The models determine the subject's decision criterion (k) values for the various experimental conditions according to the following equation:

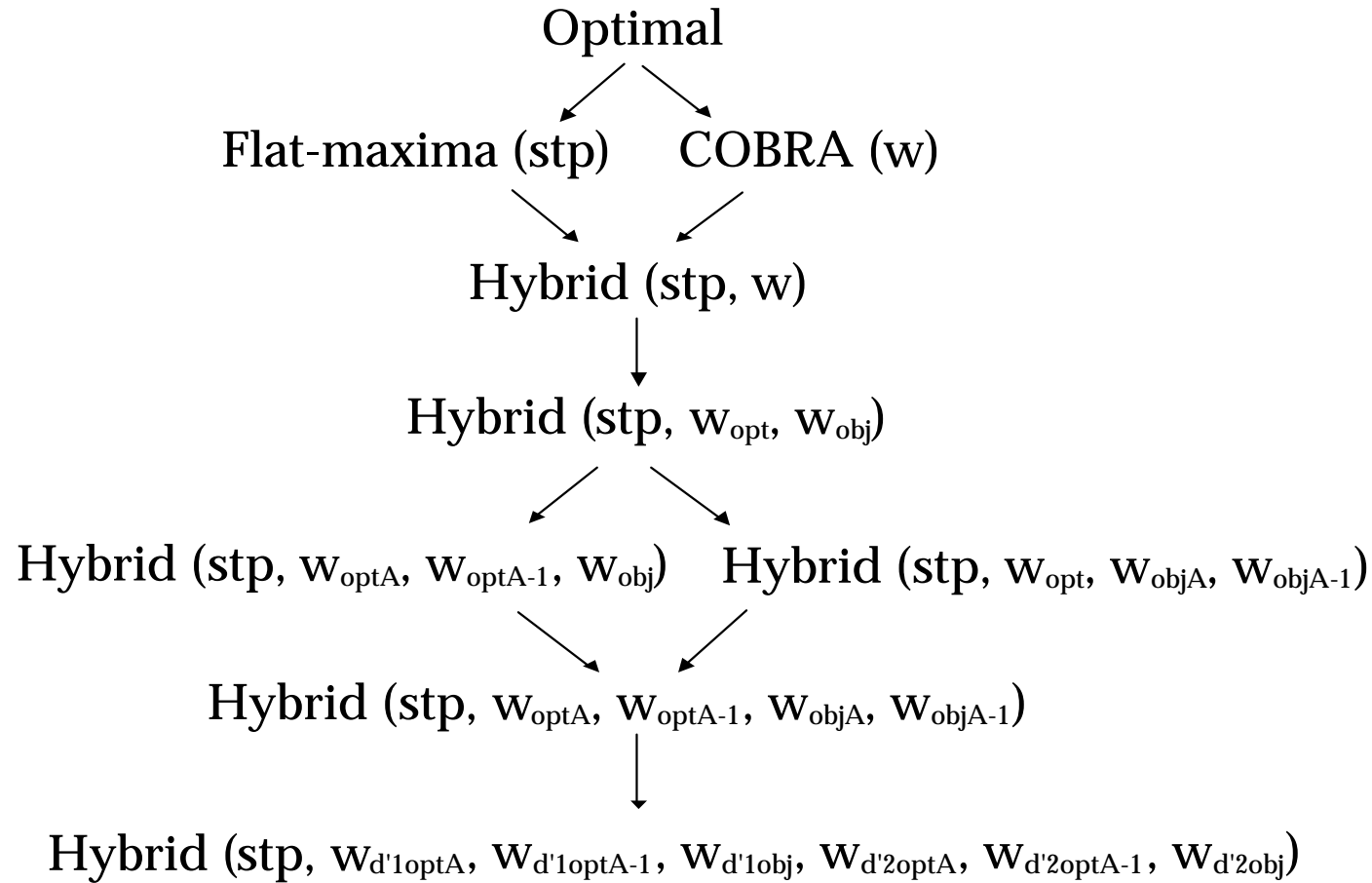
$$k_i = w k_{\text{accuracy}} + (1-w) k_{\text{reward}}$$

where k_i = the criterion used in condition i (values in red can be fixed or vary freely, depending on the model; k_{reward} is actually derived from a free steepness parameter).

Four main types of models (additional Hybrid model variants shown below)

Model	k_{reward}	w
Optimal	k_{optimal}	0
Flat-maxima	Derived from free steepness parameter	0
COBRA	k_{optimal}	Free ($0 < w < 1$)
Hybrid Flat-maxima/COBRA	Derived from free steepness parameter	Free ($0 < w < 1$)

Nested relationship among the decision bound models applied to each data set



Models indicate that subjects can be divided into 2 groups:
(based on class of most parsimonious model for 2 of 3 blocks)

"COBRA subjects" (7 of 15 subjects)

- Subjects **were not affected** by feedback or cost manipulations.

"Hybrid subjects" (7 of 15 subjects)

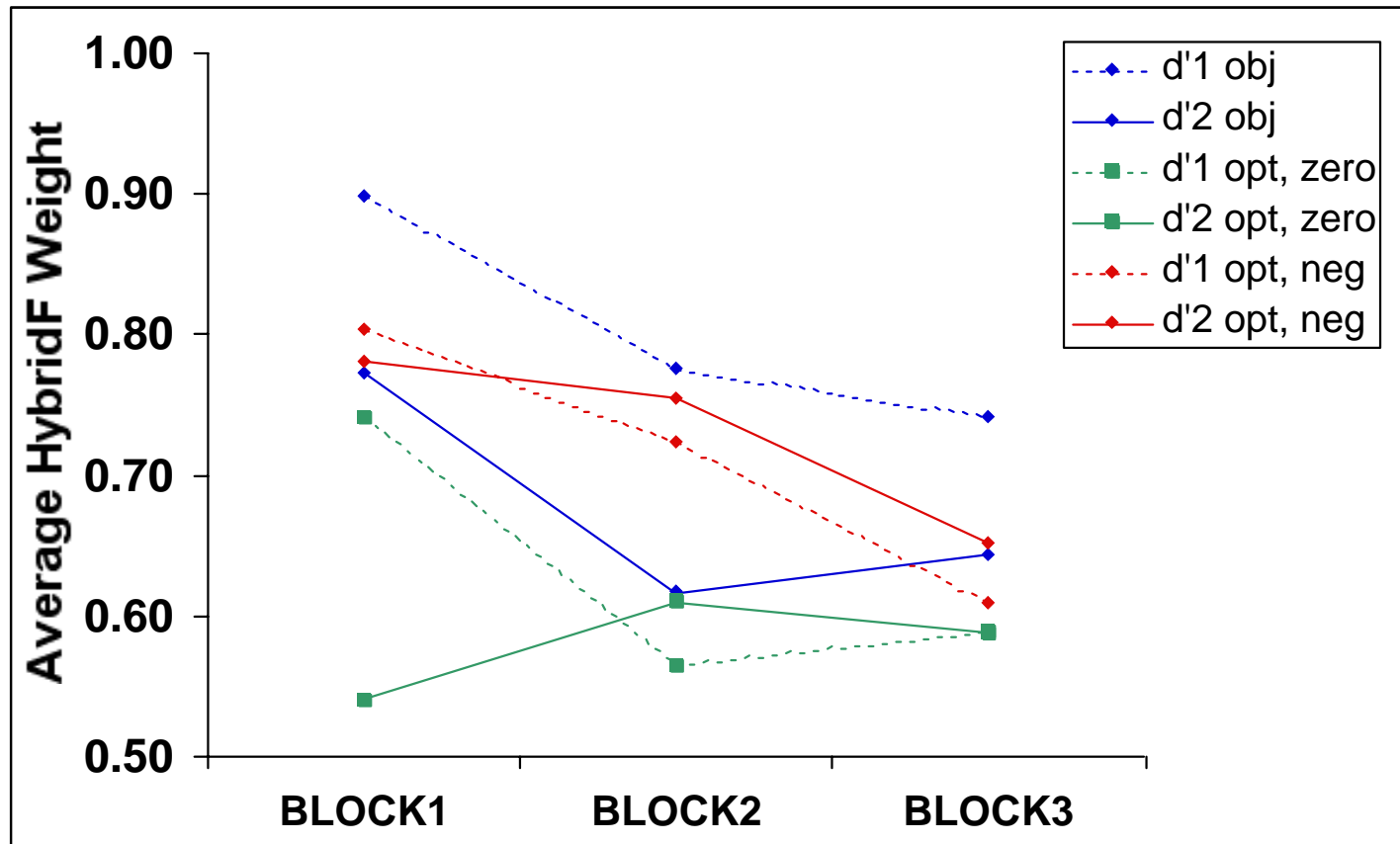
- Subjects **were affected** by feedback & cost manipulations:
 - Separate weights for objective & optimal-classifier feedback conditions
 - Separate weights for zero-cost & negative-cost conditions

Number of times separate matrix weights were needed:

- Optimal-classifier feedback: 16 of 20 cases
- Objective feedback: 4 of 20 cases

Average weight parameter values for "Hybrid subjects"

Based on parameter values from the most general Hybrid model (provided the best global account of results for these subjects)



- Generally, more weight on accuracy with objective feedback than with optimal- classifier feedback
- More weight on accuracy with negative-costs than with zero-costs

CONCLUSIONS

1. Modeling analyses illuminate **individual differences** obscured by aggregate data analysis.
 - Half affected by feedback and cost manipulations
 - Half unaffected by feedback and cost manipulations
2. Less weight on accuracy in optimal-classifier feedback conditions than in objective feedback conditions.
3. Less weight on accuracy in zero-cost conditions than in negative cost conditions.
4. **Less weight on accuracy corresponds to smaller deviations from optimal reward (smaller k-ko values).**

References

- Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*, New York: Wiley.
- Maddox, W. T. & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*, 49-70.
- Maddox, W. T. & Bohil, C. J. (1998). Base-rate and payoff effects in multidimensional perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *24*, 1459-1482.
- Maddox, W. T., & Bohil, C. J. (2001). Feedback effects on cost-benefit learning in perceptual categorization. *Memory & Cognition*, *29*, 598-615.
- Maddox, W. T., & Dodd, J. L. (2001). On the relationship between base-rate and cost-benefit learning in simulated medical diagnosis. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, accepted for publication.
- von Winterfeldt, D. & Edwards, W. (1982). Costs and payoffs in perceptual research. *Psychological Bulletin*, *91*, 609-622.

Acknowledgements

Research supported by National Science Foundation Grant SBR-9796206 and Grant # 5 R01 MH59196 from the National Institute of Mental Health, National Institutes of Health.